

Securely Extending Tag Sets to Improve Usability in a Video-Based Human Interactive Proof

Master's Thesis Proposal

Kurt Alfred Kluever
Department of Computer Science
Rochester Institute of Technology
Rochester, NY 14623 USA
kak2112@cs.rit.edu

Chair:	Dr. Richard Zanibbi	rlaz@cs.rit.edu
Reader:	Dr. Roxanne L. Canosa	rlc@cs.rit.edu
Observer:	Dr. Zack Butler	zjb@cs.rit.edu

May 13, 2008

1 Summary

Human Interactive Proofs (HIPs) are a class of automated challenges used to differentiate between legitimate human users and automated, malicious robots on the internet. HIPs have many practical security applications, including preventing the abuse of online services such as free email providers. The term HIP is preferred in this thesis over the more common (and unfortunately trademarked) term, *Completely Automated Public Turing tests to tell Computers and Humans Apart* (CAPTCHAs). HIP challenges should be easy for a machine to automatically generate, easy for a human to solve, and difficult, or impossible, for another machine to solve. A more detailed list of desirable properties can be found in the next section. The key to developing a successful HIP challenge is to choose a difficult artificial intelligence problem where a gap exists between human and machine capabilities.

Current implementations are based on hard artificial intelligence problems such as natural language processing [88], character recognition [25], image understanding [26], and speech recognition [46]. Most commercial implementations require the user to transcribe a string of distorted characters with background noise. This type of HIP poses unsolvable challenges to blind users, and, for all practical purposes, can be considered broken through techniques such as shape matching [56] [82], distortion estimation [57], and pattern recognition [92]. For example, Moy *et al.* wrote a program to successfully solve EZ-Gimpy with a 99% success rate [57]. The need for a more robust and user friendly HIP arises.

While the current trend in online media is towards streaming video, the use of video in HIP challenges has not yet been explored. This thesis introduces the concept of *Video Human Interactive Proofs*. In the proposed HIP implementation, users will be prompted to view a challenge video and then appropriately annotate (or tag) it. The challenges will be graded via exact matching of the user's response against a database of ground truth tags for the video. The database of ground truth tags will be populated using numerous sources: author-supplied tags, author-supplied title, tags of neighboring videos, alternate forms of the author-supplied tags (stemming [52] [65], synonyms, etc.), and combinations of this list. A user study will be conducted to answer many questions that are important in the field of *human computer interactions* (HCI). This thesis will evaluate and compare the usability (determined by the user study) and security

(determined by success of a database attack) of the video HIP when the different tag sources are used as ground truth.

2 Hypothesis, Problem, or Question

In order to evaluate the success of the proposed HIP, a list of desirable properties must first be established. Building off of suggested desirable properties from researchers at CMU [86], PARC [10], and Microsoft Research [70], the following set of amended properties are proposed and explained:

Automation Challenges must be easy for a machine to automatically generate and grade.

Open The database(s) and algorithm(s) used to generate the challenges must be publicly available to ensure that the difficulty of the HIP stems from a hard AI problem and not a secret algorithm. This is in accordance with Kerckhoffs' Principle, which states that a system should remain secure even if everything about the system (except the key) is public knowledge [45].

Usability Challenges should be easily and quickly solved by humans. The test should be independent of the user's language, physical location, educational background, and perceptual disabilities.

Security The underlying AI problem must be a well-known and well-studied problem where the best existing techniques are far from solving the problem. No-effort and brute-force attacks must not be successful.

Thesis Statement: *A novel video HIP where the database of ground truth tags has been intelligently extended has better usability for humans and is at least as secure against a frequency-based database attack than one where the database of ground truth tags consists only of author-supplied tags.*

The video HIP will be evaluated against the above list of properties. It is easy to show that the video HIP satisfies two of the properties (automation and open). The remaining two properties (usability and security) must be evaluated through experiments.

Automation The challenges can be easily generated by harvesting videos and tags from online video databases, or a large, local, dynamic database of videos may be used if available. The database of ground truth tags can also be easily generated using the algorithms and procedures detailed in this thesis. The challenges are also easy to grade using a set of well-defined scoring rules. Scores are computed using exact matching of user responses against the database of ground truth tags. A functional video HIP satisfies this requirement.

Open Challenges will be generated using videos from publicly available video databases. The algorithms and procedures used to generate the database of ground truth tags will be thoroughly explained in this thesis. Therefore, this requirement is satisfied.

Usability To evaluate whether the video HIP is easy for humans, 30 participants of various genders, ages, and educational backgrounds will be asked to tag 30-50 video. The following will be recorded: time required for the user to submit a tag and the tag which the user submits. Using this data, the pass/fail decision of the separate sources mentioned above can be computed in a post-processing fashion. The participants will also be asked a subjective question of how difficult the tagging task was. Note that the users will not receive a pass/fail decision during the user study (this will be computed offline at a later date). While the proposed video HIP is not language independent, localized versions could be developed by serving appropriate videos from localized versions of the online video database (YouTube.de, YouTube.fr, etc.).

Security To evaluate whether the video HIP is hard for machines, a database attack will be attempted against the video HIP. Since the underlying video database is public, attackers may attempt to replicate the database. Online video databases index tens of millions of videos (for example, YouTube.com indexes over 83.4 million videos as of April 9, 2008), which makes such an attack difficult. Nevertheless, such an attack vector will be evaluated. Furthermore, attackers may attempt to utilize important facts about the publicly available video selection algorithm and ground truth generation. Videos which will be included in the challenges must meet certain criteria (length, appropriate content, number of tags,

popularity, etc.). A clever attacker may attempt to selectively replicate the database to reduce the search space and to use frequently occurring tags. Such an attack will also be evaluated.

It is hypothesized that the human success rate (usability) for the control will be less than that of the extended method and that the attack success rate (security) for the extended method will be no worse than that of the control.

3 Prior Research on the Topic

3.1 Motivation

Early internet hackers of the 1980s are often credited with the idea of obscuring text to foil content filtering devices. When trying to obscure sensitive data online, these hackers would use the simple technique of substituting numbers for text (e.g., $I \rightarrow 1$, $A \rightarrow 4$, etc.). The algorithms they used were extremely rudimentary but would fool content filters and web crawlers. The technique was originally used to get around profanity filters which were installed on online commenting systems and forums. These filters were extremely simple and banned a predefined set of inappropriate words. The end result was something that a fellow human could read with a minor degree of difficulty but a computer could not. This concept eventually evolved into what is currently known as “133tspeak” (elite speak).

With the recent massive increase in the amounts of spam being delivered to inboxes around the world, people have been concerned with publishing their email address online. Spam bots can process thousands of web pages per hour, scanning for email addresses in clear text. Many people have resorted to attempting to fool spam bots by posting their email addresses in a human readable but unconventional form. For example, the text string “kurt AT kloover DOT com” is not easily parsed or understood by spam bots which crawl for email addresses in conventional forms. However, humans can easily identify the intended email address. This solution is not perfect, as any text-based data that humans can easily read can also be easily read by a machine. A simple set of regular expressions could easily match many predefined text masquerading patterns and successfully harvest the obscured email address. Thus, a need to protect data against automated machine processing arose.

Differentiating between a user and machine over the internet has significant importance in the fields of internet security, artificial intelligence, and human computer interactions. Examples of where HIPs are useful include the following applications:

- Preventing automated account registration (email, online marketplaces, etc.) [63]
- Preventing artificial inflation of product ratings or voting in online polls
- Preventing outgoing spam [40]
- Preventing blog comment or forum spam [24]
- Preventing automated mining or duplication of a website’s content
- Preventing denial of service attacks against web servers [55] [44]
- Preventing brute force password cracking [62] [89] [32]
- Preventing phishing (acquiring sensitive information by masquerading as a legitimate service) [34]
- Automating document authentication [36]
- Tagging images through the use of collaborative filtering [31]
- Digitizing physical books [87]

HIPs exploit the fact that many artificial intelligence problems remain unsolved. As observed by Luis von Ahn, HIPs that are based off of interesting AI problems present a win-win situation [85]. Either the HIP is successful and provides a method for differentiating between humans and machines on the internet, or the HIP is broken and the underlying hard artificial intelligence problem is also solved. The struggle between researchers and attackers continues to evolve this set of dynamic problems and advance the field of artificial intelligence.

