Exploiting the Human-Machine Gap in Image Recognition for Designing CAPTCHAs

Ritendra Datta, *Member, IEEE,* Jia Li, *Senior Member, IEEE,* and James Z. Wang, *Senior Member, IEEE*

***Abstract*— Security researchers have, for long, devised mechanisms to prevent adversaries from conducting automated network attacks, such as denial-of-service, which lead to significant wastage of resources. On the other hand, several attempts have been made to automatically recognize generic images, make them semantically searchable by content, annotate them, and associate them with linguistic indexes. In the course of these attempts, the limitations of state-of-the-art algorithms in mimicking human vision have become exposed. In this paper, we explore the exploitation of this limitation for potentially preventing automated network attacks. While undistorted natural images have been shown to be algorithmically recognizable and searchable by content to moderate levels, controlled distortions of specific type and strength can potentially make machine recognition harder without affecting human recognition. This difference in recognizability makes it a promising candidate for automated Turing tests called CAPTCHAs which can differentiate humans from machines. We empirically study the application of controlled distortions of varying nature and strength, and their effect on human and machine recognizability. While human recognizability is measured on the basis of an extensive user study, machine recognizability is based on memory-based content-based image retrieval (CBIR) and matching algorithms. We give a detailed description of our experimental image CAPTCHA system, IMAGINATION, that uses systematic distortions at its core. A significant research topic within signal analysis, CBIR is actually conceived here as a tool for an adversary, so as to help us design more foolproof image CAPTCHAs.**

***Index Terms*— Automated Turing tests, CAPTCHAs, Systematic network attacks, Image recognition.**

I. Introduction

Robust image understanding remains an open problem. The gap between human and computational ability to recognizing visual content has been termed by Smeulders et al. [30] as the *semantic gap*. A key area of research that would greatly benefit from the narrowing of this gap is content-based image retrieval (CBIR). Over more than a decade, attempts have been made to build tools and systems that can retrieve images (from repositories) that are semantically similar to query images, which have enjoyed moderate success [7], [30]. While the inability to bridge the semantic gap highlights the limitations of the state-of-the-art in image content analysis, we see in it an opportunity for *system security*. This, and any task that humans are better at performing than the best computational means, can be treated as an 'automated Turing test' [1], [32] that tells humans and computers apart. Typically referred to as HIP (Human Interactive Proof) or CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) [3], they help reduce e-mail spam, stop automated blog and forum responses, save resources, and prevent denial-of-service (DoS) attacks on Web servers [23], among others. In general, DoS attacks involve generating a large number of automated (machine) requests to one or more network devices (e.g., servers) for resources in some form, with the goal of overwhelming them and preventing legitimate (human) users from getting their service. In a distributed DoS, multiple machines are compromised and used for coordinated automated attacks, making it hard to detect and block the attack sources. To prevent such forms of attacks and save resources, the servers or other network devices can require CAPTCHA solutions to accompany each request, thus forcing human intervention, and hence, in the very least, reducing the intensity of the attacks. Because CAPTCHAs can potentially play a very critical role in system security, it is imperative that the design and implementation of CAPTCHAs be relatively foolproof.

There has been sizable research output in designing as well as *breaking* CAPTCHAs. In both these efforts, computing research stands to benefit. A better CAPTCHA design means greater security for computing systems, and the breaking of an existing CAPTCHA usually means the advancement of artificial intelligence (AI). While text-based CAPTCHAs have been traditionally used in real-world applications (Yahoo! Mail Sign up, PayPal Sign up, Ticketmaster search, Blogger Comment posting, etc.), their vulnerability has been repeatedly shown by computer vision researchers [24], [31], [4], [25], reporting over 90% success rate. Among the earliest commercial ones, the Yahoo! CAPTCHA has also been reportedly compromised, with a success rate of 35% [29], allowing e-mail accounts to be opened automatically, and encouraging e-mail spam.

In principle, there exist many hard AI problems that can replace text-based CAPTCHAs, but in order to have general appeal and accessibility, recognition of image content has been an oft-suggested alternative [1], [5], [8], [9], [27]. While automatic image recognition is usually considered to be a much harder problem than text recognition (which is a reason for it to be suggested as an alternative to text CAPTCHAs), it has also enjoyed moderate success as part of computer vision research. This implies that a straightforward replacement of

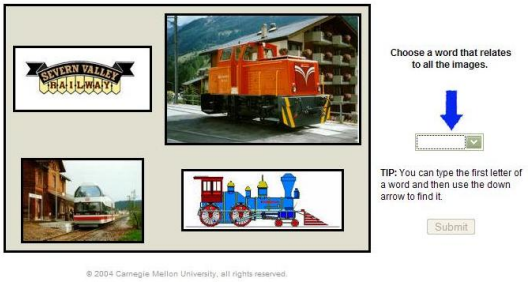
R. Datta is with the Department of Computer Science and Engineering, The Pennsylvania State University, University Park, PA 16802, USA. Phone: (814) 865-6168. E-mail: datta@cse.psu.edu. Fax: (814) 865-6426.

J. Li is with the Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA. Phone: (814) 863-3074. E-mail: jiali@stat.psu.edu. Fax: (814) 865-6426.

J. Z. Wang is with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA. Phone: (814) 865-7889. E-mail: jwang@ist.psu.edu. Fax: (814) 865-6426.

(a) Yahoo's EZ-Gimpy

(b) PayPal's CAPTCHA



(c) CMU's ESP-PIX

(d) Microsoft's Asirra

Fig. 1. Sample CAPTCHAs proposed or in real-world use. (a)-(b) Text-based CAPTCHAs in public use. (c) Image-based CAPTCHA proposed by CMU's Captcha Project. User is asked to choose an appropriate label from a list. (d) Asirra presents pictures of cats and dogs and asks users to select all the cats.

text with images may subject it to similar risks of being 'broken' by image recognition techniques. Techniques such as near-duplicate image matching [15], content-based image retrieval [30], and real-time automatic image annotation [19] are all potential attack tools for an adversary. One approach that can potentially make it harder for automated attack while maintaining recognizability by humans is *systematic distortion*. A brief mention of the use of distortions in the context of image CAPTCHAs has been made in the literature [5], but this has not been followed up by any study or implementation. Furthermore, while there have been ample studies on the algorithmic ability to handle noisy signals (occlusion, low light, clutter, noise), most often to test robustness of recognition methods, their behavior under strong artificial distortions has been rarely studied systematically.

In this work, we explore the use of systematic image distortion in designing CAPTCHAs, for inclusion in our experimental system called IMAGINATION. We compare human and machine recognizability of images under distortion based on extensive user studies and image matching algorithms respectively. The criteria for a distortion to be eligible for image CAPTCHA design are that when applied, they

1) make it difficult for algorithmic recognition, and
2) have minor effect on recognizability by humans.

Formally, let $\mathcal{H}$ denote a representative set of humans, and let $\mathcal{M}$ denote one particular algorithm of demonstrated image recognition capability. We introduce a *recognizability function* $\rho_X(I)$ to indicate whether image I has been correctly recognized by X or not. Thus, $\rho_{\mathcal{H}}(I)$ and $\rho_{\mathcal{M}}(I)$ are human and machine recognizabilities respectively, and we refer to $|\rho_{\mathcal{H}}(I) - \rho_{\mathcal{M}}(I)|$ as the *recognizability gap* with respect to image I. This image can be visually distorted to varying degrees. We define a distortion function $\delta_y(\cdot)$ that can be applied to a natural image, the degree of distortion being abstractly represented by parameter y. This study focuses on analyzing (a) recognizability, and (b) recognizability gap, of distorted images $\delta_y(I)$, over a large number of natural images. The following are of interest:

- Current state-of-the-art in image recognition typically test and report results on undistorted natural images, and on minor distortions. The 'breaking' of an image CAPTCHA, in the absence of distortion, is therefore roughly as likely as the performance of these image recognition techniques.
- On application of a distortion, the image recognition performance is expected to degrade. There has been no comprehensive study on the effect of various artificial distortions on image recognizability.
- Distortion also affects human recognizability of images. It is safe to assume, though, that humans are relatively more resilient to distortion; they can 'see through' clutter and fill in the missing pieces, owing to their power of imagination.
- In CAPTCHA design, the goal is to evade recognition by machines while being easily recognizable by humans. It is therefore important to be able to figure out the types and strengths of distortion on images that keep human recognizability high while significantly affecting machine recognizability.

While the primary aim of this work is the systematic design of a security mechanism, the results from the study (See, e.g., Figs. 7, 8, 9, 10, and 11) also reveal to us some of the shortcomings of image matching algorithms, i.e., how the application of certain distortions makes it difficult for even state-of-the-art image matching methods to pair up distorted images with their originals. Furthermore, through large-scale user studies, we are also made aware of the kinds of distortions that make image recognition difficult for humans. These peripheral observations may find use in other research domains.

A Note on CAPTCHA-based Security: Besides specific attempts to break CAPTCHAs by solving hard AI problems, in the recent times, adversaries have used a method which greatly undermines their strength: using humans to solve them. As reported recently [13], humans are being used to solve

them, either in a well-organized manner commercially (in low labor cost regions), or by the use of games and other methods whereby humans are unaware that their responses are being used for malicious purposes. These attempts make it futile to create harder AI problems, because in principle, a CAPTCHA should be solvable by virtually all humans, regardless of their intent. Nonetheless, CAPTCHAs are and will continue to remain deployed until alternate, unbreakable, human identity verification methods become practical. Till then, they should, at the very least, serve to impede the intensity of human-guided breaking of CAPTCHAs. Our work continues the mission of designing CAPTCHAs resilient to *automated* attacks. A key strategy involved in preventing automated attacks is to incorporate random distortions as much as possible, effectively making the space of CAPTCHA problems infinite, thus rendering any attempt to build a dictionary of answers infeasible. Our approach is based on such a strategy.

