

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*

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 .

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 .

J.Z. Wang is with the College of Information Sciences and Technology, The Pennsylvania State University, University Park, PA 16802, USA. He is also with Carnegie Mellon University, Pittsburgh, PA 15213, USA. Phone: (814) 865-7889. E-mail: jwang@ist.psu.edu .

Abstract

The multimedia research community has, for long, attempted to recognize generic images, make them 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 has become well-known. In this paper, we explore the exploitation of this limitation for solving a security problem. 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 and 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 three 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 the multimedia community, CBIR is actually conceived here as a tool for an adversary, so as to help us design secure image CAPTCHAs by understanding their limitations.

Index Terms

Image recognition, CBIR applications, CAPTCHAs, distortions.

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. [23] 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 [6], [23]. 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], [25] 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 and prevent denial-of-service attacks on Web servers [18], among others.



(a) Yahoo's EZ-Gimpy

(b) PayPal's CAPTCHA



(c) CMU's ESP-PIX

Choose a word that relates to all the images.

↓

TIP: You can type the first letter of a word and then use the down arrow to find it.

Submit



(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.

There has been sizable research output in making 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 means (by definition) the advancement of artificial intelligence. 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 [19], [24], [4], [20], citing over 90% success rate, thus advancing OCR technology. Among the earliest commercial ones, the Yahoo! CAPTCHA has also been reportedly compromised, with a success rate of 35% [27], 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], [7], [8], [22]. 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 text with images may subject it to similar risks of being 'broken' by image recognition techniques. Techniques such as near-duplicate image matching [11], content-based image

retrieval [23], and real-time automatic image annotation [14] are all potential attack tools. 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 a formal study. Furthermore, while there has 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. To re-iterate, 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 considerations 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 types/strengths of 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 up the missing pieces,

owing to their power of imagination.

- For the purpose of designing CAPTCHAs, 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.

Besides the design of security mechanisms, this study can also be viewed as highlighting the shortcomings of image matching algorithms, especially the use of low-level cues for such high-level tasks, and the limits of the human vision system to recognize entities under strong, random, artificial distortions.

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 recognizeability. In Sec. III, we describe our experimental system IMAGINATION. In Sec. IV, we present extensive experimental results on the effect of distortions on human and machine recognizability. We conclude in Sec. V.

II. IMAGE RECOGNIZABILITY UNDER DISTORTION

For the purpose of understanding the effect of distortions and for the designing of image CAPTCHAs, let us assume that we have a collection of natural images with a single dominant subject, such that given a limited set of options (say 15), choosing a label is unambiguous. First, we concretely define how machine and human recognizability are measured. Then we 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 [23] 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* [11], which focus on finding marginally modified/distorted copyrighted images, also seems to be a good choice here. This is part of our future work. *Automatic image annotation* and *scene recognition*

techniques [6] also have potential, but given the current state-of-the-art, these methods are very 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., $I \in \mathcal{X} \forall I$. We can think of the image retrieval algorithm as a function that takes in a pair of images and produces a distance measure $g(I_1, I_2)$ (that hopefully correlates well with their semantic distance). Define a rank function

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

We relax the criteria for machine recognizability, treating image I_1 as recognizable if $\text{rank}_g(I_1, I_2, \mathcal{X})$ 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 a given distortion $\delta_y(\cdot)$, where machine is equivalent to an image retrieval system modeled as $g(\cdot, \cdot)$, is given as

$$\overline{\rho}_g(\delta_y) = \frac{1}{|\mathcal{X}|} \sum_{I \in \mathcal{X}} \mathcal{I}(\text{rank}_g(I, \delta_y(I), \mathcal{X}) \leq K) \quad (2)$$

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 (In our experiments, all test images are of the same dimensions). If the two images are I and I' , then

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

where $|I|$ here denotes the total number of pixels in the image. This measure clearly lacks robustness, and is expect to be sensitive even to very small distortions. Second, we employ the Earth Mover's Distance (EMD) [21] (which is essentially the earlier proposed Mallow's Distance [16]) based on global color features and a robust, true distance metric. First, we employ the more recent IRM distance which forms the backbone of the SIMPLicity system [26]. 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 survey. An image I is sampled from \mathcal{X} , subjected to distortion $\delta_y(\cdot)$, and then presented to an user, along with a set of 15 word choices, one of which is unambiguously an appropriate label. 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 the entire image set \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 that were subjected $\delta_y(\cdot)$, then

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

where \mathcal{I} is the indicator function. The implicit assumptions made here, under which the term $\overline{\rho_{\mathcal{H}}}(\delta_y)$ can be fairly compared to $\overline{\rho_{emd}}(\delta_y)$ or $\overline{\rho_{irm}}(\delta_y)$, are that (a) all users are independent and identical instances of the ‘human’ prototype, and (b) with sufficient, but not necessarily identical number of user responses, the average recognizability measures converge to their true value.