HIPs have received criticism from the HCI community [53] [15]. Unfortunately most web users are unaware of the reasons why the websites are forcing them to pass these “annoying” challenges. Some HIPs

are extraordinarily difficult and continue to frustrate legitimate users. For example, Yahoo! discontinued the use of the Gimpy HIP due to a slew of user complaints. Users with perceptual disabilities have an even larger disadvantage against HIPs. The key to developing a successful human verification technique is to correctly balance security, usability, and accessibility.

3.2 Theory Base for the Research

In 1950, Alan Turing provided an operational definition of intelligence using the imitation game [83]. The *Turing test* is administered by a human judge and taken by a human participant and a machine participant, both of which attempt to appear human over a text-only channel. If the human judge cannot reliably determine which participant is the machine and which is the human, the machine is said to have passed the Turing test. Literature describes a *reverse Turing test* as a test in which any of the above roles have been switched. HIPs are a slight modification of a reverse Turing test, where the challenge is administered by a machine and taken by a human. The burden is on the human participant to convince the machine that he is human. Furthermore, the challenge should not be solvable by any machine. Notice the paradox that this creates: the machine can automatically create, administer, and grade a test that it itself cannot pass [86].

In order to retrieve neighboring videos, a small amount of information retrieval theory will be used. Mathematical information retrieval models can be separated into 3 categories: set-theory, algebraic, or probabilistic. The most common information retrieval model is a model based on set-theory operations (union \cup , intersection \cap). However, the standard boolean model is often criticized: [77]:

1. Queries are often difficult to formulate for untrained or unexperienced users. Well formulated queries are more of an art than a science.
2. It is impossible to control the number of results from a query. Some queries may return no results while others may return an unmanageable number of results.
3. Returned items cannot be ranked. All items returned are presumed to be of equal importance.
4. Weights cannot be assigned to documents.
5. Results can be counter intuitive. For example, consider the query “A or B or ... or Z”. Documents which matches all of the terms in the query are labeled with the same importance as documents which match only a single term. Similarly, the query “A and B and ... and Z” will only return documents which match every single term and not return a document which matches all but 1 of the terms.

To overcome these shortcomings, extended boolean models have been proposed and evaluated [48]. These models require weights to be assigned to the query terms. A form of approximate matching will be used to retrieve the tags neighboring videos. It is also important to note that YouTube provides a mechanism to retrieve related videos using a proprietary search algorithm [93]. Since no public definition exists on how related videos are selected, it remains difficult to formally evaluate this feature.

3.3 Prior Work

While often uncredited, Moni Naor was the first to informally define the concept of online human verification [59]. In a 1996 unpublished draft, he theorized nine possible sources for HIPs which can be separated into 3 logical categories:

Text Based HIPs (Hard AI Problem: natural language processing)

1. **Filling in words** Given a sentence without a subject, select which noun best fits the sentence.
2. **Disambiguation** Given a set of sentences with a pronoun, determine what noun the pronoun refers to.

Image Based HIPs (Hard AI Problem: understanding semantic content in images)

3. **Gender recognition** Given an image of a face, determine which gender it is.
4. **Facial expression understanding** Given an image of a face, determine the mood of the person.
5. **Find body parts** Given an image of a body, locate a certain body part.

6. **Deciding nudity** Given several images, determine which image contains the nude person.
7. **Native drawing understanding** Given an image, determine what the subject of the image is.
8. **Handwriting understanding** Given an image of a handwritten word with noise added, transcribe the word.

Audio Based HIPs (Hard AI Problem: understanding spoken languages)

9. **Speech recognition** Given a clip of spoken audio, transcribe the text.

It is important to note that all popular HIP implementations that have been suggested in the past 12 years of literature have been either based on, or are a slight variation from, the above list. Therefore, existing HIP research can be classified into one of the above three logical categories.

3.3.1 Text Based HIPs

An important distinction must be made between true text based HIPs and image based HIPs containing text. Text based HIPs include challenges that are rendered in standard character encoding schemes, such as ASCII. The text appears as normal, inline text on a web page. Humans who are both blind and deaf still frequently use the internet through the aid of Braille readers [61], but are often locked out of many services which rely on HIPs which do not fall into this category. The HIPs in this category are the only ones that can pride themselves on being truly accessible to all humans regardless of perceptual disabilities.

The problem with developing such HIPs is that they are extremely difficult to automatically generate. Common techniques utilize a finite set of phrase templates in which variables are substituted. This finite set of templates increases the likelihood of a successful attack. HIPs of this type are also vulnerable against machine attack because they do not require any audio or image understanding abilities. This type of HIP only requires a trivial amount of natural language processing. Due to this, very little research has been devoted to this type of HIP.

At the HIP 2002 workshop, the concept of text-based HIPs was presented in two extended abstracts [39] [66]. Several months later in a preliminary manuscript, the benefits and difficulties of constructing a *Natural Language CAPTCHA* (NLC) [38] were explored. A NLC could be presented in a variety of formats: rendered in an image, spoken as audio, or in plain text. When presented in plain text, it could be solved by users with visual and/or hearing disabilities. While the appeal of a NLC is strong, the author was not able to find a suitable generation algorithm, and further warns that a NLC based on recognizing coherent natural language is most likely impossible using current models of natural language.

In 2004, researchers attempted to use word-sense ambiguity to construct a HIP [14], similar to Naor's 2nd recommendation. They attempted to exploit the fact that different words can have similar meaning, depending on the surrounding context. They also showed that it is possible to use input from previous tests to train the underlying linguistic model (similar to collaborative filtering).

In 2006, a natural language HIP was presented that required the user to determine the funny knock-knock joke out of a set of 3 jokes (1 real and 2 constructed jokes) [88]. While it was demonstrated that a gap between humans and machines did exist, a machine can pass such a challenge with 33.3% by random guessing. Requiring a user to successfully pass such a challenge twice in a row (serial repetition) drops the machine success rate to 11.1%, but this is still unacceptably high.

The question of whether any other hard AI problem exists that is expressible in natural language and fits the requirements of a HIP remains open.

3.3.2 Image Based HIPs

By far the most popular media format choice for HIPs has been images, specifically those containing strings of distorted text (a variation of Naor's 8th recommendation).

In 1997, one year after Naor's recommendations, AltaVista developed the first concrete implementation of a HIP. AltaVista had recently been receiving many automated URL submissions to their search engine database by spam bots. A group of researchers from the Digital Equipment Systems Research Center were contracted to develop a solution to prevent such an attack. To combat this, the team of developers created

an image based verification system based on Naors suggestion of recognizing handwritten images. However, they soon realized that although an image containing text was a step in the right direction, it could easily be foiled by use of OCR software. *Optical Character Recognition* (OCR) software is designed to translate images of text into a machine editable form. The team researched the limitations of scanners with OCR capabilities, and exploited the weaknesses of the OCR systems when rendering their HIPs. In order to improve OCR results, the manual suggested using similar typefaces, plain backgrounds, and no skew or rotation. To create an image that was resilient to OCR, they did the exact opposite of the suggestions.

In the summer of 2000, Yahoo! began to experience a similar problem where their chat rooms were being spammed by chatbots. Dr. Udi Manber, Yahoo!'s chief scientist, contacted researchers at Carnegie Mellon University to develop a solution to this problem. This request gave birth to the CAPTCHA project, which was led by a cryptographer and theoretician Manuel Blum and his graduate student, Luis von Ahn. *Completely Automated Public Turing test to tell Computers and Humans Apart* (CAPTCHA), a play on the word "gotcha", is defined as any automated, public method for differentiating between humans and computers on the internet. The trademarked acronym has become widely used, however the term HIP is preferred here. The CMU team is credited with the first formal mathematical definition of the problem [85].