The rest of this paper is arranged as follows. In Sec. II, we discuss the metrics for measurement of recognizability under distortion for both humans and machines, and potential candidate distortions that can affect recognizability. In Sec. III, we describe our experimental system IMAGINATION, which we then compare comprehensively with existing CAPTCHAs, in Sec. IV. Experimental results on the effect of distortions on human and machine recognizability are presented in Sec. V. We conclude in Sec. VI.

II. Image Recognizability Under Distortion

Let us assume that we have a collection of natural images, each with a dominant subject, such that given a set of options (say 15), choosing a label is unambiguous. We first define machine/human recognizability concretely, and then discuss distortions that can potentially satisfy the CAPTCHA requirements.

A. Algorithmic Recognizability

Algorithms that attempt to perform image recognition under distortion can be viewed from two different angles here. First, they can be thought of as methods that potential *adversaries* may employ in order to break image CAPTCHAs. Second, they can be considered as intelligent vision systems. Because the images in question can be widely varying and be part of a large image repository, content-based image retrieval (CBIR) systems [30] seem apt. Essentially a memory-based method of attack, the assumption is that the adversary has access to the original (undistorted) images (which happens to be a requirement [3] of CAPTCHAs) for matching with the distorted image presented. While our experiments focus on image matching algorithms, other types of algorithms also seem plausible attack strategies. *Near-duplicate detection* [15], which focus on finding marginally modified/distorted copyrighted images, seems to be a potential choices as well. This is part of our future work. *Automatic image annotation* and *scene recognition* techniques [7] have potential, but given the current state-of-the-art, these methods are unlikely to do better than direct image-to-image matching.

Recognition of a distorted image $\delta_y(I)$ is thus achieved as follows: Let the adversary have at hand the entire database $\mathcal{X}$ of possible images, i.e., $\forall I,\, I \in \mathcal{X}$. We can think of the image retrieval algorithm as a function that takes in a pair of images and produces a distance measure $g(I_1, I_2)$ (which hopefully correlates well with their semantic distance). Define a rank function

$$rank_g(I_1, I_2, \mathcal{X}) = \text{Rank of } I_1 \text{ w.r.t. } I_2 \text{ in } \mathcal{X} \text{ using } g(\cdot,\cdot) \tag{1}$$

We relax the criteria for machine recognizability, treating image I_1 as recognizable if $rank_g(I_1, I_2, \mathcal{X})$ is within the top K ranks. This is done since the adversary, being a machine, can iterate over a small set K of images quickly to produce a successful attack. Thus, we define *average machine recognizability* under distortion $\delta_y(\cdot)$, where machine in this case is an image retrieval system modeled as $g(\cdot,\cdot)$, as

$$\overline{\rho_g}(\delta_y) = \frac{1}{|\mathcal{X}|} \sum_{I \in \mathcal{X}} \mathcal{I}\big(rank_g(I, \delta_y(I), \mathcal{X}) \le K\big) \tag{2}$$

where $\mathcal{I}(\cdot)$ is the indicator function. For our experiments, we consider a very simple image similarity metric, and two well-known and widely used image retrieval systems that use different low-level image representation and compute pairwise image distance in different ways. First, we use the simplest possible image similarity metric; the average of the norm of the pixel-wise difference (PWD) between the two images. Given two images, the larger image is first scaled to the smaller one to match its dimensions. If the two images are I and I', then

$$pwd(I, I') = \frac{1}{|I|} \sum_{x,y} \sum_{c \in \{R,G,B\}} \big(I_c(x,y) - I'_c(x,y)\big)^2 \tag{3}$$

where $|I|$ here denotes the total number of pixels in the image. This measure clearly lacks robustness, and is expected to be sensitive even to very small distortions. Second, we employ the Earth Mover's Distance (EMD) [26] (which is essentially the half-century old Kantorovich Distance, also known as the Mallows Distance [21]) based on global color features and a robust, true distance metric. Finally, we employ the more recent IRM distance which forms the backbone of the SIMPLIcity system [34]. This distance performs region segmentation and takes into consideration color, texture, and shape of regions, going on to compute a robust distance between a variable number of region descriptors across a pair of images. In these two cases, color similarity is computed in the CIE-LAB and CIE-LUV spaces respectively, thus adding to their robustness to chromatic distortions. Both methods, while being fairly distinct, have been independently shown to yield good retrieval performance under distortion. The generic distance function $g(\cdot,\cdot)$ is specifically denoted here as $pwd(\cdot,\cdot)$, $emd(\cdot,\cdot)$, and $irm(\cdot,\cdot)$ respectively. Thus, under distortion $\delta_y(\cdot)$, we denote their average recognizability by $\overline{\rho_{pwd}}(\delta_y)$, $\overline{\rho_{emd}}(\delta_y)$, and $\overline{\rho_{irm}}(\delta_y)$ respectively.

B. Human Recognizability

We measure human recognizability under distortion using a controlled user study. An image I is sampled from $\mathcal{X}$, sub-

jected to distortion $\delta_y(\cdot)$, and then presented to a user, along with a set of 15 word choices, one of which is unambiguously an appropriate label. While higher than 15 choices makes it harder to solve automatically, too many choices also makes it more challenging for humans and hence affects usability. The user choice, made from the word list, is recorded alongside the particular image category and distortion type. Since it is difficult to get user responses for each distortion type over all images $\mathcal{X}$, we measure the average recognizability for a given distortion using the following. If $\mathcal{U}(\delta_y)$ is the set of all images presented to users subjected to $\delta_y(\cdot)$,

$$\overline{\rho_{\mathcal{H}}}(\delta_y) = \frac{1}{|\mathcal{U}(\delta_y)|} \sum_{I \in \mathcal{U}(\delta_y)} \mathcal{I}\big(I \text{ is correctly recognized}\big) \quad (4)$$

where $\mathcal{I}$ is the indicator function. The implicit assumptions made here, under which the term $\overline{\rho_{\mathcal{H}}}(\delta_y)$ is comparable to $\overline{\rho_{emd}}(\delta_y)$ or $\overline{\rho_{irm}}(\delta_y)$ is that (a) all users independently assess recognizability of a distorted image (since they are presented privately, one at a time), and (b) with sufficient, but not necessarily identical number of responses, the average recognizability measures converge to their true value.

Assessing Recognizability with User Study: The user study we use in order to measure what we term as the average human recognizability $\overline{\rho_{\mathcal{H}}}(\delta_y)$ under distortion δ_y, is only one of many ways to assess the ability of humans to recognize images in clutter. This metric is designed specifically to assess the usability of CAPTCHAs, and may not reflect on general human vision. Furthermore, the study simply asks users to choose one appropriate image label from a list of 15 words, and recognizability is measured as the fraction of times the various users made the correct choice. While correct selection may mean that the user recognized the object in the image correctly, it could also mean that it was the only choice perceived to be correct, by elimination of choices (i.e., best among many poor matches), or even a random draw from a reduced set of potential matches. Furthermore, using the averaged responses over multiple users could mean that the CAPTCHA may still be unusable by some fraction of the population. While it is very difficult to assess true recognizability, our metric serves the purpose it is used for: the ability of users to pick one correct label from a list of choices, given a distorted image, and hence we use these averaged values in the CAPTCHA design. Furthermore, the user study consists of roughly the same number of responses from over 250 random users, making the average recognizability metric fairly representative. Later in Sec. V, we will see that there is sufficient room for relaxing the intensity of distortions so as to ensure high recognizability for most users, without compromising on security.

C. Candidate Distortions

We look at image distortion candidates that are relevant in designing image CAPTCHAs. With the exception of the requirement that the distortion should obfuscate machine vision more than human vision, the space of possible distortions $\delta_y(\cdot)$ is unlimited. Any choice of distortion gets further support if simple filtering or other pre-processing steps are ineffective in undoing the distortion. Furthermore, we avoid non-linear transformations on the images so as to retain basic shape information, which can severely affect human recognizability. For the same reason we do not use other images or templates to distort an image. Pseudo-randomly generated distortions are particularly useful here, as with text CAPTCHAs.

For the purpose of making it harder for machine recognition to undo the effect of distortion, we need to also consider the approaches taken in computer vision for this task. In the literature, the fundamental step in generic recognition tasks has been *low-level feature extraction* from the images [30], [7]. In fact, this is the only part of the recognition process that we have the power to affect. The subsequent steps typically involve deriving mid to high level features representations from them, performing pair-wise image feature matching, matching them to learned models, etc. Because of their dependence on low-level features, we expect them to weaken or fail when feature extraction is negatively affected. Some of the fundamental features and the corresponding distortions (describe below) that typically affect their extraction, are presented in Table I. For each feature, we consider only well-established extraction methodologies (e.g., SIFT [20] for interest point detection) when deciding which distortions affect them.

We formalize the notion of image distortions as follows. Suppose we have a set of fundamental or 'atomic' distortion types (denoted δ), e.g., adjustment of image luminance, quantization of colors, dithering, or addition of noise. These distortions are parameterized (parameter denoted y), so a particularly distortion is completely specified by (type, parameter) tuples, denoted δ_y. The set of possible distortions Δ, which is countably infinite if parameter y is discrete, is formalized as follows:

- Atomic distortions $\{Quantize_y(\cdot), Dither_y(\cdot), \cdots\} \in \Delta$.
- If $\delta_y(\cdot)$ and $\delta'_y(\cdot) \in \Delta$, then $\delta_y(\delta'_y(\cdot))$ and $\delta'_y(\delta_y(\cdot)) \in \Delta$.