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. If that was the case, then machine recognition would be able to proceed normally with the insertion of such a pre-processing step, which is not desirable. 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

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

step in generic recognition tasks has been *low-level feature extraction* from the images [23], [6]. 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 pairwise 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 [15] 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 particular 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 $\{Quantize_y(\cdot), Dither_y(\cdot), \dots\} \in \Delta$.
- 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 by definition. Here, we list the atomic distortions (and their parametrization) that we considered for the purpose of 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 less 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 [10]. 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 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 for dithering that region. This leaves a block or segment-wise dithering effect on the image, which is difficult to undo by filtering. 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 simple take a portion of one of the four side 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. 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 as noise, 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*. It is worthwhile to mention here though that *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. IV).

III. EXPERIMENTAL SYSTEM: IMAGINATION

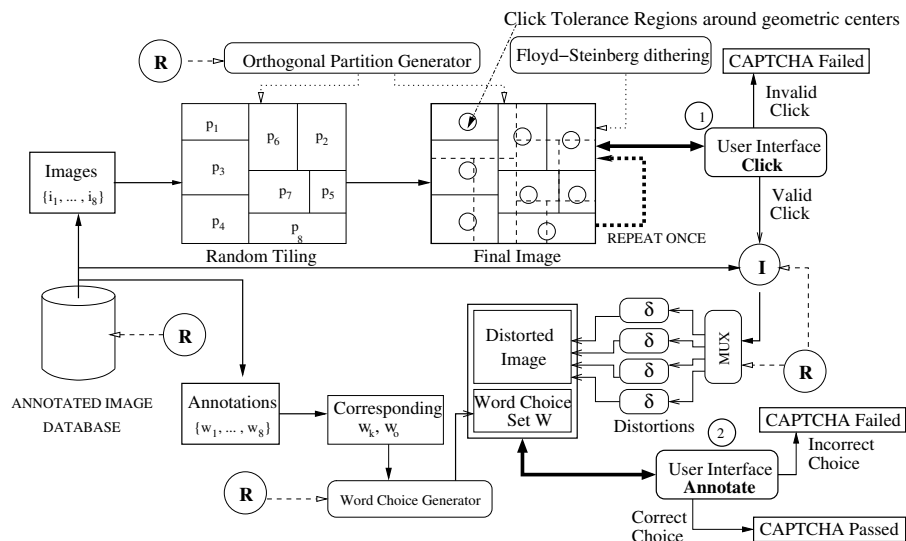


Fig. 2. Architecture of the IMAGINATION system.

So as to put the implications of the distortion experiments into perspective, we first briefly describe our

experimental system IMAGINATION¹ (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 up the 'gaps' introduced by distortion, aided by past experience and exposure.

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 orthogonal 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 is sampled from the repository, distorted by one of four methods and appropriate parameterizations (discussed in detail in Sec. IV), and presented along with a list of word choices to the user of selection. 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 an orthogonal partition generated. 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 [9]. 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 revert back to 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. On 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. IV), 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. Alternation between these two steps is part of our design, but having two successive click or annotate steps is likely to be as effective. Either way, the convenience of this interface lies in the fact that no typing is necessary. Authentication

¹A working version of the IMAGINATION system, primarily meant as an experimental testbed, can be found at <http://riemann.ist.psu.edu/imagination/>.

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 creates an unambiguous list of 15 words, inclusive of the correct label, in a very simple manner. For this, we make use of a WordNet-based [17] word similarity measure proposed by Leacock and Chodorow [13]. 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 [7], but we found that for limited pools of words, this simpler strategy was equally effective. Furthermore, the WordNet-based similarity measures are only trustable locally.