Meanwhile, researchers at Georgia Institute of Technology developed similar technique as CMU. A string of characters is rendered into an image form which humans can easily transcribe through the use of human pattern recognition skills, but an automated robot (a Turing machine) cannot [90]. They compared the problem to a one-way trapdoor (non-reversible) hash function, which they termed a Turing-resistant hash.

In 2001, researchers at the Xerox Palo Alto Research Center and the University of California at Berkeley synthesized low quality images of machine printed text using a range of words, fonts, and image degradations [29] [5]. Following Baird's quantitative stochastic model of document image quality [3] and a list of problematic OCR examples [58], noise was introduced into the rendered strings by using a set of image-degradation parameters.

A couple of years later, again at the University of California at Berkeley and the Xerox Palo Alto Research Center, a reading based HIP known as BaffleText was developed [25] [7]. BaffleText exercised the Gestalt perception abilities of humans. Gestalt perception states that humans are extremely good at recognizing and understanding pictures despite incomplete, sparse, or fragmented information. Again, Baird's model was used to apply three types of mask operations to binary images: addition, subtraction, and difference. Addition is the equivalent of boolean OR, subtraction is equivalent to NOT-AND, and difference is equivalent to XOR. A significant contribution was the use of pronounceable, non-dictionary character strings. The character strings were generated by a character trigram Markov model which was trained on the Brown corpus [47]. This prohibits a simple dictionary attack while still being somewhat user friendly.

Typically, OCR systems separate the recognition task into two sub tasks: segmentation and classification. In 2003, researchers at Microsoft Research argued that the segmentation task is much more difficult than the classification task for OCR systems. They developed a HIP which represent hard segmentation problems, as opposed to hard classification problems [80]. Another contribution was the observation that HIP researchers have the advantage in the battle against HIP attackers. This is because HIP generation is a synthesis task while attacking a HIP is an analysis task. Analysis is orders of magnitude more difficult than synthesis, especially during lossy synthesis. In the synthesis task, the researcher has the ability to use randomness and creativity, whilst in the analysis task, the attackers are tightly constrained by the decisions made by the creator.

Building off of the idea that segmentation is more difficult than recognition, ScatterType was developed at Lehigh University [9] [11] [8]. The challenges are pseudorandomly generated images of text which has been fragmented using horizontal and vertical cuts and scattered using horizontal and vertical displacements. To defend against dictionary attacks, the text strings are English-like, but non-dictionary words (similar to BaffleText). A human study was performed and showed that human legibility averaged 53% and exceeded 73% on the easiest challenges.

A formal study of user friendliness for transcription tasks was conducted at Microsoft Research [22] [20] [21]. They studied the effects of varying the distortion parameters and attempted to determine the optimal parameters where the HIPs prove hard for machines but easy for humans. As researchers found in the past,

the most effective HIPs are segmentation based challenges, which continues to be a computationally difficult task.

In 2006, a formal evaluation of string generation methods was conducted at Avaya Labs [13]. The three main challenge string generation choices were considered: dictionary words, Markov text, and random strings. The authors argued that all of the existing models exhibit substantial weaknesses. It was suggested to use *consonant-vowel-consonant* (CVC) trigrams of psychology as an improvement to the Markov text model. The three original methods and the improved method was evaluated using the ScatterType system [12] to determine the correct balance of image degradations and familiarity.

A researcher from the National Chengchi University in Taiwan has developed an interesting HIP reading challenge using textured patterns [49]. Instead of rendering characters in a different color than the background, the foreground is rendered as a different texture as the background. Humans have a relatively easy task of reading the textured text from the image. As stated before, segmentation continues to be the more difficult task for most OCR systems. Segmenting such an image is extremely hard because the texture of the characters must be analyzed.

Researchers at the University of Buffalo, CEDAR have contributed a great deal of research building off of Naor’s 8th recommendation (“Handwriting recognition”) [72] [73] [74] [75] [76]. Features are removed and non-textual strokes/other noise is added to images of handwritten words, while preserving Gestalt segmentation and grouping principles. Attempts at attacking the HIP using state of the art handwritten word recognizers yields a success rate of less than 10%, while the accuracy of human readers is over 75%.

However, requiring a user to recognize characters is not the only reliable image based method. Semantic image understanding tasks have also been proposed. Researchers from the University of California at Berkeley investigated a set of three image recognition tasks [26] [27]:

1. **Naming images** Determine the common term associated with a set of 6 images
2. **Distinguishing images** Determine if two sets of images contain the same subject
3. **Identifying anomalies** Identify the “out one out” from a set of 6 images

Note that the excellent work presented by Chew and Tygar (specifically the “Naming images” task) served as a major inspiration for the video HIP. The problems which affected human performance were evaluated and tested during an in-depth user study. Two formal metrics for evaluating HIPs were also proposed as well as attacks on the three image recognition based HIP implementations. The first metric evaluated HIP efficacy with respect to the number of rounds of a HIP and the second metric measured the expected time required for a human to pass the HIP. Further details can be found in the appendix of [27].

In late 2003, researchers at Microsoft Research argued that the most familiar objects to humans are human faces. They developed a HIP designed to confuse face recognition algorithms while still being easy to use [68] [69] [70] [71]. The images are automatically synthesized from facial models, but end up looking rather eerie to many users. For this reason, the system was never adopted.

A novel approach to image based HIPs was presented by researchers at the University of Memphis where they used a 3D model to generate a 2D image based HIP [43]. Random rotations, distortions, translations, lighting effects, and warping were applied to the 3D model to result in a 2D image. After these manipulations, the image is displayed to the user who is prompted to determine the subject contained in the image. By rendering and morphing the images from models, the image database is infinite in size. However, the downside is that the user must choose a label from a fixed, finite set of approximately 30 words. While the test data is infinite, the subject selection is severely limited.

An interesting attempt was made to harness the faults in the human visual system [17]. Optical illusions occur due to failures in our visual system or brain and make us see and believe something that in fact is not true. Conventional machine vision systems do not suffer from such visual illusions (but some have been specifically designed to). All existing HIPs have been problems where humans have surpassed a machine’s abilities. However, visual illusions present a situation in where the visual system fails in comparison to machines. This gap, albeit a negative gap, can still be exploited in the form of a HIP.

Building off of Chew’s approach, a robust image based HIP was developed in 2005 [33]. The system, called IMAGINATION, performs controlled distortions on randomly chosen images and presents them to

the user for annotation. The distortions are chosen to have a negligible effect on human performance, while simultaneously confusing content based image retrieval systems. The word list was carefully chosen to avoid ambiguous labels while still providing security against attacks.

A similar approach to a face recognition based HIP was developed in 2006 by researchers at George Mason University [54]. Photographs of human faces were mined from a public database and distorted. The user is then prompted to match distorted photographs of several different humans. This HIP has the bonus of being completely language independent (not counting the minute amount of textual instructions required to complete the task).

A recent image based HIP, also by Microsoft Research, prompts the user to identify images of cats from a set of 8 images of cats and dogs [35]. The task is remarkably simple, yet continues to baffle even the best attempts using computer vision and content based image retrieval. However, the one downside of the approach is that the underlying image database has not been publicized. Through a partnership with PetFinder.com, Microsoft is able to harness the power of their large database of manually labeled images.

3.3.3 Audio Based HIPs

Automatic speech recognition (ASR) is severely affected by background noise, music, or other speech while human perception of speech in a noisy environment is very robust. In fact, humans need only a *signal-to-noise* (SNR) ratio of approximately 1.5 dB to recognize and understand speech [81], while even the most state of the art ASR systems require a SNR between 5 and 15 dB [84]. This large gap between human perception and ASR systems provides a great opportunity for a HIP.