Put in plain words, any combination of an atomic distortion (applied in a specific order) is a new distortion. Here, we list the atomic distortions (and their parametrization) that we considered for this study.

- **Luminance:** Being one of the fundamental global properties of images, we seek to adjust it. Increasing and decreasing ambient light within an image is expected to affect recognizability. A scale factor parameter controls this in the following way. The RGB components of each pixel are scaled by scale factor, such that the average luminance over the entire image is also scaled by this scale factor. Too much or too little brightness are both expected to affected recognizability.
- **Color Quantization:** Instead of allowing the full color range, we quantize the color space for image representation. For each image, we transform pixels from RGB to CIE-LUV color space. The resultant color points, represented in $\mathbb{R}^3$ space, are subject to k-means clustering with k-center initialization [14]. A parameter controls the number of color clusters generated by the k-means algorithm. All colors are then mapped to this reduced set of colors. A lower number of color clusters translates to

TABLE I
SOME FEATURES AND DISTORTIONS THAT AFFECT THEIR EXTRACTION

Feature	Affected by	Not Affected by
Local Color	Quantization, Dithering, Luminance, Noise	Cut/rescale
Color Histogram	Luminance, Noise, Cut/rescale	Quantization, Dithering
Texture	Quantization, Dithering, Noise	Luminance, Cut/rescale
Edges	Noise, Dithering	Quantization, Luminance, Cut/rescale
Segmentation & Shape	Dithering, Noise, Quantization	Luminance, Cut/rescale
Interest Points	Noise, Dithering, Quantization	Luminance, Cut/rescale

loss of information and hence lower recognizability.

- **Dithering:** Similar to half-toning of the printing industry, color dithering is a digital equivalent that uses a few colors to produce the illusion of color depth. This is a particularly attractive distortion method here, since it affects low-level feature extraction (on which machine recognition is dependent) while having, by design, minimal effect on human vision. Straightforward application of dithering is, however, ineffective for this purpose since a simple *mean filter* can restore much of the original image. Instead, we randomly partition the image in the following two ways:
  - Multiple random orthogonal partitions.
  - Image segments, generated using k-means clustering with k-center initialization on color, followed by connected component labeling.

  In either case, for each such partition, we randomly select y colors (being the parameter for this distortion) and use them to dither that region. This leaves a segment-wise dithering effect on the image, which is difficult to undo. We expect automatic image segmentation to be particularly affected. Distortion tends to have a more severe effect on recognizability at lower values of y.
- **Cutting and Re-scaling:** For machine recognition methods that rely on pixel-to-pixel correspondence based matching, scaling and translation helps making them ineffective. We simply take a portion of one of the four sides of the image, cut out between $10 - 20\%$ from the edge (chosen at random), and re-scale the remainder to bring it back to the original image dimensions. This is rarely disruptive to human recognition, since items of interest occupy the central region in our image set. On the other hand, it breaks the pixel correspondence. Which side to cut is also selected at random.
- **Line and Curve Noise:** Addition of pixel-wide noise to images is typically reversible by median filtering, unless very large quantities are added, in which case human recognizability also drops. Instead, we add stronger noise elements on to the image, at random. In particular, thick lines, sinusoids, and higher-order curves are added. Technically, we do not set the color of these lines and curves to zero; instead, to make detection and removal harder, we reduce the RGB components of each such line or curve by a randomly drawn factor, giving the illusion of being dark but not necessarily zero. The density of noisy lines and curves are controlled by parameter y. Lines and sinusoids are generated orthogonal to each axis, spaced by density parameter y. For higher order curves, y specifies the number of them to be added, each added at random positions and orientations.

These distortions are by no means exhaustive, as mentioned before. However, they are hand-picked to be representative of distortions that are potentially good candidates. We experimented with each of them individually, and their simultaneous application on images to produce *composite distortions*. None of the atomic distortions by themselves yielded results promising enough to satisfy the requirements. Hence composite distortions were the only way out. We give specific details of the composite distortions that proved effective for CAPTCHA design, in the results section (Sec. V).

III. EXPERIMENTAL SYSTEM: IMAGINATION

So as to put the implications of the distortion experiments into perspective, we first briefly describe our experimental system IMAGINATION[1] (IMAge Generation for INternet AuthenticaTION). The nomenclature is inspired by the fact that the system's success inherently depends on the imagination power of humans, to help them 'see through' distortion and fill in the 'gaps' introduced by distortion.

The overall system architecture of our system is shown in Fig. 2. Assume the availability of an image repository $\mathcal{R}$, each labeled with an appropriate word, and an *orthogonal partition generator* that randomly breaks up a rectangle of a given dimension into 8 partitions. The system generates a tiled image, dithers it to make automatic boundary detection hard, and asks the user to select near the center of one of the images. This is the **click** step. On success, an image

[1]A working version of the IMAGINATION system can be found at `http://alipr.com/captcha/`.

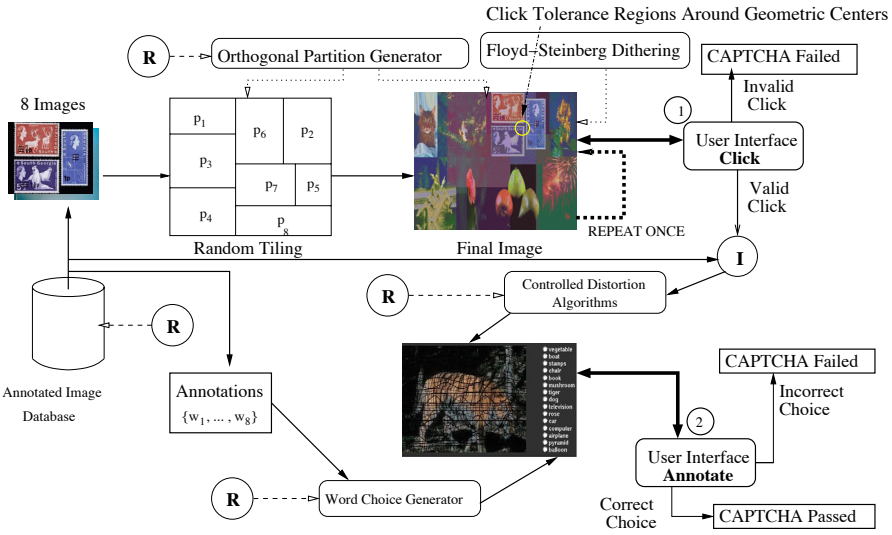


Fig. 2. Architecture of the IMAGINATION system. The circled 'R' components represent randomizations.

is randomly sampled, distorted by one of four methods and appropriate parameterizations (discussed in detail in Sec. V), and presented to the user along with a list of word choices, for labeling. This is the **annotation** step. These two steps are detailed below:

- **Click:** A single image is created on-the-fly by sampling 8 images from $\mathcal{R}$ and tiling them according to a randomly generated orthogonal partition. This image is then similarly partitioned twice over. Each time, and for each partition, 18 colors are chosen at random from the RGB space and are used to dither that partition using the two-stage Floyd-Steinberg error-diffusion algorithm [11]. The two rounds of dithering are employed to ensure that there is increased ambiguity in image borders (more candidate 'edges'), and to make it much more difficult to infer the original layout. An example of such an image is shown in Fig. 3. What the user needs to do is select near the physical center of any one of the 8 images. Upon successfully clicking within a tolerance radius r of one of the 8 image centers, the user is allowed to proceed. Otherwise, authentication is considered failed.
- **Annotate:** Here, an image is sampled from $\mathcal{R}$, a distortion type and strength is chosen (from among those that satisfy the requirements - we find this out experimentally, as described in Sec. V), applied to the image and presented to the user along with an unambiguous choice of 15 words (generated automatically). A sample screenshot is presented in Fig. 4. If the user fails in image recognition, authentication is immediately considered failed and re-start from step 1 is necessary.

These two click-annotate steps are repeated once more for added security. The convenience of this interface lies in the fact that no typing is necessary. Authentication is completed using essentially four mouse clicks. The word choices can be translated automatically to other languages if needed.

Word Choice Generator: The word choice generator quickly creates an unambiguous list of 15 words, inclusive of the correct label. For this, we make use of a WordNet-based [22] word similarity measure proposed by Leacock and Chodorow [17]. The 14 incorrect choices are generated by sampling from the word pool, avoiding any one that is too similar semantically (determined by a threshold on similarity) to the correct label. A more elaborate strategy was proposed in [8], but we found that for limited pools of words, this simpler strategy was equally effective.

Orthogonal Partition Generator: Optimal rectangle packing (within a larger rectangle), with minimum possible waste of space, is an NP-complete problem. Approximate solutions to this problem have been attempted before, such as in recent work of R.E. Korf [16] However, waste of space is not an issue for us, nor are rectangles to pack rigid, i.e., linear stretching is allowed. Our approach is as follows.