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 [12]. However, waste of space is not an issue for us, nor are rectangles to pack rigid, i.e., linear stretching is allowed, with some limitations on the size and aspect ratio. We take a simple approach. We recursively partition the rectangular region (with constraints), and resize the chosen images to fit them in.

Analysis: The size of the tiled image in the click stage is fixed at 800×600 , and the tolerance radius r is pegged at 25 pixels, which corresponds 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 $(\frac{8\pi r^2}{800 \times 600} \frac{1}{15})^2$, or 0.00048%, 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, with the bottleneck being the latter step. For faster processing, a large set of distorted images over varied parameter settings can be pre-generated and stored.

IV. EXPERIMENTAL RESULTS

Experiments 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’. Recognition is considered failed if the latter is chosen, or if an incorrect label is chosen. The following summarizes recognizabilities



Fig. 3. A sample tiled image presented in the **Click** step of authentication. The tiled image is randomly partitioned orthogonally and dithered using different color sets. The user must click near the center of one of the images.

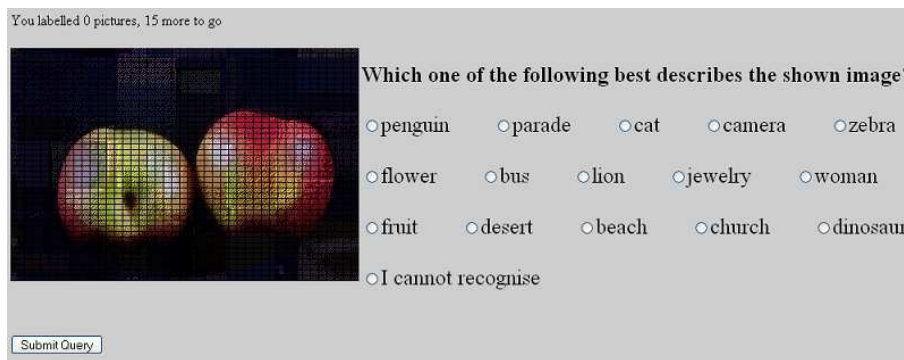


Fig. 4. Screenshot of the **Annotate** step, where a distorted image is presented, and the user must select an appropriate label from a list of choices.

of humans and machines, and their *recognizability gap*.

A. 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 These results are presented in Figures 5, 6, 7, and 8 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. 7 and 8). 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.

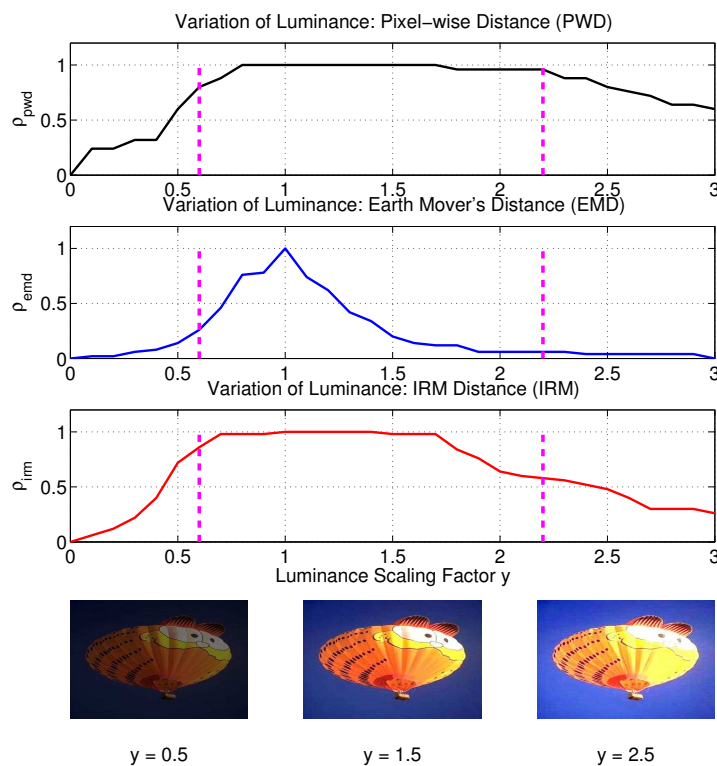


Fig. 5. Variation of average machine recognizability with change in luminance scaling factor. Human recognizability is high within the Magenta lines.