At the First International Workshop Human Interactive Proofs [16], two audio based HIPs were presented: one by a team from Bell Laboratories [51] and one by Nancy Chan of the City University of Hong Kong [18]. Chan’s system overlaid white noise and other distractions onto a clip of spoken text to baffle ASR systems. The team from Bell Labs also chose to exploit the gap between humans and machines in natural language processing. Answering a simple spoken question requires processing the audio stream as well as understanding the semantic meaning of the question. A further analysis of 18 different sets of distortions on speech data was later performed at Bell Laboratories [46].

A year later, Chan attempted to break the audio based HIP using a *Hidden Markov Model* (HMM) [19]. While the results demonstrate that a small gap does exist in the understanding of synthesized speech with background noise, the gap is so small that Chan actually discourages against building audio based HIPs until the naturalness of synthesized speech improves. It seems that this recommendation carried a significant amount of weight, and very little research was performed in the area of audio based HIPs until 2007.

A new research group from the Sharif University of Technology has recently begun publishing papers in the area of HIPs. Their approach to audio based HIPs is very similar to the Bell Labs approach from 2002. They suggest creating a large database of question template which are then rendered into an audio clip using a *Text-To-Speech* (TTS) synthesizer [79]. An example question from this HIP: “There are 5 cats, 3 apples, and 4 dogs on a table. How many pets are there on the table?”. This challenge is trivial for humans, but computers would be required to possess three difficult abilities: speech recognition, natural language processing, and reasoning. The problem with such a HIP is that the templates for the questions must be manually generated (even if random values can be substituted into the templates). The list of possible challenge questions is therefore finite and prone to attack. They have also applied such a technique in the application of mobile phones [78].

3.3.4 Hybrid HIPs

In January 2005, researchers thought that current HIPs were too demanding of legitimate human users. Instead, they proposed *Implicit CAPTCHAs* which require as little as a single click [4]. Implicit CAPTCHAs are carefully woven into the expected sequence of natural browsing events. The challenges are so elementary that a failed challenge indicates an attempted bot attack. The authors suggest disguising necessary browsing links in images and claim that bots would not be able to find these hidden links. While the usability of the system is attractive, the system could easily be attacked on a case-by-case basis. Furthermore, it would

likely be susceptible to replay attacks, where the attack manually solves the challenge once and then replays the challenge as many times as necessary. This type of HIP may work for low traffic or low value services, but it would never survive in a large scale application.

At the 2005 HIP workshop, Lopresti presented his plans to leverage the HIP problem by constructing HIP challenges which offer simultaneous benefits to both security and pattern recognition [50]. The goal is to harness the cognitive “horsepower” of the millions of legitimate users who are required to solve the HIPs. Instead of constructing a contrived challenge, use real world problems which current computers struggle with and also use prior user input as ground truth training data.

Also at the 2005 HIP workshop, Chew and Tygar presented a similar idea as Lopresti known as *Collaborative Filtering CAPTCHAs*. Collaborative filtering CAPTCHAs are challenges where the ground truth answer is not known [28]. Instead, challenges are graded by comparing their response to the responses of other users. They suggest designing a challenge which evokes some aspect of humanity which is difficult to quantify such as quality in art, emotion, philosophical views, or humor. While their results are inconclusive, the direction for future work is an exciting area.

An interesting project known as reCAPTCHA, which is based on Lopresti’s and Chew’s ideas, has been developed at CMU [87]. The project is implemented as a web service (first suggested in [30]) that provides images of words from physical books as HIP challenges. The images are of words which modern OCR systems have failed to recognize. Since no ground truth is known for the word, two words are presented: 1 which the ground truth is known for and 1 which the ground truth is not known for. If the user correctly transcribes the word for which the ground truth is known, it is assumed that the other transcription is correct as well. After several people have transcribed an unknown word as the same string, it can then be labeled as ground truth with the transcription.

In October 2007, researchers attempted to avoid *laundry attacks* by animating the answer of the HIP inside the test itself [2]. A laundry attack is when the attacker posts the challenge to a malicious site and convinces unsuspecting visitors to solve the challenge for them (also known as the *pornographer-in-the-middle attack* in [50]). The HIP proposed is a Java applet where various objects are randomly animated inside the test window. The user is given a question and told to click on the correct answer. The animation prevents the unsuspecting user from inadvertently telling the attack where the answer is (as the object is constantly moving).

In hopes of providing an alternate for blind users, researchers at Towson University and the University of Notre Dame developed an audio and image based HIP [42]. The proposed HIP combines a photo of an object and the sound which the object makes (e.g., a photo of a cow and a clip of “moo-ing”). The challenge then becomes to identify the object that they see and hear. While this type of HIP may be easy for humans and hard for machines, these challenges cannot be automatically generated. Currently only 15 different image/audio combinations have been proposed. Since these challenges must be manually constructed and the list is finite in length, an attack against this HIP would be incredibly trivial. Therefore, while this is a valiant effort to provide accessibility to those with disabilities, it fails to meet the requirement that challenges can be automatically generated.

3.3.5 Attacks on HIPs

While there has been a lot of research devoted to creating successful HIPs, there has also been a considerable amount of interesting research devoted to the breaking of HIPs [67]. Note that the definition of “breaking a HIP” is to solve the HIP with an automated computer program (see [91] for further clarification of this ambiguity).

In June 2003, Mori and Malik used shape context matching to solve EZ-Gimpy with 92.1% accuracy and Gimpy with 33% accuracy [56]. They are credited with the first successful attack against image based HIPs which require the transcription of distorted text.

Also in June 2003, the shape context matching was used again to achieve 93.2% accuracy on the EZ-Gimpy challenge [82]. This technique computed the cost between templates and the images as the average symmetric chamfer distance of the letters and the variance of the letter distances. The use of two templates raised their accuracy from 89.5% to a Mori-and-Malik-beating 93.2%.

In June 2004, distortion estimation techniques were used to solve EZ-Gimpy with 99% accuracy and Gimpy-r with 78% accuracy [57]. Due to the limited and fixed size of EZ-Gimpy’s dictionary, every challenge image was easily tested against a template database. The distorted template image with the best correlation was returned as the result. Gimpy-r does not rely on a dictionary, and therefore required local distortions to be removed via distortion estimation techniques.

Also in June 2004, the first attempt was made at solving an image-based HIP which did not require the transcription of distorted text. The HIP required users to tell time from a distorted clock face. However, researchers were able to successfully solve this HIP with 87.4% accuracy [94]. Their method used 4 types of Harr wavelet-like features and AdaBoost to detect the clock face.

Later in 2004, researchers at Microsoft Research attacked several commercial HIP implementations with surprisingly high accuracy (80%-95%) [23]. They learned that most HIPs are pure classification tasks which can be broken with machine learning. Convolutional neural networks were used to perform character recognition. Their attacks had the most difficulty with the segmentation task, not the recognition task. Therefore, they suggested that researchers focus their efforts on building HIPs which rely on the segmentation task instead of the recognition task (ScatterType used this recommendation). It was later confirmed in July 2005 that computers are as good as, or better than humans at recognizing single characters under common distortion and clutter techniques [21].

In March 2005, the Holiday Inn Priority Club HIP was broken using linear regression and rotation of axes to undo the rotation of the string and bivariate Haar wavelet filters to recognize the characters [1].

In December 2007, simple pattern recognition algorithms were used to exploit design errors in several commercially available HIPs [92] and achieved near 100% accuracy. All previous attacks have focused on computer vision or machine learning algorithms, but researchers at Newcastle University chose to use a much simpler approach: letter-pixel counting. They found that several implementations distorted characters, but the pixel counts of the letters remained constant.