The full rectangular area is first partitioned vertically or horizontally (chosen randomly) into two equal rectangles. The sub-rectangles so formed are further partitioned recursively, strictly alternating between horizontal and vertical. The point of partition is sampled uniformly at random along a given length. We stop when the required number of sub-rectangles (8 in our experiments) are formed. In the Appendix, we explain in greater detail this process of generating partitions. It is important that the adversary be unable to take advantage of any non-uniformity in the way the partitions are generated. In other words, the center coordinates of sub-rectangles so formed should be drawn from a jointly uniform distribution, such that within the plausible region that an image center may lie, every point is equally probable. In the Appendix, we show that the way we generate the partitions guarantees this uniformity.

A. Vulnerability of 'Click' Stage to Attack

In the experiments (Sec. V), we primarily analyze vulnerability of the 'Annotate' stage to automatic image recognition. If

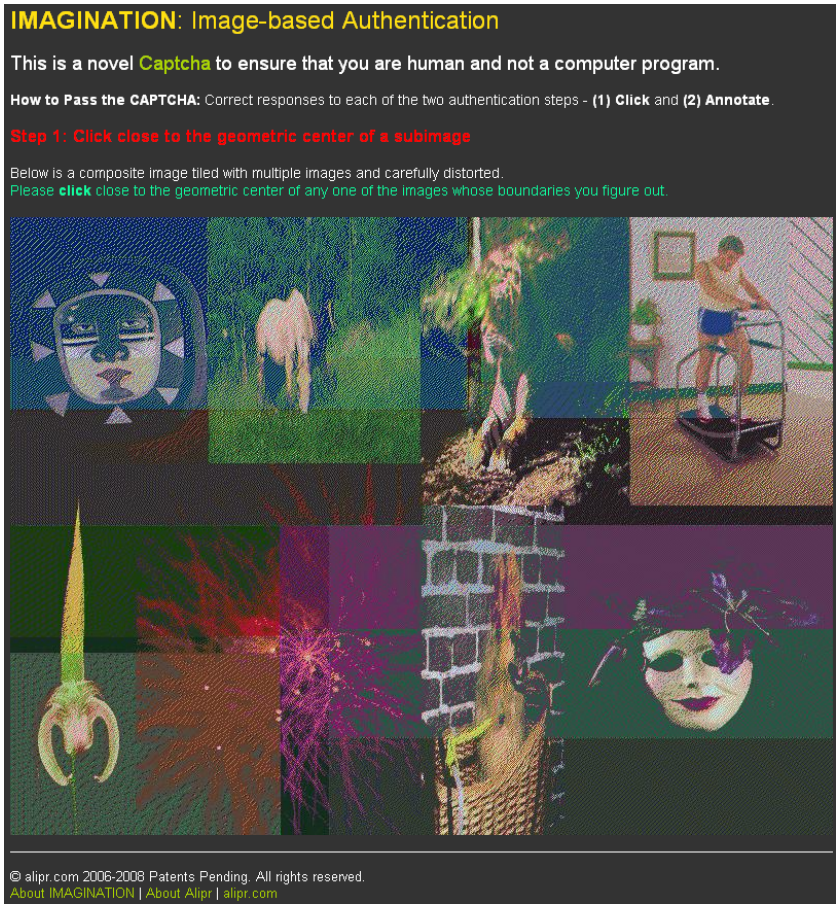


Fig. 3. Screenshot of the **Click** step of authentication in the IMAGINATION system. The tiled image is randomly partitioned orthogonally and dithered using different color sets, to make it harder for automated identification of the image boundaries. The user must click near the center of one of the images to get past this step.

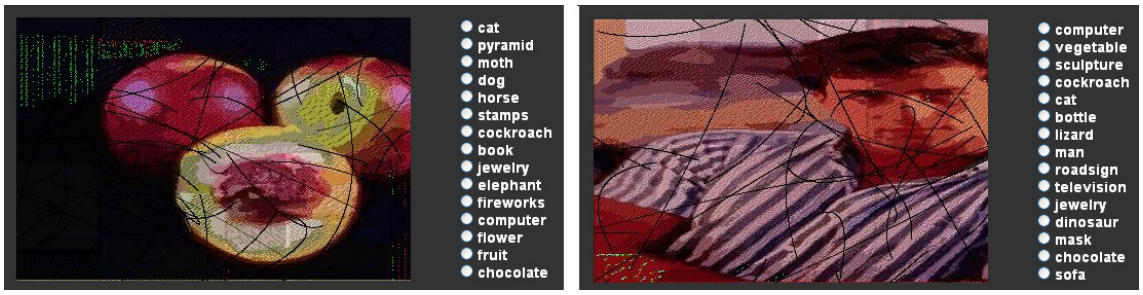


Fig. 4. Two screenshots of the **Annotate** step in the IMAGINATION system, where a distorted image is presented, and the user must select an appropriate label from a list of choices.

the adversary has full access to the image dataset, she should be able to use brute force to align one of the eight images within the tiled image, and determine its center. However, this will be an extremely expensive operation.

For an attack endeavor to be successful, it is essential to be able to find the corner coordinates of one of the 8 images. Exact match with an image in the database is made challenging by the following factors:

- After tiling the images, the full image is re-partitioned and dithered twice (described previously), fragmenting the color composition of each image.
- The images are scaled to fit the generated partitions, not maintaining aspect ratio.

Because partial matches do not reveal the corner coordinates, the only way to get at them is to do a brute-force search over the tiled image, at various scales. This is still subject to the fact that given a perfect alignment, there exists an image similarity metric, which despite the dithering, decisively reveals the match. To get a pessimistic estimate of the threat (from the designer point of view), let us assume the adversary does have such a metric. Let us consider a tiled image of size 800×600, and that the adversary is attempting to match with one of the 8 images whose top-left corner coincides with the tiled image (Φ_{111} in the Appendix). Its bottom-right corner could

be placed anywhere from $(0,0)$ to $(X/2,Y)$, which in this case is $(400,600)$. Again, assume that the adversary can skip 5 pixels in each dimension without missing the exact match. In our database, there are currently 1050 images. This size can grow relatively easily. The number of image matches to be attempted is $400/5 \times 600/5 \times 1050 = 10080000$. Assuming that the matching metric takes 100μs for each image pair, it will take 1008 seconds, or about 17 minutes, to find one pair of corners.

While this analysis clearly indicates that the brute-force image matching approach to automatically solving the 'Click' stage is infeasible, there are some caveats with respect to commercial implementations and actual attempts at breaking them. First, the number of images in the database should be far more than 1050. Assuming that a realistic commercial implementation use a 1 million image database, it will take 11+ days to conduct a successful attack. Second, given the heavy multi-stage distortion applied to the tiled image, a metric for image matching may not always reveal the perfect alignment positions. Third, a more robust image matching algorithm may be necessary, which is likely to take more than 100μs to process, which would mean that attack in this manner will be more expensive than estimated here. On the other hand, more robust methods used in subimage matching [15] may be able to significantly reduce brute-force search. However, such methods need to be improved upon, since they are only robust to minimal distortions and limited, aspect-ratio-preserving rescaling of images.

B. Overall Success Rate of Random Attack

The size of the tiled image in the click stage is fixed at 800×600. The choice of the tolerance radius r is an important trade-off between ease of use and threat of random attacks. In Sec. V, we empirically show the impact of this choice on users. For now, let us assume that $r = 25$ is a reasonable choice, which corresponds very roughly to one-tenth the width of each contained image. Assuming that we are able to produce dithering and distortions that make it no easier to attack than by random guess, the success rate is approximately $\left(\frac{8\pi r^2}{800\times600}\frac{1}{15}\right)^2$ (see Appendix), or about 1 in 210,312. which can be considered quite costly for opening one e-mail account, for example. The tiled image, the word choices, and the final distorted image together take about 1 second to generate. For faster processing, a large set of distorted images over varied parameter settings can be pre-generated and stored.

IV. Qualitative Comparison with Existing CAPTCHAs

Before quantifying the efficacy of the click-annotate steps of the proposed IMAGINATION system, we draw qualitative comparison with existing CAPTCHA systems, both in public-domain existence and proposed in research publications. Reiterating that their purpose is to authenticate users as human without being disruptive or time-consuming, the basis for comparison among CAPTCHAs broadly includes (a) vulnerability to attacks, and (b) user-friendliness. Vulnerability, leading to the failure of such systems, can be due to one of the following:

1) The AI problem posed is fundamentally solved;
2) Use of cheap labor to solve the problem; and
3) Problematic implementation.

While (3) can be avoided by foolproof design and rigorous testing, and (2) cannot be avoided in principle, neither of them depend on the nature of the CAPTCHA. Our comparison, therefore, focuses on (1) which is the availability of tools to solve the AI problem posed. User-friendliness of such systems can be attributed to the following characteristics:

1) Time taken to solve a problem;
2) Chances of human failure;
3) Culture/language/educational bias; and
4) Accessibility (blind, deaf).

In the case of user-friendliness of a CAPTCHA, all these factors are important, so we consider each of them in the ensuing comparison. In order to make the comparisons concise, we group CAPTCHAs into broad classes, as follows:

- **Text based:** Text characters, typically in English, distorted in various ways (Fig. 1.a).
- **Generic image based:** Images of easily recognizable objects, shown to be labeled (Fig. 1.c).
- **Speciality image based:** Image classes easily distinguished by humans, hard for machines (Fig. 1.d).
- **Knowledge based:** Questions in a language, which require 'common sense' responses.
- **Audio based:** For people with vision problems, audio clips are presented for recognition.