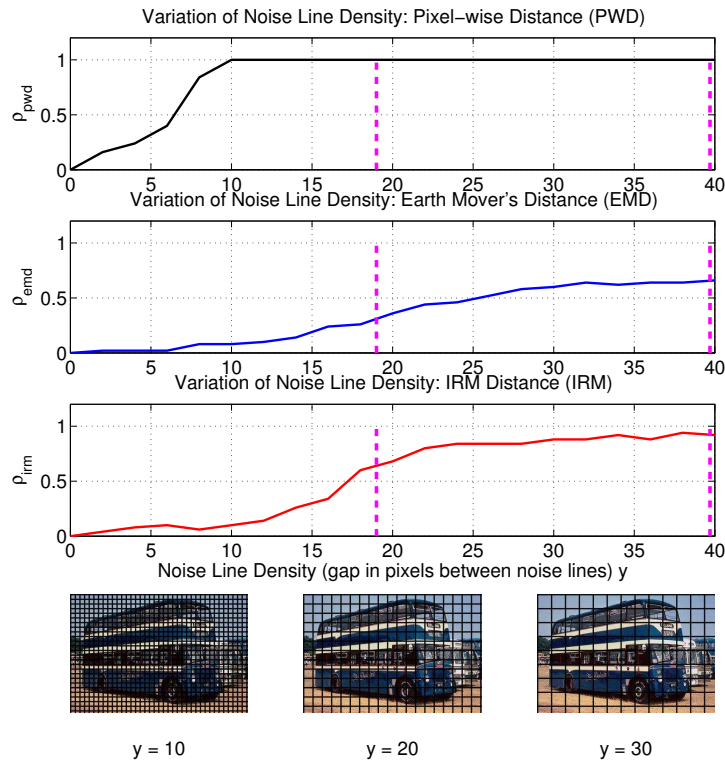


Fig. 6. Variation of average machine recognizability with change in density of noisy lines added, represented in pixels specifying the gap between consecutive lines. Human recognizability is high within the Magenta lines.

B. 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. 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. This step would require feedback from multiple users over a multitude of images.