Many non-academic related attacks have also been informally documented on blogs and websites. Security analysts also recently confirmed that automated attacks have been reasonably successful at solving the Google, Yahoo! and Microsoft HIPs. While there is no published literature on the attacks, logs show that spammers are becoming more and more successful at attacking existing HIPs. Security experts are also concerned with the *pornographer-in-the-middle* attack where a spammer captures a HIP challenge and serves it to an unsuspecting user who is trying to access a website which the spammer has control over [50]. Furthermore, it has been theorized that spammers are outsourcing the solving of HIPs to third world countries. While this does require a spammer to pay a human for the answers to these HIPs, the value of the email account is often much more than the cost required to solve the HIP.

4 Research Approach and Methodology

The experiment will be evaluating the usability and security of the standard tag set and the extended tag set. The standard tag set consists only of author-supplied tags on the given video. The extended tag set consists of the standard tag set, words in the title of the video, stems of the tags, synonyms of the tags, and tags from neighboring videos. The two tag sets are formalized below.

4.1 Definitions

Author Supplied Tags $AST(v) = \{t_1, t_2 \dots t_n\}$ where t_i is a tag of video v . This is the set of tags chosen by the author who originally uploaded the video to the online video website. Since there is no requirement on the quality or number of tags required during the upload process, they can be a poor representation of the video. However, authors are motivated to correctly tag their video because poorly tagged videos will not show up in relative searches.

Author Supplied Title $ASL(v) = \{t\}$ where t is the title of video v . The title is chosen by the author during the upload process. The same data reliability issues are present with the title as are with the tags.

Stemming of Tag $STM(t) = \{s\}$ where s is the stem of tag t . The porter2 algorithm [64] (based on [65]) will be used.

Synonyms of Tag $SYN(t) = \{s_1, s_2, \dots, s_k\}$ where s_i is a synonym of tag t . An online thesaurus will be queried to accomplish this task.

Neighboring Videos $N(v) = \{n_1, n_2, \dots, n_l\}$ where n is a neighboring video of video v where a neighboring video is defined as: $n \in N(v)$ if $\exists t$ such that $t \in AST(v)$ and $t \in AST(n)$. This set is partially ordered on the strength of the link, where videos that have more overlapping tags come first. Informally, neighboring videos are two videos which share at least 1 author supplied tag. The hope is that videos that share tags are about related concepts.

Challenge Response $RES(v) = \{t\}$ where t is the user’s response to video v . This set contains a single value, which is the user’s guess at the correct tag for video v .

Score Set $S = \{s_1, s_2, \dots, s_r\}$ where s_i is the user’s score from round i . Given the scoring rules defined below, it is possible that multiple rounds of the video HIP will be required before a user is considered to have passed the HIP. This set consists of the scores which they achieve in each round. There is no maximum number of rounds. Brute force attacks will be prevented using the scoring rules below.

4.2 Experiment

The experiment will use the data collected during the user study to validate the hypothesis. The hope is that the extended tag set will produce an easier video HIP for humans. To determine if a user has passed the HIP, a score is computed using their responses. The score for round i can be computed as follows:

$$s_i = \begin{cases} w_{AST} & \text{if } RES(v) \in AST(v) \\ w_{ASL} & \text{if } RES(v) \in ASL(v) \\ w_{STM} & \text{if } RES(v) \in STM(ASL(v)) \\ w_{SYN} & \text{if } RES(v) \in SYN(ASL(v)) \\ w_{NV} & \text{if } RES(v) \in AST(N(v)) \\ 0.0 & \text{else} \end{cases}$$

The tag database for the control will consist only of author-supplied tags. The user will be required to exactly match their tag to one of the author supplied tags in a single round. Since only 1 round exists in this method (there are no partial points and therefore only 1 round), the pass/fail can be determined if they scored a 1.0 in any round. More formally, the control uses the following assignments:

$$|S| = 1; w_{AST} = 1.0; w_{ASL} = 0; w_{STM} = 0; w_{SYN} = 0; w_{NV} = 0.$$

The database of ground truth tags will then be extended to include more than just the author supplied tags. As mentioned above, the database will also include tags of neighboring videos, alternate forms of the author supplied tags, and words from the author supplied title. These methods will attempt to be more flexible to human responses.

The above weights (w_x) will be determine after the user study by searching the space to find the appropriate values. However, they must satisfy the requirement that $w_x \leq 1.0$. An estimated interval for the values is: $0.5 \leq w_x < 1.0$. Informally, this means that a user will pass the HIP if they guess appropriate tags for two consecutive videos ($|S| = 2$). To pass the HIP, users will need to achieve a score of 1.0 or better as a sum over the previous 2 rounds. For example, if during the first round the user scores a 0.6 and during the second round the user scores a 0.5, they are considered to have passed. However, a user with $S = \{0.6, 0.3\}$ will not pass because the score from two consecutive rounds do not sum to ≥ 1.0 . Intuitively, this prevents brute force attacks. More formally, the pass/fail can be defined as:

$$\begin{cases} PASS & \text{if } \exists s_i s_{i+1} \in S \text{ such that } s_i + s_{i+1} \geq 1.0 \\ FAIL & \text{else} \end{cases}$$

The tag relationships can easily be modeled using a weighted undirected graph. Videos and tags correspond to nodes, while the scores correspond to the weights of the edges between the nodes. Determining

the score for a given round would consist of a simple *breadth-first search* (BFS) of the graph starting at the video.

4.3 Details of User Study

A user study including at least 30 participants will be held to collect challenge responses of 30-50 videos. The results of the user study are extremely important and will ultimately help validate or invalidate the hypothesis. Participants will be unpaid volunteers who are willing to contribute 30 minutes of their time. The experiments will be conducted on campus in a controlled environment. The user study will measure a number of important values:

- The participant’s demographics (age, gender, educational background, etc).
- Time required to submit answers for each challenge.
- Answer submitted for each challenge.
- A subjective rating of how easy each challenge was.

The users will not be provided with the pass/fail decision from the HIP. The pass/fail decision will be computed in an offline, post-processing fashion using the different sources of tags. This will allow for early data collection and the offline validation of the hypothesis.

4.4 Proposed Attack Vector

The user study will only be able to determine if the extended method is easier for humans. A task which is trivial for humans and also trivial for machines is not a valuable HIP. Due to time limitations, a single attack will be attempted: a tag-based frequency attack.

The tag-based frequency attack would require the attacker to mine the underlying tag database and use the most frequently occurring tag. This would guarantee a relatively high success rate as compared to other potential tags. To simulate such an attack, a large sampling of video/tag sets will be gathered [37]. The data will be gathered by using the publicly available YouTube API. Preliminary investigation shows that generating a truly random sampling of videos is not possible due to limitations in the API [60]. Large social networks are typically sampling using the *snowball sampling* method [41] where new entities are discovered by traversing relationships of existing entities. This method severely biases well-connected entities. To mitigate this, multiple starting points into the network will be used. The frequency-based attack will be parameterized by a probability distribution on tags and the results will be presented using the reverse ROC method.

Care will be made in the design of the system to reduce the success of such an attack. Using the psuedo-random sampling of videos described above, frequency analysis of tags will be performed and used to prune frequently occurring tags out of the database of ground truth tags. While this has the potential to reduce usability, tags which frequently occur will often be generalizations which carry very little descriptive information. Other modes of attack are possible, but will not be considered (see limitations section below).

5 Limitations and Key Assumptions

5.1 Limitations

This research will not explore all possible attack vectors on the video HIP. While it is relatively easy to prove that a specific attack was unsuccessful against the video HIP (for example, a frequency based attack), it is impossible to show that an implementation is secure against all possible attacks. The author acknowledges that computer vision and content based video retrieval attacks could be a successful attack vector. While it is out of the scope of this research to attempt attacking the video HIP in such a way, the author gladly welcomes future attacks by other researchers.

Very few HIPs are capable of being accessible by users with perceptual disabilities (see section on Text based HIPs above). Image-based HIPs are inaccessible by blind users. Audio-based HIPs are inaccessible by

deaf users. Because video-based HIPs provided both visual and auditory cues, a video HIP has the potential to be more accessible than existing methods.