Our IMAGINATION system falls roughly within the category of 'generic image based', but with some vital differences. In Table II, we compare it with the various classes of CAPTCHAs. We observe that for a majority of factors, our system is favorable compared to the rest. Furthermore, our system has so much randomness that it is not possible to encounter same or similar problems repeatedly, ruling out 'answer collection' as an attack strategy. Note that blindness poses a challenge to all visual CAPTCHAs. While the 'audio based' systems primarily serve to solve this issue, they are somewhat orthogonal in design, and hence are not compared. In the following sections, we also make a more detailed comparison of IMAGINATION with text and image based CAPTCHAs.

A. Comparison with Text-based CAPTCHAs

Text-based systems, such as the ones shown in Fig. 1 a. and b., remain arguably the most widely deployed forms of CAPTCHAs, nine years since their inception. The AI challenge involved is essentially optical character recognition (OCR), with the additional challenge that the characters are randomly distorted. Thanks to years of research in OCR, hand-writing recognition, and computer vision, a number of research articles, such as [24], [25], [31], have shown that such CAPTCHAs can be solved or 'broken', with reportedly over 90% success rate. This is a key motivation for exploring alternate paradigms [6]. The AI challenge posed by our

TABLE II
QUALITATIVE COMPARISON OF IMAGINATION WITH OTHER CLASSES OF CAPTCHAS

CAPTCHA Classes:	**Text based**	**Generic image based**	**Speciality image based**	**Knowledge based**	**IMAGINATION**
Examples	EZ-Gimpy	ESP-PIX [3]	Asirra [9], ARTiFACIAL [27]	NoSpam! [33]	-
Automated Solutions	Yes - [24], [25], [31]	Yes - Exact Match, Approximate Match [34], [7], [30], Automatic Annotation [19]	Yes - Asirra [12], [10]	Likely	Unlikely (by design)
Randomization	Moderate	Low	Moderate	Moderate	High (multi-stage)
Adversary advantage with dataset access	Word dictionary to prune guesses - [25], [24]	Exact/approximate match with image collection	Exact match - speciality DB like Petfinder [28]	Commonsense datasets, e.g. Cyc [18] helps answer questions	Unlikely (tested assuming dataset is available)
Input modality	Keyboard	Mouse	Mouse	Keyboard	Mouse
Time taken to solve	Quick	Quick	Moderate	Quick	Quick
Chances of human failure	Medium (distortion-attack tradeoff)	Low (no distortion)	Medium (distortion-attack tradeoff)	High (knowledge-dependent)	Medium (distortion-attack tradeoff)
Language bias	Medium (must recognize letters)	Low (via automatic translations)	Low	High (must comprehend sentences)	Low (via automatic translations)
Educational bias	Low	Low	Low	High (knowledge-dependent)	Low

IMAGINATION system is that of image recognition, which is arguably [1] a much harder problem to solve than OCR. Therefore, in principle, IMAGINATION is likely to be more resilient to automated attacks. Furthermore, text-based CAPTCHAs require typing of letters in a given language, say English, and its internationalization may require considerable effort which includes regenerating the CAPTCHA images. With IMAGINATION, the 'click' stage is language-independent, and for the 'annotate' stage, any standard language translator software can map the options from English to another language.

B. Comparison with Other Image-based CAPTCHAs

Within the paradigm of image-based CAPTCHAs, there are distinct examples, such as simple image recognition CAPTCHAs [5] which present users with undistorted generic images to be labeled using the provided word lists, Asirra [12], which present images of 12 cats and dogs and users are required to identify the cats among them, and ARTiFACIAL [27] which generates facial images and asks users to pinpoint facial features in them. Problems with presenting undistorted images, or arbitrarily distorted images, in these systems, are

- It is a security requirement of CAPTCHA systems to make data publicly available, in this case the set of labeled images used. Given this, a straightforward pixel-by-pixel matching algorithm should be sufficient to answer the image labeling question in such CAPTCHAs.
- If the dataset is indeed not made available, even then there is a problem. Real-time automatic image annotation systems such as Alipr [19], or other object recognition systems [7], [30], may be able to tag images at a moderate level of accuracy, especially when they are trained and used for concepts in a limited domain. This level will then translate into the success rate of attacks, if they are used by the adversary. For the more specialized image CAPTCHAs like Asirra, work on cat detection [10], [12] can go a long way in undermining their effectiveness.
- In case the images are distorted arbitrarily, approximate matching algorithms [34], [7] may be sufficient to match them to their originals, and hence obtain the results.

Our proposed IMAGINATION system has the advantage of posing a hard AI problem (image recognition) while avoiding the pitfalls of the other image recognition based systems. Distortions are applied so that presented images cannot be matched exactly to the image dataset, and these distortions are generated in a controlled manner such that approximate matching methods cannot be successfully applied. While researchers in computer vision have had success in image recognition for specific images classes, such as cats [10], recognition of generic image classes is considered a challenging open problem that is unlikely to be solved in the near future. Because our system works with an unrestricted range of image categories, the threat of this AI problem getting solved sometime soon is extremely low.

V. EXPERIMENTAL RESULTS

Large scale experiments were conducted using our publicly available IMAGINATION system[2], as well as our internal testbed. We obtained empirical results for both the **click** and the **annotate** steps, based on actual usage. We describe the setup and results below.

A. Stage 1: Click

The main variable component of this stage is the choice of r, the radius of tolerance around each image center that is

[2]http://alipr.com/captcha/

considered a valid click. We therefore wanted to see whether valid human users were able to click near the geometric centers or not, and if so, how near or far they clicked.

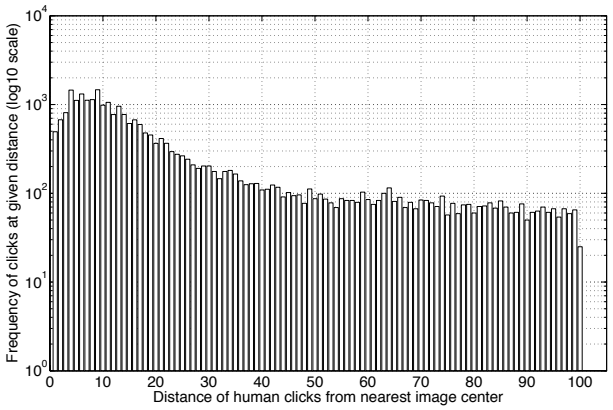


Fig. 5. Plot of distribution of distances of user clicks from the nearest image centers within the composite images. This distribution (26, 152 points), plotted in base-10 logarithmic scale, shows that a majority of the clicks are within a short distance of a center, followed by a long tail.

For this, we used the click data obtained from visitors to our public demo. Since there was evidence that a fraction of users attempted automated attacks on the system by randomly generating coordinates and trying them out, we had to denoise the data such that the majority of the clicks were genuine attempts at succeeding in the task. To achieve this, we randomly picked a single click data for each unique IP address. This way, multiple automated clicks from one machine will have been eliminated from consideration. After denoising the data, we had $26,152$ data points corresponding to as many unique IP addresses.

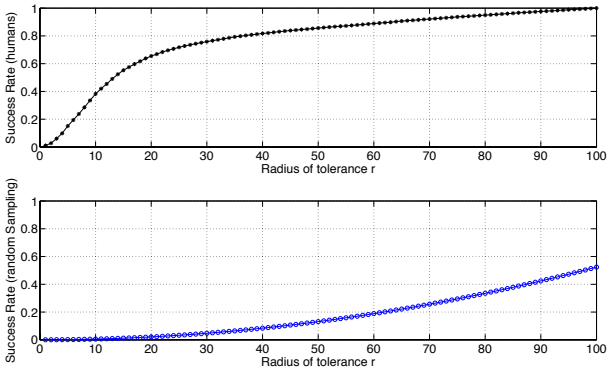


Fig. 6. Variation of success rates by human (above) and automated randomly sampled clicks (below) with varying tolerance radius r. This graph helps to choose r given a desired trade-off between human ease of use and insulation from attacks.

The distribution of the distance of the human clicks from their nearest image centers is shown in Fig. 5. We see that a majority of the clicks are in the vicinity of a valid image center, adding to the confidence that this step of the CAPTCHA is a reasonable one for humans. In order to choose a tolerance radius r, useful reference graphs are those plotted in Fig. 6. Here we see the empirical distribution of success rates over this set of user clicks, as r is varied. Assuming an attack strategy that involves uniformly sampling at random a pair of coordinates to artificially click at, its success rates are only dependent on r, and these are plotted below. Together, these two plots help determine a value that gives desirable security as well as usability. We can see that 25, which leads to a 70% user success rate and very low random attack success rate, is one good choice and hence is currently used in the IMAGINATION demo. The two outcomes of this experiment were (a) the verification of plausibility of this step, and (b) the selection of a desirable value for parameter r.

B. Stage 2: Annotate

The experiments related to the annotate step consisted of distorting images and measuring human and machine recognizability, over a set of 1050 Corel images covering 35 easily identifiable categories. Machine recognizabilities was based on the similarity measures PWD, EMD, and IRM (detailed in Sec. II). Human recognizability was measured based on a user study consisting of over 250 individuals, receiving over 4700 responses. The user study consisted of presenting distorted images and a list of 15 words to each user (See Fig. 4), allowing them to select an appropriate label, or choose 'I cannot recognise' (enabled only during experimentation). Recognition is considered failed if the latter is chosen, or if an incorrect label is chosen. The following summarizes recognizabilities of humans and machines, and their *recognizability gap*.

1) Atomic Distortions: We first analyzed results obtained from the application of atomic distortions on images. In particular, the effect of luminance adjustment, noise addition, color quantization, and dithering, each in isolation, were studied. For the latter two distortions, cut/rescale was also applied for comparison. These results are presented in Figures 7, 8, 9, and 10 respectively. Dithering here is based on orthogonal block partitioning. In each case, the range of values for which human recognizability exceeds 0.9 are shown within Magenta colored dashed lines. They help understand how human and machine recognizabilities contrast.