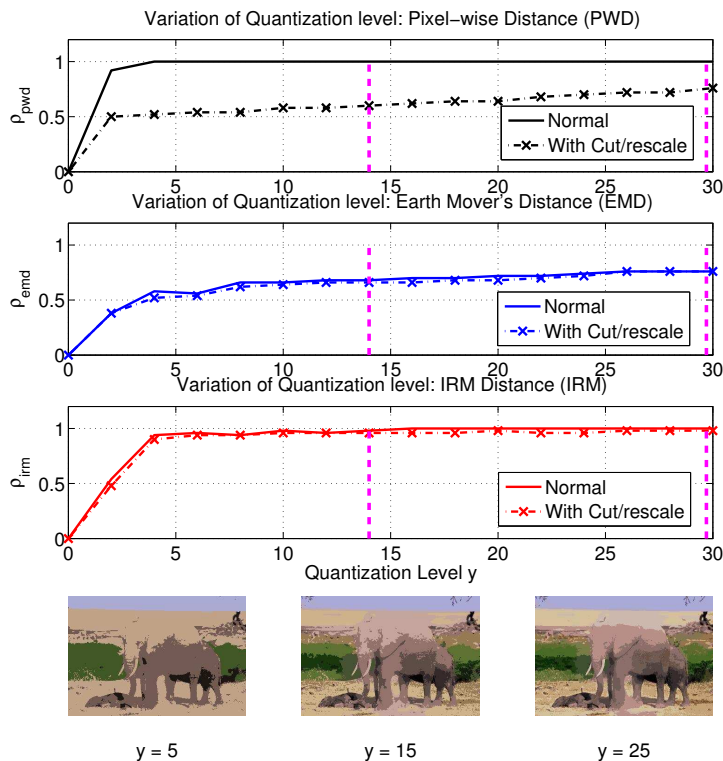


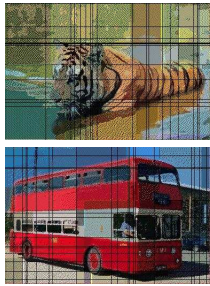
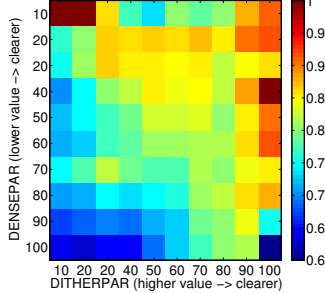
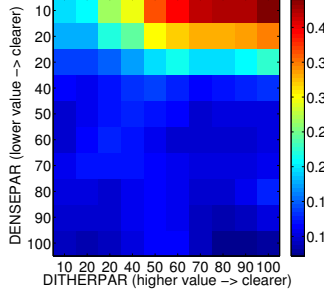
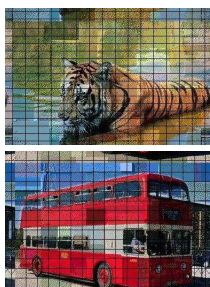
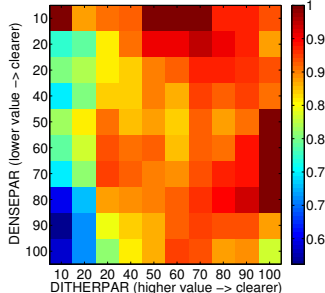
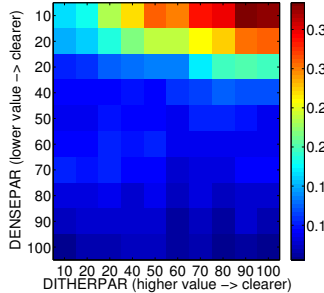
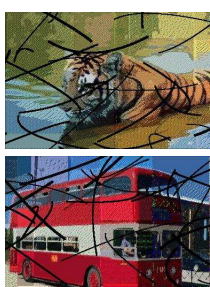
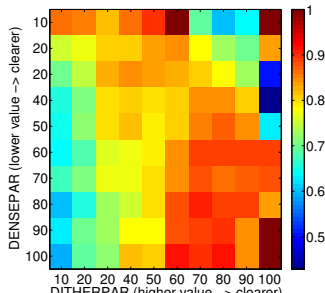
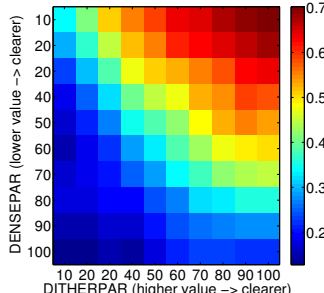
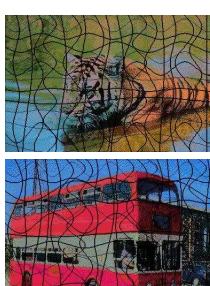
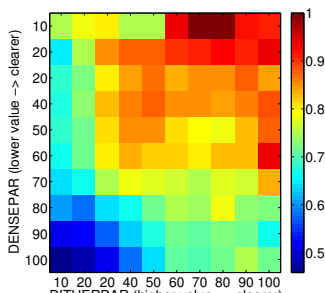
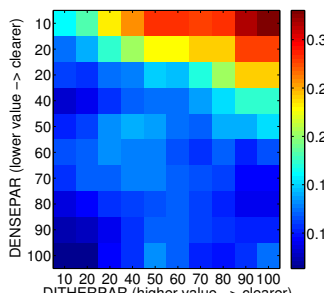
Fig. 7. Variation of average machine recognizability with change in quantization level, specified in terms of the number of color clusters generated and (centroids) used for mapping. Human recognizability is high within the Magenta lines.

- **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 Table II, 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. 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

TABLE II
FOUR DISTORTIONS THAT ARE PART OF THE IMAGINATION SYSTEM

Distortion Steps	Sample Images	Human Recognizability (User Study)	Machine Recognizability (PWD + EMD + IRM)
<ol style="list-style-type: none"> 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 <i>block</i> with an independently drawn random set of DITHERPAR colors. 5. Draw DENSEPAR lines parallel to each axis, <i>randomly spaced</i>. 6. Perform 10 – 20% cut/rescale on a randomly chosen side. 			
<ol style="list-style-type: none"> 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 <i>block</i> with an independently drawn random set of DITHERPAR colors. 5. Draw DENSEPAR lines parallel to each axis, <i>equally spaced</i>. 6. Perform 10 – 20% cut/rescale on a randomly chosen side. 			
<ol style="list-style-type: none"> 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 <i>block</i> with an independently drawn random set of DITHERPAR colors. 5. Draw DENSEPAR <i>third-order curves</i>, 1-3 pixels thick, <i>randomly positioned</i>. 6. Perform 10 – 20% cut/rescale on a randomly chosen side. 			
<ol style="list-style-type: none"> 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 <i>segment</i> with random set of DITHERPAR colors. 5. Draw DENSEPAR <i>sinusoids</i> with axes parallel to each axis, <i>randomly spaced</i>. 6. Perform 10 – 20% cut/rescale on a randomly chosen side. 			