5.2 Assumptions

It is assumed that it is impossible and impractical to replicate the entire database of online videos. The amount of storage and bandwidth required to perform such a task would be prohibitively expensive. Online video databases continue to grow at an astronomical rate (for example, it was reported in May 2006 that over 60,000 new videos are uploaded to YouTube every day).

Furthermore, it is assumed that an actual implementation of a video HIP would proxy the videos in some fashion. Streaming the content directly from the video content providers to the user exposes the source of the video stream. An attacker would easily be able to grab the author supplied tags from the original page. Proxying the video would require the HIP service to cache videos from many popular video content providers. When a HIP challenge is requested, the video would be streamed from the HIP service instead of from the video content provider in order to mask the original source.

6 Contributions

The first major contribution will be the concept of a video HIP, as video has not yet been explored as a presentation format for HIPs. An example of a video HIP will be prototyped and evaluated. The vulnerability of the system against a frequency based database attack will be assessed and documented. The results of the user study will determine whether users are able to successfully pass the video HIP.

Regardless of the outcomes of the user study and the attempts at attacking the video HIP, a knowledge contribution will be made:

	Easy for Machines	Hard for Machines
Easy for Humans	Minor contribution	Significant contribution
Hard for Humans	Minor contribution	Minor contribution

If the video HIP is resilient against attack and easy for humans to solve, the contribution will be a secure and user friendly HIP. All other outcomes will result in a minor contribution, which demonstrate the limitations of the video HIP. Regardless of the outcomes of the user study and attack attempt, several minor contributions will be made.

As described in the above section, a HIP has the following desirable properties: Automation, Open, Usability, and Security. The first two properties are clearly met. The second two properties will be evaluated by the user study and database attack attempt.

The following table relates the results of the user study and the attack experiment to the overall outcome. Notation: U_C is the usability (human success rate) for the control, U_E is the usability (human success rate) for the extended method, S_C is the security (attack success rate) for the control, and S_E is the security (attack success rate) for the extended method.

	$U_C \geq U_E$	$U_C < U_E$
$S_C \leq S_E$	Negative Result	Positive Result
$S_C > S_E$	Negative Result	Negative Result

7 Proposed Chapters and Timeline

7.1 Introduction

The general problem area, the specific problem, why the topic is important, research approach, limitations and key assumptions, and contribution to be made by the research are described. Thesis statement is presented here.

7.2 Background

A complete survey of theory base, motivations, and prior research of HIPs.

7.3 Methodology

Describe how the hypothesis was tested (experiments, data collection techniques, measurement techniques, participant selection, participant demographics, etc).

7.4 Results

Present the data from the experiments. Analyze and explain the results and draw conclusions in a discussion section.

7.5 Conclusion

Summarize the thesis and with an emphasis on the results and contributions. Restate the hypothesis and validate the hypothesis with the results. Make suggestions for future work.

7.6 Timeline (Tentative)

The writing process is an on-going process and will occur in parallel with the major milestones listed below.

Date	Item
May 1, 2008	Thesis proposal completed and submitted to reader for approval.
June 1, 2008	Begin data collection from experiments.
July 1, 2008	Run, collect, and analyze results.
August 8, 2008	Final defense.

References

- [1] Edward Aboufadel, Julia Olsen, and Jesse Windle. Breaking the holiday inn priority club captcha. *College Mathematics Journal*, 36:101–108, March 2005.
- [2] Elias Athanasopoulos and Spiros Antonatos. Enhanced captchas: Using animation to tell humans and computers apart. In *Communications and Multimedia Security, 10th IFIP TC-6 TC-11 International Conference, CMS 2006, Proceedings*, number 4237 in Lecture Notes in Computer Science, pages 97–108, Heraklion, Crete, Greece, October 2006. Springer.
- [3] Henry S. Baird. Document image defect models and their uses. In *Document Analysis and Recognition, 1993., Proceedings of the Second International Conference on*, pages 62–67, Tsukuba Science City, Japan, October 1993.
- [4] Henry S. Baird and Jon Louis Bentley. Implicit captchas. In Elisa H. Barney Smith and Kazem Taghva, editors, *Proceedings of the IST SPIE Document Recognition and Retrieval XII Conference*, volume 5676, San Jose, CA, USA, January 2005.
- [5] Henry S. Baird, Allison L. Coates, and Richard J. Fateman. Pessimprint: A reverse turing test. *International Journal of Document Analysis and Recognition*, 5(2-3):158–163, April 2003.
- [6] Henry S. Baird and Daniel P. Lopresti, editors. *Human Interactive Proofs, Second International Workshop*, volume 3517 of *Lecture Notes in Computer Science*, Bethlehem, PA, USA, May 19-20 2005. Springer.

- [7] Henry S. Baird and Mark Luk. Protecting websites with reading-based captcha. In *2nd International Web Document Analysis Workshop (WDA'03)*, Edinburgh, Scotland, August 2003.
- [8] Henry S. Baird, Michael A. Moll, and Sui-Yu Wang. A highly legible captcha that resists segmentation attacks. In Baird and Lopresti [6], pages 27–41.
- [9] Henry S. Baird, Michael A. Moll, and Sui-Yu Wang. Scattertype: A legible but hard-to-segment captcha. In *ICDAR '05: Proceedings of the Eighth International Conference on Document Analysis and Recognition*, pages 935–939, Washington, DC, USA, August 2005. IEEE Computer Society.
- [10] Henry S. Baird and Kris Popat. Human interactive proofs and document image analysis. In *Proceedings of the 5th International Workshop on Document Analysis Systems*, volume LNCS 2423, pages 507–518, Princeton, NJ, August 2002. Springer-Verlag (Berlin).
- [11] Henry S. Baird and Terry Riopka. Scattertype: A reading captcha resistant to segmentation attack. In Elisa H. Barney Smith and Kazem Taghva, editors, *Proceedings of the IST SPIE Document Recognition and Retrieval XII Conference*, volume 5676, pages 197–207, San Jose, CA, USA, January 2005.
- [12] Henry S. Baird, Sui-Yu Wang, and Jon Louis Bentley. Captcha challenge tradeoffs: Familiarity of strings versus degradation of images. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 3, pages 164–167, 2006.
- [13] Jon Louis Bentley and Colin Mallows. Captcha challenge strings: Problems and improvements. In Kazem Taghva and Xiaofan Lin, editors, *IST SPIE Document Recognition and Retrieval XIII Conference*, volume 6067. SPIE, January 2006.
- [14] Richard Bergmair and Stefan Katzenbeisser. Towards human interactive proofs in the text-domain (using the problem of sense-ambiguity for security). In *Proceedings of the 7th International Conference, ISC 2004*, pages 257–267, Palo Alto, CA, USA, September 2004.
- [15] Jeffrey P. Bigham, Maxwell B. Aller, Jeremy T. Brudvik, Jessica O. Leung, Lindsay A. Yazzolino, and Richard E. Ladner. Inspiring blind high school students to pursue computer science with instant messaging chatbots. In *SIGCSE '08: Proceedings of the 39th SIGCSE technical symposium on Computer science education*, pages 449–453, New York, NY, USA, March 2008. ACM.
- [16] Manuel Blum and Henry S. Baird, editors. *Human Interactive Proofs, First International Workshop*, Palo Alto, CA, January 2002. Xerox Palo Alto Research Center.
- [17] Gavin Brelstaff and Francesca Chessa. Practical application of visual illusions: errare humanum est. In *APGV '05: Proceedings of the 2nd symposium on Applied perception in graphics and visualization*, pages 161–161, New York, NY, USA, August 2005. ACM.
- [18] Nancy Chan. Abstract of sound oriented captcha. In Blum and Baird [16], page 35.
- [19] Tsz-Yan Chan. Using a text-to-speech synthesizer to generate a reverse turing test. In *Tools with Artificial Intelligence, 2003. Proceedings. 15th IEEE International Conference on*, pages 226–232, Los Alamitos, CA, USA, November 2003. IEEE Computer Society.
- [20] Kumar Chellapilla, Kevin Larson, Patrice Y. Simard, and Mary Czerwinski. Building segmentation based human-friendly human interaction proofs (hips). In Baird and Lopresti [6], pages 1–26.
- [21] Kumar Chellapilla, Kevin Larson, Patrice Y. Simard, and Mary Czerwinski. Computers beat humans at single character recognition in reading based human interaction proofs (hips). In *In Proceedings of the Second Conference on Email and Anti-Spam (CEAS)*, Palo Alto, CA, July 2005.
- [22] Kumar Chellapilla, Kevin Larson, Patrice Y. Simard, and Mary Czerwinski. Designing human friendly human interaction proofs (hips). In *CHI '05: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 711–720, New York, NY, USA, April 2005. ACM.