When pixel correspondence is unaffected, the pixel-wise distance (PWD) performed quite well. However, with the cut/rescale addition, this correspondence is broken and we see significant degradation of PWD's performance (Fig. 9 and 10). In general IRM shows well-balanced performance, making it a good general-purpose attack tool. Note also that in all these atomic distortion cases, the range where human recognizability is high, at least one of the machine-based methods show high recognizability as well. From this observation, we conclude that any one atomic distortion, does not provide the requisite security from attacks while still being able to maintain human recognizability. This leads us to searching the space of composite distortions. Nonetheless, the results of atomic distortion give a clear insights and help build intuitions about how to combine them effectively.

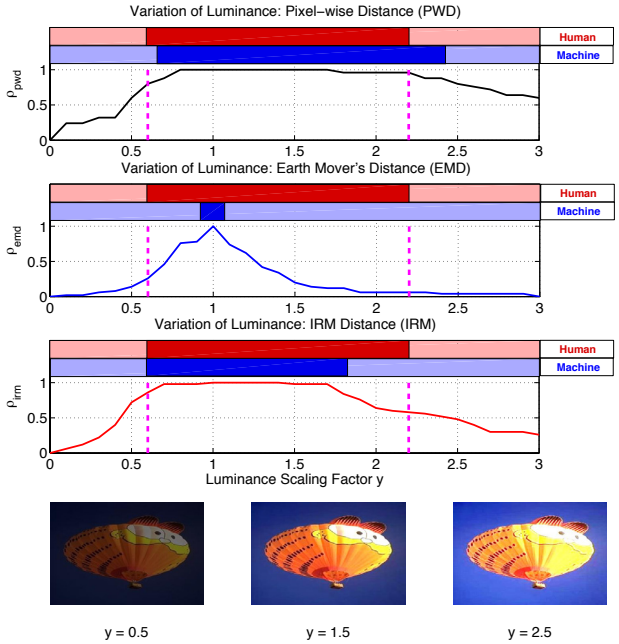


Fig. 7. Variation of average machine recognizability with change in luminance scaling factor. Human recognizability is high within the Magenta lines. The red and blue regions above the graphs show ranges (and overlaps) of human and machine recognizability. Dark and light shades indicate 'high' and 'low' recognizability respectively. Machine recognizability is considered 'high' for $\rho \geq 0.8$.

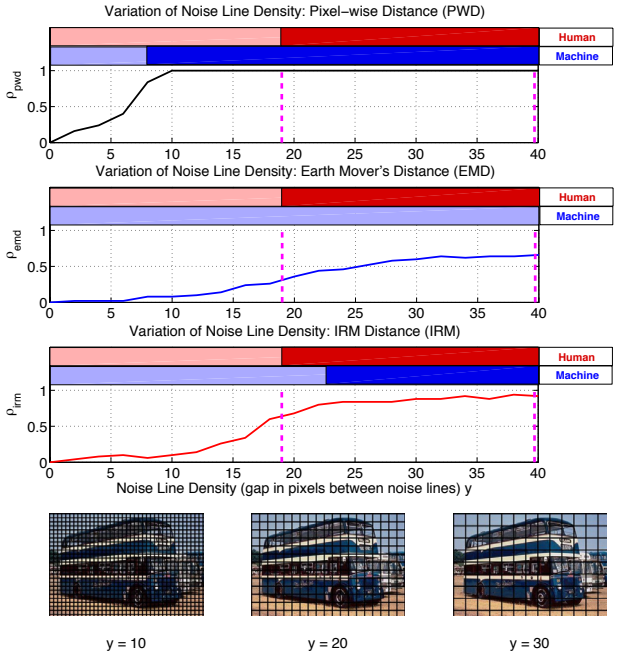


Fig. 8. Variation of average machine recognizability with change in density of noisy lines added, represented in pixels specifying the gap between consecutive lines. Human recognizability is high within the Magenta lines. The red and blue regions above the graphs show ranges (and overlaps) of human and machine recognizability. Dark and light shades indicate 'high' and 'low' recognizability respectively. Machine recognizability is considered 'high' for $\rho \geq 0.8$.

2) Composite Distortions: An exhaustive search for composite distortions is prohibitively expensive. One may be able to think of algorithmic means to arrive at a composite distortion that satisfies the image CAPTCHA requirements.

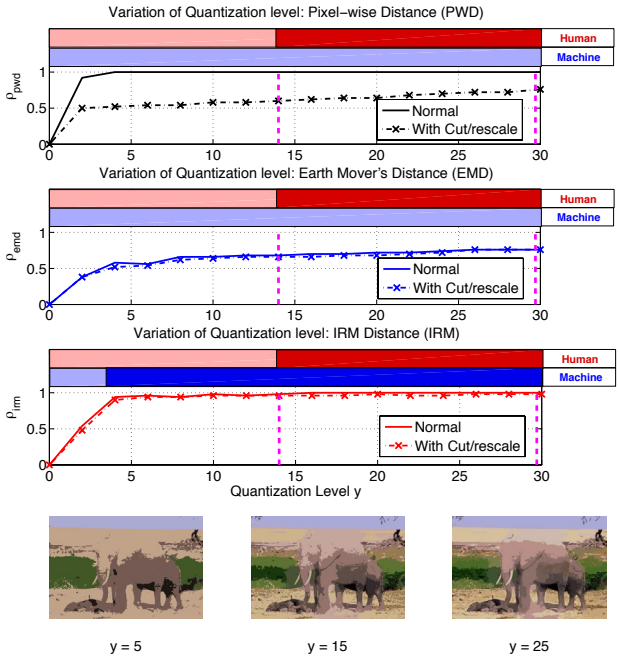


Fig. 9. Variation of average machine recognizability with change in quantization level, specified in terms of the number of color clusters generated and (centroids) used for mapping. Human recognizability is high within the Magenta lines. The red and blue regions above the graphs show ranges (and overlaps) of human and machine recognizability. Dark and light shades indicate 'high' and 'low' recognizability respectively. Machine recognizability is considered 'high' for $\rho \geq 0.8$, and we show the cut/rescale case here.

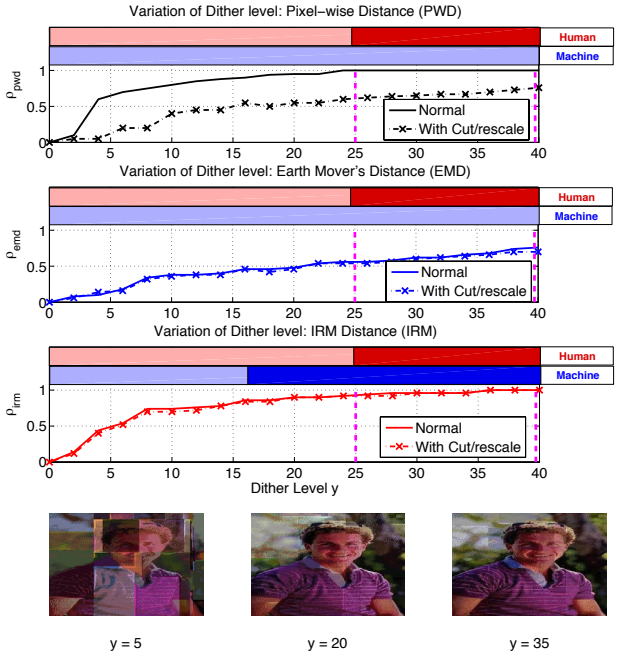


Fig. 10. Variation of average machine recognizability with change in dithering level, specified in terms of the number of colors available for dithering each partition. Human recognizability is high within the Magenta lines. The red and blue regions above the graphs show ranges (and overlaps) of human and machine recognizability. Dark and light shades indicate 'high' and 'low' recognizability respectively. Machine recognizability is considered 'high' for $\rho \geq 0.8$, and we show the cut/rescale case here.

For example, if atomic distortions are considered analogous to features in a learning problem, then forward-backward selection [2] seems to be an appropriate choice, adding and removing atomic distortions (ordered), testing recognizability, and stopping on satisfactory performance. The bottlenecks to systematic search for acceptable composite distortions are:

- **Search space is large:** Not only are there many possible atomic distortions, they are also parameterized. Each add/remove step need also iterate over the possible parameter values. For each distortion-parameter pair, machine recognizability needs to be measure over multiple test images.
- **Humans in the loop:** The search space being so large, what is even more problematic is measuring human recognizability at each step, requiring feedback from multiple users over multiple images.
- **Lack of Analytical Solution:** Given its nature, it is difficult to formulate it theoretically as an optimization problem, without which analytical solutions are not possible.

Instead, we heuristically selected permutations of the atomic distortions and experimented with them. Based on preliminary investigation, four composite distortions seemed particularly attractive, and we conducted large-scale experimentation on them.

Detailed description of each of the four chosen composite distortions are presented in Fig. 11, along with the corresponding experimental results. Each of them are controlled by parameters DITHERPAR, which controls the extent of dithering, and DENSEPAR, which controls the density of noise elements added. The color-coded matrices represent the corresponding degrees of human/machine recognizability. To better visualize the recognizability gap as well as make the problem harder, the three types of machine recognition are combined together in the following way. If any one of PWD, EMD, or IRM recognizes an image, it is considered as successful machine recognition. We find that for a limited range of parameter values in each of them, human recognizability is high (exceeds 0.9) while machine recognizability is low (below 0.1). These distortion type and parameter value/range combinations are appropriate for inclusion into our experimental system IMAGINATION. The few cases where machine recognizability exceeds human recognizability are also worth exploring, but they are beyond the scope of this paper.