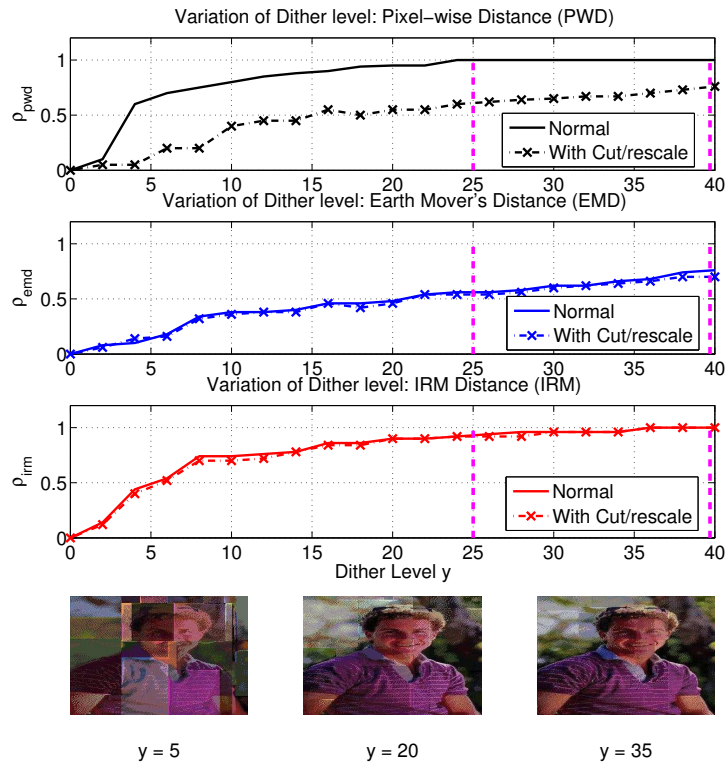


Fig. 8. Variation of average machine recognizability with change in dithering level, specified in terms of the number of colors available for dithering each partition. Human recognizability is high within the Magenta lines.

recognizability exceeds human recognizability are also interesting and worth exploring, but that is beyond the scope of this paper.

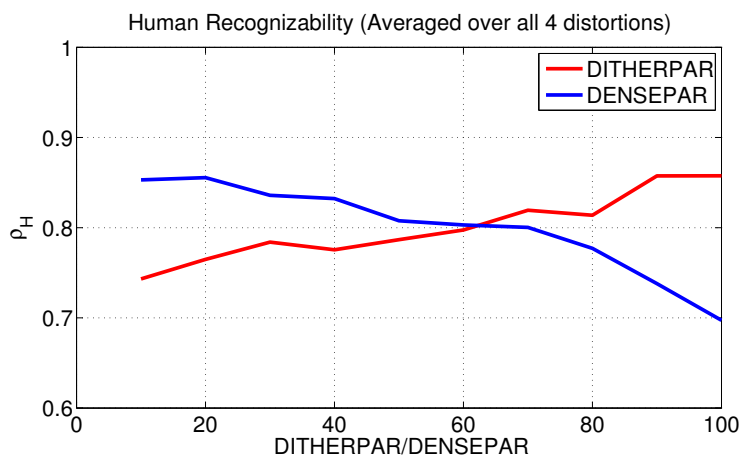


Fig. 9. Overall variation of human recognizability with dithering parameter DITHERPAR and noise density parameter DENSEPAR, taken across all four composite distortion methods.