- [23] Kumar Chellapilla and Patrice Y. Simard. Using machine learning to break visual human interaction proofs (HIPs). In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 17 (NIPS 2004)*, pages 265–272, Cambridge, MA, December 2004. MIT Press.
- [24] Casey Chesnut. Using AI to beat CAPTCHA and post comment spam. Online <http://www.brains-n-brawn.com/aiCaptcha>, January 2005.
- [25] Monica Chew and Henry S. Baird. Baffletext: A human interactive proof. In *IST/SPIE Document Recognition and Retrieval X Conference*, pages 305–316, January 2003.
- [26] Monica Chew and J. Doug Tygar. Image recognition captchas. In Kan Zhang and Yuliang Zheng, editors, *In Proceedings of the 7th International Information Security Conference (ISC 2004)*, volume 3225 of *Lecture Notes in Computer Science*, pages 268–279, Palo Alto, CA, September 2004. Springer Berlin / Heidelberg.
- [27] Monica Chew and J. Doug Tygar. Image recognition captchas. Technical Report UCB/CSD-04-1333, EECS Department, University of California, Berkeley, August 2004.
- [28] Monica Chew and J. Doug Tygar. Collaborative filtering captchas. In Baird and Lopresti [6], pages 66–81.
- [29] Allison L. Coates, Henry S. Baird, and Richard J. Fateman. Pessimist print: A reverse turing test. In *Proceedings, IAPR 6th Int'l Conf. on Document Analysis and Recognition*, pages 1154–1158, Seattle, WA, September 2001. IEEE Computer Society.
- [30] Tim Converse. Captcha generation as a web service. In Baird and Lopresti [6], pages 82–96.
- [31] Bruno Norberto da Silva and Ana Cristina Bicharra Garcia. A hybrid method for image taxonomy: Using captcha for collaborative knowledge acquisition. In *Proceedings of the AAAI 2006 Fall Symposium on Semantic Web for Collaborative Knowledge Acquisition*, pages 17–23, Arlington, VA, October 2006.
- [32] M Dailey and C. Namprempe. A text graphics character captcha for password authentication. In *TENCON 2004. 2004 IEEE Region 10 Conference*, volume 2B, pages 45–48, November 2004.
- [33] Ritendra Datta, Jia Li, and James Z. Wang. Imagination: a robust image-based captcha generation system. In *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, pages 331–334, New York, NY, USA, November 2005. ACM.
- [34] Rachna Dhamija and J. Doug Tygar. Phish and hips: Human interactive proofs to detect phishing attacks. In Baird and Lopresti [6], pages 127–141.
- [35] John Douceur, Jeremy Elson, Jon Howell, and Jared Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. In *CCS '07: Proceedings of the 14th ACM Conference on Computer and Communications Security*, pages 366–374, New York, NY, USA, October 2007. ACM.
- [36] Igor Fischer and Thorsten Herfet. Visual captchas for document authentication. In *Multimedia Signal Processing, 2006 IEEE 8th Workshop on*, pages 471–474, October 2006.
- [37] Gary Geisler and Sam Burns. Tagging video: conventions and strategies of the youtube community. In *JCDL '07: Proceedings of the 7th ACM/IEEE joint conference on Digital libraries*, pages 480–480, New York, NY, USA, June 2007. ACM.
- [38] Philip Brighten Godfrey. Natural language captchas. Preliminary Manuscript, April 2002.
- [39] Philip Brighten Godfrey. Text-based captcha algorithms. In Blum and Baird [16].

- [40] Joshua T. Goodman and Robert Rounthwaite. Stopping outgoing spam. In *Electronic Commerce '04: Proceedings of the 5th ACM conference on Electronic commerce*, pages 30–39, New York, NY, USA, May 2004. ACM.
- [41] Leo A. Goodman. Snowball sampling. *The Annals of Mathematical Statistics*, 32(1):148–170, March 1961.
- [42] Jonathan Holman, Jonathan Lazar, Jinjuan Heidi Feng, and John D’Arcy. Developing usable captchas for blind users. In *Assets '07: Proceedings of the 9th international ACM SIGACCESS conference on Computers and accessibility*, pages 245–246, New York, NY, USA, October 2007. ACM.
- [43] Mohammed E. Hoque, David J. Russomanno, and Mohammed Yeasin. 2d captchas from 3d models. In *SoutheastCon, 2006. Proceedings of the IEEE*, pages 165–170, April 2006.
- [44] Srikanth Kandula, Dina Katabi, Matthias Jacob, and Arthur W. Berger. Botz-4-sale: Surviving organized ddos attacks that mimic flash crowds. In *2nd Symposium on Networked Systems Design and Implementation (NSDI)*, Boston, MA, May 2005.
- [45] Auguste Kerckhoffs. La cryptographie militaire. *Journal des sciences militaires*, 9:161–191, January 1883.
- [46] Greg Kochanski, Daniel P. Lopresti, and Chilin Shih. A reverse turing test using speech. In *Proceedings of ICSLP2002, Seventh International Conference on Spoken Language Processing*, pages 1357–1360, Denver, Colorado, September 2002.
- [47] Henry Kucera and W. Nelson Francis. Computational analysis of present-day american english. *International Journal of American Linguistics*, 35(1):71–75, January 1969.
- [48] Joon Ho Lee. Properties of extended boolean models in information retrieval. In *SIGIR '94: Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 182–190, New York, NY, USA, July 1994. Springer-Verlag New York, Inc.
- [49] Wen-Hung Liao and Chi-Chih Chang. Embedding information within dynamic visual patterns. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 2, pages 895–898, June 2004.
- [50] Daniel P. Lopresti. Leveraging the captcha problem. In Baird and Lopresti [6], pages 97–110.
- [51] Daniel P. Lopresti, C. Shih, and G. Kochanski. Human interactive proofs for spoken language interfaces. In Blum and Baird [16], pages 30–34.
- [52] Julie Beth Lovins. Development of a stemming algorithm. *Mechanical Translation and Computational Linguistics*, 11:22–31, March 1968.
- [53] Matt May. Inaccessibility of CAPTCHA. <http://www.w3.org/TR/turingtest/>, November 2005.
- [54] Deapesh Misra and Kris Gaj. Face recognition captchas. In *Telecommunications, 2006. AICT-ICIW '06. International Conference on Internet and Web Applications and Services/Advanced International Conference on*, page 122, Washington, DC, USA, February 2006. IEEE Computer Society.
- [55] William G. Morein, Angelos Stavrou, Debra L. Cook, Angelos D. Keromytis, Vishal Misra, and Dan Rubenstein. Using graphic turing tests to counter automated ddos attacks against web servers. In *Computer and Communications Security '03: Proceedings of the 10th ACM conference on Computer and communications security*, pages 8–19, New York, NY, USA, October 2003. ACM.
- [56] Greg Mori and Jitendra Malik. Recognizing objects in adversarial clutter: breaking a visual captcha. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages 134–141, June 2003.