To further the investigation and help design the IMAGINATION system better, we studied the trends of human recognizability from the user responses. Figure 12 presented the variation of recognizability with parameter values across all four distortion types, revealing the general trend associated with DITHERPAR and DENSEPAR regardless of the distortion type. More specifically, a greater number of dithering colors tend to help humans recognize image content better, while greater quantities of noise hinder their recognition. Figure 13 reveals yet another aspect of the recognition process, namely the average human recognizability per concept, taken over varying distortion type and strength. As can be seen, some concepts (e.g., parade, vegetable) are inherently harder

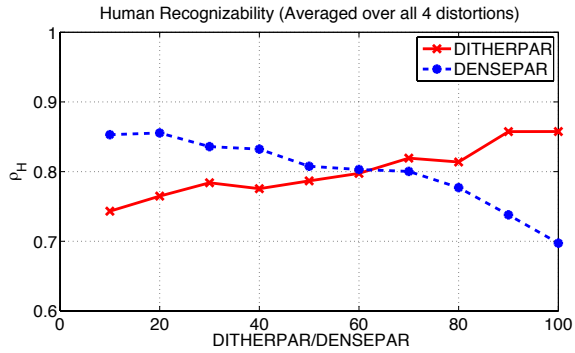


Fig. 12. Overall variation of human recognizability with dithering parameter DITHERPAR and noise density parameter DENSEPAR, taken across all four composite distortion methods.

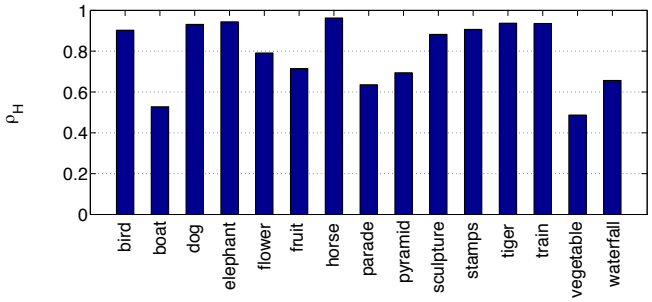


Fig. 13. Overall variation of human recognizability with the image concept, taken across all four composite distortion methods and their parameterizations. The fifteen most frequently sampled concepts are shown here.

to identify than others, regardless of distortion.

The results we presented here are over-optimistic from the point of view of attacks. This is because human recognizability only involves identifying the entity and not 'matching' any specific pair of images. If we increase the number of images in the repository $\mathcal{R}$, machine recognizability is bound to suffer, while human recognizability should remains at about the same level as reported here. A real-world system implementation will have many more than 1050 in its repository, and will thus be more secure. Also note that with a 15 word choice list, the distortions never need to reduce machine recognizability to less than $1/15$, since randomly selecting a word without even considering the image would yield a $1/15$ chance.

VI. Conclusions

We have presented a novel way to distinguish humans from machines by an image recognition test, one that has far-reaching implications in computer and information security. The key point is that image recognition, especially under missing or pseudo information, is still largely unsolved, and this fact can be exploited for the purpose of building better CAPTCHA systems than the vulnerable text-based CAPTCHAs that are in use today. We have explored the space of systematic distortions as a means of making automated image matching and recognition a very hard AI problem. Without on-the-fly distortion, and with the original images publicly available, image recognition by matching is a trivial task. We have learned that atomic distortions are largely ineffective

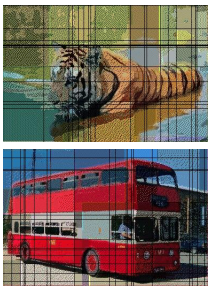
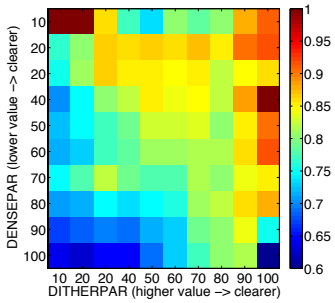
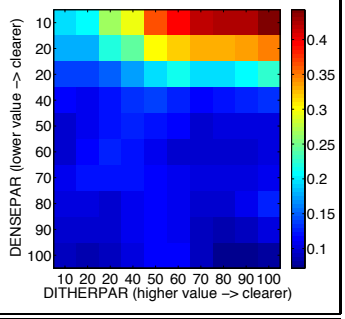
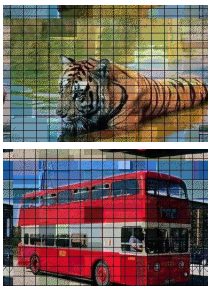
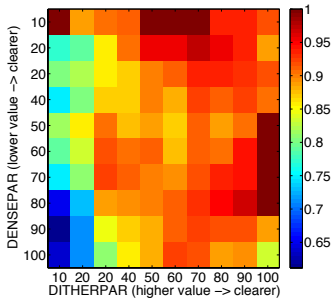
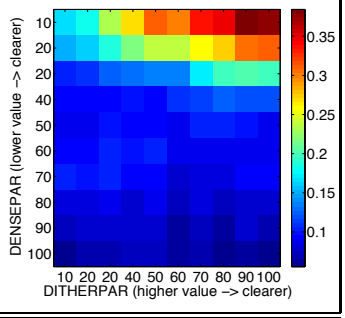
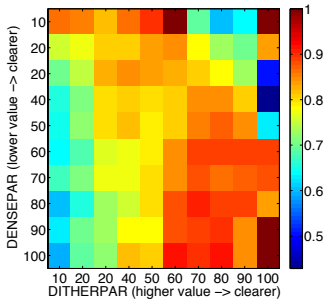
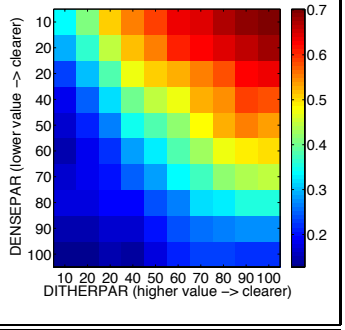
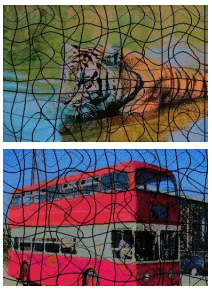
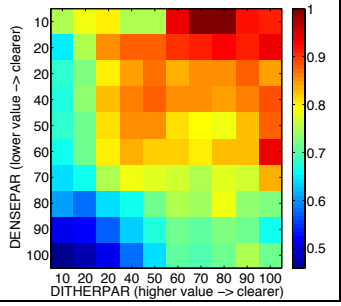
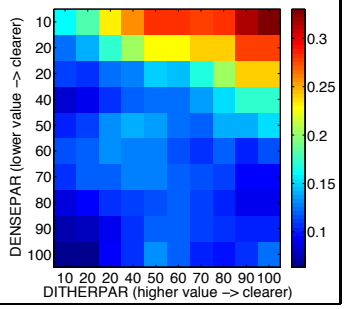
Distortion Steps	**Sample Images**	**Human Recognizability (User Study)**	**Machine Recognizability (PWD + EMD + IRM)**
1. Perform k-center/k-means based segmentation (k=15). 2. Use cluster centroids to quantize image. 3. Create block partitioning of the image using the Orthogonal Partition Generator. 4. Dither each *block* with an independently drawn random set of DITHERPAR colors. 5. Draw DENSEPAR lines parallel to each axis, *randomly spaced*. 6. Perform $10-20\%$ cut/rescale on a randomly chosen side.			
1. Perform k-center/k-means based segmentation (k=15). 2. Use cluster centroids to quantize image. 3. Create block partitioning of the image using the Orthogonal Partition Generator. 4. Dither each *block* with an independently drawn random set of DITHERPAR colors. 5. Draw DENSEPAR lines parallel to each axis, *equally spaced*. 6. Perform $10-20\%$ cut/rescale on a randomly chosen side.			
1. Perform k-center/k-means based segmentation (k=15). 2. Use cluster centroids to quantize image. 3. Create block partitioning of the image using the Orthogonal Partition Generator. 4. Dither each *block* with an independently drawn random set of DITHERPAR colors. 5. Draw DENSEPAR *third-order curves*, 1-3 pixels thick, *randomly positioned*. 6. Perform $10-20\%$ cut/rescale on a randomly chosen side.			
1. Perform k-center/k-means based segmentation (k=15). 2. Use cluster centroids to quantize image. 3. Perform connected component labeling to get image segments. 4. Dither each *segment* with random set of DITHERPAR colors. 5. Draw DENSEPAR *sinusoids* with axes parallel to each axis, *randomly spaced*. 6. Perform $10-20\%$ cut/rescale on a randomly chosen side.			

Fig. 11. Four Distortions that are part of the IMAGINATION System.

in reducing machine-based attacks, but when multiple atomic distortions combine, their effect significantly reduce machine recognizability.

Our study, while in no way encompassing the entire space of distortions (or algorithms that can recognize under distortion), presents one way to understand the effects of distortion on the recognizability of images in general, and more specifically to help design image CAPTCHA systems. Furthermore, it attempts to expose the weaknesses of low-level feature extraction to very simple artificial distortions. As a bi-product, an understanding of the difference in recognizability of algorithms and humans under similar conditions also provides an opportunity for better feature extraction and matching distance design.