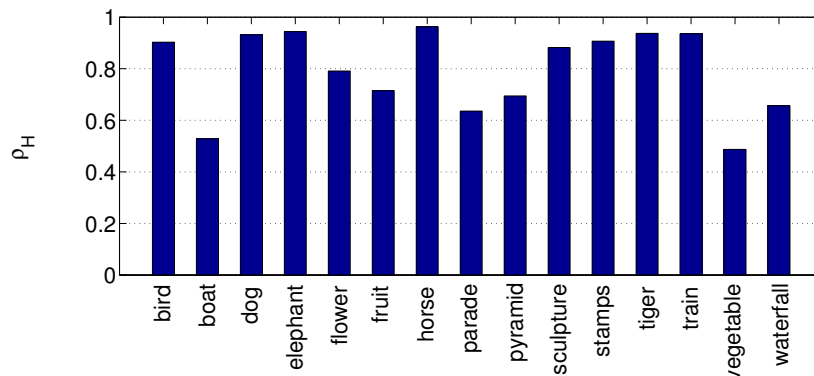


Fig. 10. Overall variation of human recognizability with the image concept, taken across all four composite distortion methods and their parameterizations. The fifteen most frequently sampled concepts are shown here.

To further the investigation and help design the IMAGINATION system better, we studied the trends of human recognizability from the user responses. Figure 9 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 10 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 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 remain 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.

V. 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 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. An understanding of the difference in recognizability of algorithms and humans under similar conditions provides an opportunity for better feature extraction design.

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] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image Retrieval: Ideas, Influences, and Trends of the New Age," *ACM Computing Surveys*, to appear.
- [7] R. Datta, J. Li, and J. Z. Wang, "IMAGINATION: A Robust Image-based CAPTCHA Generation System," *Proc. ACM Multimedia*, 2005.
- [8] J. Elson, J. R. Douceur, J. Howell, and J. Saul, "Asirra: A CAPTCHA that Exploits Interest-Aligned Manual Image Categorization," *Proc. ACM Conference on Computer and Communications Security*, 2007.
- [9] R.W. Floyd and L. Steinberg, "An Adaptive Algorithm for Spatial Grey Scale," *Proc. Society of Information Display*, 17:75-77, 1976.
- [10] A.K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Prentice Hall, 1988.
- [11] Y. Ke, R. Sukthankar, and L. Huston, "Efficient Near-duplicate Detection and Subimage Retrieval," *Proc. ACM Multimedia*, 2004.
- [12] R.E. Korf, "Optimal Rectangle Packing: New Results," *Proc. ICAPS*, 2004.
- [13] C. Leacock and M. Chodorow, "Combining Local Context and WordNet Similarity for Word Sense Identification," *Fellbaum*, 1998.
- [14] J. Li and J.Z. Wang, "Real-time Computerized Annotation of Pictures," *IEEE Trans. Pattern Analysis and Machine Intelligence*, to appear.
- [15] D.G. Lowe, "Object Recognition from Local Scale-invariant Features," *Proc. ICCV*, 1999.

- [16] C.L. Mallows, "A Note on Asymptotic Joint Normality," *Annals of Mathematical Statistics*, 43(2):508–515, 1972.
- [17] G. Miller, "WordNet: A Lexical Database for English," *Communications of the ACM*, 38(11):39-41, 1995.
- [18] 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 Conference on Computer and Communications Security*, 2003.
- [19] G. Mori and J. Malik, "Recognizing Objects in Adversarial Clutter: Breaking a Visual CAPTCHA," *Proc. IEEE CVPR*, 2003.
- [20] G. Moy, N. Jones, C. Harkless, and R. Potter, "Distortion Estimation Techniques in Solving Visual CAPTCHAs," *Proc. IEEE CVPR*, 2004.
- [21] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *International Journal of Computer Vision*, 40(2):99-121, 2000.
- [22] Y. Rui and Z. Liu, "ARTiFACIAL: Automated Reverse Turing Test using FACIAL Features," *Proc. ACM Multimedia*, 2003.
- [23] 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.
- [24] A. Thayananthan, B. Stenger, P. H. S. Torr, and R. Cipolla, "Shape Context and Chamfer Matching in Cluttered Scenes," *Proc. IEEE CVPR*, 2003.
- [25] A. Turing, "Computing Machinery and Intelligence," *Mind*, 59(236):433-460, 1950.
- [26] 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.
- [27] Slashdot, "Yahoo CAPTCHA Hacked", <http://it.slashdot.org/article.pl?sid=08/01/30/0037254>. Retrieved on 29th January, 2008.