- [57] Gabriel Moy, Nathan Jones, Curt Harkless, and Randall Potter. Distortion estimation techniques in solving visual captchas. In *Conference on Computer Vision and Pattern Recognition (CVPR'04)*, volume 02, pages 23–28, Los Alamitos, CA, USA, June 2004. IEEE Computer Society.
- [58] George L. Nagy, Stephen V. Rice, and Thomas A. Nartker. *Optical Character Recognition: An Illustrated Guide to the Frontier*. Kluwer Academic Publishers, Norwell, Massachusetts, USA, May 1999.
- [59] Moni Naor. Verification of a human in the loop or identification via the turing test. Unpublished manuscript, Sept 1996.
- [60] John C. Paolillo. Structure and network in the youtube core. In *HICSS '08: Proceedings of the Proceedings of the 41st Annual Hawaii International Conference on System Sciences*, pages 156–166, Washington, DC, USA, January 2008. IEEE Computer Society.
- [61] Linda Dailey Paulson. Blind, deaf engineer develops computerized braille machine. *Computer*, 35(12):27–27, December 2002.
- [62] Benny Pinkas and Tomas Sander. Securing passwords against dictionary attacks. In *CCS '02: Proceedings of the 9th ACM conference on Computer and communications security*, pages 161–170, New York, NY, USA, November 2002. ACM.
- [63] Clark Pope and Khushpreet Kaur. Is it human or computer? defending e-commerce with captchas. *IT Professional*, 7(2):43–49, March 2005.
- [64] Martin F. Porter. The english (porter2) stemming algorithm. Online <http://snowball.tartarus.org/algorithms/english/stemmer.html>.
- [65] Martin F. Porter. An algorithm for suffix stripping. *Program*, 14(3):130–137, July 1980.
- [66] Bartosz Przydatek. On the (im)possibility of a text-only captcha. In Blum and Baird [16].
- [67] Sara Robinson. Up to the challenge: Computer scientists crack a set of ai-based puzzles. *SIAM News*, 35(9):23–24, November 2002.
- [68] Yong Rui and Zicheng Liu. Artificial: Automated reverse turing test using facial features. In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 295–298, New York, NY, USA, November 2003. ACM.
- [69] Yong Rui and Zicheng Liu. Excuse me, but are you human? In *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pages 462–463, New York, NY, USA, November 2003. ACM.
- [70] Yong Rui and Zicheng Liu. Artificial: Automated reverse turing test using facial features. *Multimedia Systems*, 9(6):493–502, June 2004.
- [71] Yong Rui, Zicheng Liu, Shannon Kallin, Gavin Janke, and Cem Paya. Characters or faces: A user study on ease of use for hips. In Baird and Lopresti [6], pages 53–65.
- [72] Amalia Rusu and Venu Govindaraju. Handwriting word recognition: A new captcha challenge. In *Proceedings of the Fifth International Conference on Knowledge Based Computer Systems*, pages 347–357, Hyderabad, India, December 2004.
- [73] Amalia Rusu and Venu Govindaraju. Handwritten CAPTCHA: using the difference in the abilities of humans and machines in reading handwritten words. In *Frontiers in Handwriting Recognition, 2004. Ninth International Workshop on*, pages 226–231, October 2004.

- [74] Amalia Rusu and Venu Govindaraju. Challenges that handwritten text images pose to computers and new practical applications. In Elisa H. Barney Smith and Kazem Taghva, editors, *Proceedings of the IST SPIE Document Recognition and Retrieval XII Conference*, volume 5676, pages 84–91, San Jose, CA, USA, January 2005.
- [75] Amalia Rusu and Venu Govindaraju. A human interactive proof algorithm using handwriting recognition. In *Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on*, pages 967–971 Vol. 2, Aug 2005.
- [76] Amalia Rusu and Venu Govindaraju. Visual captcha with handwritten image analysis. In Baird and Lopresti [6], pages 42–52.
- [77] Gerard Salton. The use of extended boolean logic in information retrieval. In *SIGMOD '84: Proceedings of the 1984 ACM SIGMOD international conference on Management of data*, pages 277–285, New York, NY, USA, June 1984. ACM.
- [78] Mohammad Shirali-Shahreza and M. Hassan Shirali-Shahreza. A new solution for password key transferring in steganography methods by captcha through mms technology. In *Information and Emerging Technologies, 2007. ICIET 2007. International Conference on*, pages 1–6, July 2007.
- [79] Mohammad Shirali-Shahreza and Sajad Shirali-Shahreza. Captcha for blind people. In *Signal Processing and Information Technology, 2007 IEEE International Symposium on*, pages 995–998, December 2007.
- [80] Patrice Y. Simard, Richard Szeliski, Josh Benaloh, Julien Couvreur, and Iulian Calinov. Using character recognition and segmentation to tell computers from humans. In *International Conference on Document Analysis and Recognition (ICDAR)*, volume 1, pages 418–423, Los Alamitos, CA, USA, August 2003.
- [81] Andrew Stuart and Dennis P. Phillips. Word recognition in continuous and interrupted broadband noise by young normal-hearing, older normal-hearing, and presbycusis listeners. *Ear and Hearing*, 17(6):478–489, December 1996.
- [82] Arasanathan Thayananthan, Björn Stenger, Phil H. S. Torr, and Roberto Cipolla. Shape context and chamfer matching in cluttered scenes. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages 127–133, June 2003.
- [83] Alan M. Turing. Computing Machinery and Intelligence. *Mind*, 59(236):433–460, October 1950.
- [84] Erik van Woudenberg, Frank K. Soong, and J. E. West. Acoustic echo cancellation for hands-free asr applications in noise. In *In Proceedings of the Workshop on Acoustic Echo and Noise Control*, pages 160–163, September 1999.
- [85] Luis von Ahn, Manuel Blum, Nicholas J. Hopper, and John Langford. CAPTCHA: Using Hard AI Problems for Security. In *Advances in Cryptology, Eurocrypt '03*, volume 2656 of *Lecture Notes in Computer Science*, pages 294–311, Warsaw, Poland, May 2003. Springer.
- [86] Luis von Ahn, Manuel Blum, and John Langford. Telling humans and computers apart automatically. *Communications of the ACM*, 47(2):56–60, February 2004.
- [87] Luis von Ahn, Ben Maurer, Colin McMillen, Mike Crawford, Ryan Staake, and Manuel Blum. reCAPTCHA Project. Online, <http://www.recaptcha.net>, May 2007.
- [88] Pablo Ximenes, Andre dos Santos, Marcial Fernandez, and Joaquim Celesti. A captcha in the text domain. In R. Meersman, Z. Tari, and P. Herrero, editors, *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, volume 4277/2006 of *Lecture Notes in Computer Science*, pages 605–615. Springer Berlin / Heidelberg, November 2006.

- [89] J. Xu, R. Lipton, I. Essa, M. Sung, and Y. Zhu. Mandatory human participation: a new authentication scheme for building secure systems. In *Computer Communications and Networks, 2003. ICCCN 2003. Proceedings. The 12th International Conference on*, pages 547–552, October 2003.
- [90] Jun Xu, Irfan A. Essa, and Richard J. Lipton. Hello, are you human? CC Technical Report GIT-CC-00-28, Georgia Institute of Technology, November 2000.
- [91] Jeff Yan. Bot, cyborg and automated turing test. Technical Report CS-TR-970, University of Newcastle, June 2006.
- [92] Jeff Yan and Ahmad Salah El Ahmad. Breaking visual captchas with naive pattern recognition algorithms. In *Computer Security Applications Conference, 2007. ACSAC 2007. Twenty-Third Annual*, pages 279–291, December 2007.
- [93] YouTube. Youtube glossary. Online <http://www.google.com/support/youtube/bin/answer.py?hl=en&answer=70181>.
- [94] Zhenqiu Zhang, Yong Rui, T. Huang, and C. Paya. Breaking the clock face hip [web services human interactive proofs]. In *Multimedia and Expo, 2004. ICME '04. 2004 IEEE International Conference on*, volume 3, pages 2167–2170, June 2004.