Appendix: Orthogonal Partition Generation

We now illustrate how the composite images are generated and prove that the approach leads to uniformly distributed placement of image centers. The significance of uniform distribution is that even if the adversary was aware of the

algorithm for generating the composite images, the person would not be able to improve chance of successful attack to better than random. In other words, if the algorithm generated constituent image positions non-uniformly, the adversary may predict image centers more often around regions of higher density, thereby increasing success chance even without any image processing.

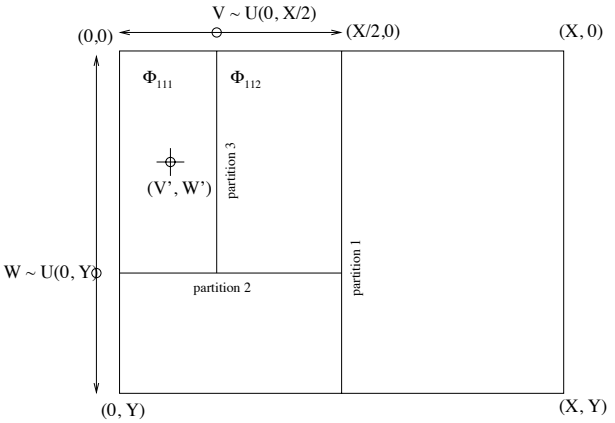


Fig. 14. Steps to orthogonal partition generation, to create 8 rectangular sub-regions for image tiling.

Let the composite image have dimensions $X \times Y$, and be denoted by Φ. Let the *uniform distribution* over an $[a, b]$ range be denoted by $U(a, b)$. To achieve uniformity in partitioning Φ to generate 8 sub-images, the following algorithm is employed.

- Partition Φ along the center, either horizontally or vertically (randomly chosen), at $X/2$ or $Y/2$ respectively, to get Φ_1 and Φ_2 respectively.
- Recursively partition Φ_1 and Φ_2 further. Here we explain the case of horizontally-split left-side rectangle Φ_1, as shown in Fig. 14. Other cases are similar.
  - Sample $V \sim U(0, Y)$ and partition Φ_1 vertically along V, to generate two more sub-rectangles Φ_{11} and Φ_{12}.
  - Sample $W \sim U(0, X/2)$ and partition the upper sub-rectangle Φ_{11} horizontally along W, to further generate sub-rectangles Φ_{111} and Φ_{112}.
- In a similar way for the remaining cases, we end up with 8 partitions Φ_{111}, Φ_{112}, Φ_{121}, Φ_{122}, Φ_{211}, Φ_{212}, Φ_{221}, and Φ_{222}.

Let us now analyze the probability distribution of the center of sub-rectangle Φ_{111}, with analysis of the other sub-rectangles being similar. The top-left corner of Φ_{111} is at $(0, 0)$. The bottom edge row is at W which is drawn uniformly at random over $[0, Y]$. Similarly, the right edge column is at V, which is drawn uniformly at random over $[0, X/2]$. Furthermore, V and W are drawn conditionally independent of each other. Therefore, the joint p.d.f. $f(v, w)$ of the random vector (V, W) is given by

$$f(v, w) = f_V(v) f_W(w) = \frac{1}{Y}\frac{2}{X} = \frac{2}{XY}$$

where $f_V(v)$ and $f_W(w)$ are the marginal densities. Furthermore, if a variable Z is drawn uniformly from $[0, c]$, having p.d.f. $\frac{1}{c}$, a new variable $T = Z/2$ is distributed uniformly over $[0, c/2]$ and has p.d.f. $\frac{2}{c}$. Therefore, given that the rectangle Φ_{111} spans $(0, 0)$ to (V, W), its center (V', W') is located at $(V/2, W/2)$, is uniformly distributed, and has a joint p.d.f. $f(V', W')$ given by

$$f(v', w') = f_{V'}(v') f_{W'}(w') = \frac{1}{2} f_V(v) \frac{1}{2} f_W(w) = \frac{8}{XY}$$

since V' and W' are conditionally independent of each other.

In other words, for a composite image Φ of size $X \times Y$, the center (V, W) of the sub-rectangle Φ_{111} (and similarly for other sub-rectangles), generated by the algorithm above, can lie anywhere in the $(0, 0) - (X/4, Y/2)$ region with equal probability. Therefore, without analyzing the composite image content, given a single shot at clicking near the center of Φ_{111} (or any other sub-rectangle of choice), the adversary's success probability is approximately $\frac{8\pi r^2}{XY}$, where r is the tolerance radius.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. 0347148, 0219272, 0705210. The authors would like to thank Dhiraj Joshi for valuable discussions and for assistance in creating a Web-based system used in evaluation of the work. Razvan Orendovici has created the Web-based front-end system which allows Web users to test the work and to provide feedback. Undergraduate students of some Penn State courses were involved in the evaluation user study.

REFERENCES

[1] L. von Ahn, M. Blum, and J. Langford, "Telling Humans and Computers Apart (Automatically) or How Lazy Cryptographers do AI," *Communications of the ACM*, 47(2):57-60, 2004.
[2] A.L. Blum and P. Langley, "Selection of Relevant Features and Examples in Machine Learning," *Artificial Intelligence*, 97(1–2):245–271,1997.
[3] "The CAPTCHA Project," *http://www.captcha.net*.
[4] K. Chellapilla and P. Y. Simard, "Using Machine Learning to Break Visual Human Interaction Proofs (HIPs)," *Proc. NIPS*, 2004.
[5] M. Chew and J. D. Tygar, "Image Recognition CAPTCHAs," *Proc. ISC*, 2004.
[6] Computerworld, "Building a better spam-blocking CAPTCHA," `http://www.computerworld.com/action/article.do?command=viewArticleBasic&articleId=9126378`. Retrieved on 01/23/2009.
[7] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, 40(2):1-60, 2008.
[8] R. Datta, J. Li, and J. Z. Wang, "IMAGINATION: A Robust Image-based CAPTCHA Generation System," *Proc. ACM Multimedia*, 2005.
[9] J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization," *Proc. ACM CCS*, 2007.
[10] F. Fleuret and D. Geman, "Stationary Features and Cat Detection," *J. Machine Learning Research*, 9:2549-2578, 2008.
[11] R.W. Floyd and L. Steinberg, "An Adaptive Algorithm for Spatial Grey Scale," *Proc. Society of Information Display*, 17:75-77, 1976.
[12] P. Golle, "Machine learning attacks against the Asirra CAPTCHA," *Proc. ACM CCS*, 2008.
[13] Guardian, "How Captcha was foiled: Are you a man or a mouse?" `http://www.guardian.co.uk/technology/2008/aug/28/internet.captcha`. Retrieved on 08/28/2008.
[14] A.K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.
[15] Y. Ke, R. Sukthankar, and L. Huston, "Efficient Near-duplicate Detection and Subimage Retrieval," *Proc. ACM Multimedia*, 2004.

[16] R.E. Korf, "Optimal Rectangle Packing: New Results," *Proc. ICAPS*, 2004.
[17] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," *Fellbaum*, 1998.
[18] D.B. Lenat, "Cyc: A Large-Scale Investment in Knowledge Infrastructure," *Comm. of the ACM*, 38(11):33-38, 1995.
[19] J. Li and J.Z. Wang, "Real-time Computerized Annotation of Pictures," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.
[20] D.G. Lowe, "Object Recognition from Local Scale-invariant Features," *Proc. ICCV*, 1999.
[21] C.L. Mallows, "A Note on Asymptotic Joint Normality," *Annals of Mathematical Statistics*, 43(2):508–515, 1972.
[22] G. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, 38(11):39-41, 1995.
[23] W.G. Morein, A. Stavrou, D.L. Cook, A.D. Keromytis, V. Mishra, and D. Rubenstein, "Using Graphic Turing Tests To Counter Automated DDoS Attacks Against Web Servers," *Proc. ACM CCS*, 2003.
[24] G. Mori and J. Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA," *Proc. IEEE CVPR*, 2003.
[25] G. Moy, N. Jones, C. Harkless, and R. Potter, "Distortion Estimation Techniques in Solving Visual CAPTCHAs," *Proc. IEEE CVPR*, 2004.
[26] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Intl. J. of Computer Vision*, 40(2):99-121, 2000.
[27] Y. Rui and Z. Liu, "ARTiFACIAL: Automated Reverse Turing Test using FACIAL Features," *Proc. ACM Multimedia*, 2003.
[28] J. Saul, "Petfinder," `http://www.petfinder.com`. Retrieved on 01/22/2009.
[29] Slashdot, "Yahoo CAPTCHA Hacked", `http://it.slashdot.org/article.pl?sid=08/01/30/0037254`. Retrieved on 01/30/2008.
[30] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years," *IEEE Trans. Pattern Analysis and Machine Intelligence* , 22(12):1349-1380, 2000.
[31] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla, "Shape Context and Chamfer Matching in Cluttered Scenes," *Proc. IEEE CVPR*, 2003.
[32] A. Turing, "Computing Machinery and Intelligence," *Mind*, 59(236):433-460, 1950.
[33] VBulletin, "NoSpam! An Alternative to CAPTCHA images," `http://www.vbulletin.org/forum/showthread.php?t=124828`. Retrieved on 01/20/2009.
[34] J.Z. Wang, J. Li, and G. Wiederhold "SIMPLIcity: Semantics-Sensitive Integrated Matching for Picture LIbraries," *IEEE Trans. Pattern Analysis and Machine Intelligence*, 23(9):947-963, 2001.