Security through a different kind of obscurity: Evaluating Distortion in Graphical Authentication Schemes

Eiji Hayashi
ehayashi@cs.cmu.edu

Jason I. Hong
jasonh@cs.cmu.edu

Nicolas Christin
nicolasc@cmu.edu

Carnegie Mellon University
5000 Forbes, Pittsburgh, PA 15213, USA

ABSTRACT

While a large body of research on image-based authentication has focused on memorability, comparatively less attention has been paid to the new security challenges these schemes may introduce. Because images can convey more information than text, image-based authentication may be more vulnerable to educated guess attacks than passwords. In this paper, we evaluate the resilience of a recognition-based graphical authentication scheme using distorted images against two types of educated guess attacks through two user studies.

The first study, consisting of 30 participants, investigates whether distortion prevents educated guess attacks primarily based on information about individual users. The second study, using Amazon Mechanical Turk, investigates whether distortion mitigates the risk of educated guess attacks based on collective information about users. Our results show that authentication images without distortion are vulnerable to educated guess attacks, especially when information about the target is known, and that distortion makes authentication images more resilient against educated guess attacks.

Author Keywords

Educated Guess Attack, User Authentication, Graphical Authentication

ACM Classification Keywords

H.5.2 Information Interfaces and Presentation: Miscellaneous

INTRODUCTION

Knowledge-based authentication, where users authenticate by evidencing knowledge of a secret, remains the most pervasive type of authentication system because of its simplicity and availability on almost any type of platform. One of the fundamental assumptions in knowledge-based authentication is that users can memorize secrets, e.g., passwords, shared between the users and the authentication system. However, because people have difficulty in memorizing many complicated passwords, a typical strategy is to choose simple, easy-to-remember passwords and/or reuse a limited number of passwords for many accounts, all of which undermine the security of password authentication [1, 7]. Researchers have proposed alternatives to traditional text-based passwords with the hopes of improving memorability. One such approach is image-based authentication, which leverages research in cognitive psychology that shows that it is significantly easier for users to recognize images than to remember text [9, 18]. In that respect, image-based authentication schemes (e.g., graphical passwords) are a promising alternative to text-based passwords.

However, while image-based authentication facilitates memorability, it also poses new security challenges. One of these challenges is the *educated guess attack*, where an attacker tries to guess a user's shared secrets based on knowledge about that user. Because graphical authentication tokens (e.g., authentication images) preserve more contextual information than text-based passwords, image-based authentication could actually be *less* resilient against educated guess attacks than its text-based counterpart.

In this paper, we investigate the security of user chosen authentication images against two types of educated guess at-

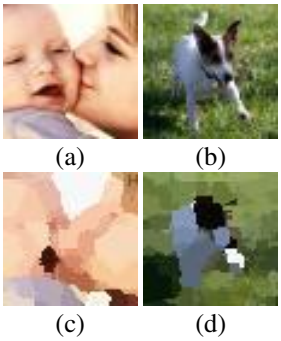


Figure 1. Authentication images, (a) and (b), can be vulnerable to educated guess attacks. For instance, if the attacker knows that users are likely to choose animal pictures as their authentication images, the attacker can guess that (b) is an authentication image (collective educated guess attack). If the attacker knows the target's wife, and sees her in (a), (a) is likely to be an authentication image (individualized educated guess attack). In contrast, if distorted pictures, (c) and (d), are used, these attacks are unlikely to work because the contextual information of the distorted pictures is obfuscated.

CHI 2011, May 7–12, 2011, Vancouver, BC, Canada.

tacks by emulating these attacks in two user studies. We also evaluate the marginal security gained against these attacks by performing image distortion, which was proposed in *Use Your Illusion* [10].

Through this investigation, we make novel contributions. First, we provide quantitative, empirical results that show that an attacker can make very accurate guesses about a user's authentication images if the attacker possesses enough information about the user. Second, we demonstrate that distorting images can mitigate educated guess attacks against recognition-based graphical authentication systems.

The rest of this paper is organized as follows. First, we discuss the two types of educated guess attacks and related work. Second, we formulate our research hypotheses, which we test in the two user studies. Third, we describe our first user study, in which we emulate educated guess attacks by asking pairs of "friends" to guess their friend's authentication image(s). Fourth, we describe our second user study using Amazon Mechanical Turk. The study evaluates whether distortion can prevent the educated guess attacks based on collective information about users. Finally, we discuss the implications of our findings.

EDUCATED GUESS ATTACKS

In educated guess attacks, an attacker tries to guess a user's shared secret (e.g., a password) based on the user's information. Most educated guess attacks can be categorized into two types, *collective educated guess attacks* and *individualized educated guess attacks* (Figure 1). The collective educated guess attack rely on knowledge about the entire population of users, while the individualized educated guess attacks rely on knowledge specific to a given user.

Collective educated guess

A lot of past work has investigated collective educated guess attacks on graphical authentication schemes [14, 21–24]. The collective educated guess attacks rely on knowledge about the entire population of users. For instance, in the context of recognition-based schemes (e.g., Use Your Illusion), if an attacker knows that most people prefer to select pictures of animals as their authentication images, the attacker could infer that pictures of animals in the challenge set (Figure 1 (b)) have a higher probability of being authentication images.

Individualized educated guess

The individualized educated guess attacks rely on knowledge about a specific user. For instance, if an attacker knows that a user has a wife and a baby and finds a picture showing a woman and a baby in a challenge set, the attacker can guess that the picture has a high probability of being one of the authentication images (Figure 1 (a)).

As described above, there has been a great deal of past work investigating collective educated guess attacks. However, there has been little work on individualized educated guess attacks against graphical authentication schemes. As such, one of the primary contributions of this paper is to quantify the risk of the individualized educated guess attacks on recognition-based graphical authentication schemes.

RELATED WORK

Based on the observation that humans are considerably better at remembering images than they are at remembering text [18], much work has been devoted to investigating graphical password authentication (e.g., [5, 11, 16]). Evaluations of these techniques have shown that graphical password authentication can achieve higher memorability, but also have some security drawbacks. We outline various schemes below, and discuss some of their strengths and weaknesses.

Recall-based schemes

Recall is one of the cognitive processes utilized in graphical authentication schemes. Recall is the ability to remember items from memory without help. For example the Draw-A-Secret scheme requires users to draw a pre-determined image on a grid in order to authenticate [11]. The drawings can provide a larger password space, and more memorable passwords than a text-based password does; however, users have a tendency to choose drawings with high symmetry and a small number of strokes [21, 22], which are analogous to choosing simple passwords in text-based authentication. An attacker can thus use collective educated guess attacks, i.e., an attacker do better than randomly guessing by trying symmetric drawings based on the knowledge that users *in general* are likely to choose specific types of drawing.

Cued-recall-based schemes

Other authentication systems rely on *cued recall*, where retrieval cues are provided to a user to aide the recall task. As an example, in the PassPoints graphical password scheme, users authenticate by clicking on points, or selecting regions, of an image that were previously chosen by the user [2, 16, 24]. In this case, the image itself serves as a cue to the regions of the image that a user must recall. This scheme however produces predictable authentication sequences. This is because users tend to choose "hot spots," i.e., regions that are often selected because they are most memorable or most obvious [23, 24]. As a result, an attacker is more likely to succeed by guessing hot spots rather than all possible positions of clicks. Again, biases common to many users biases let attackers launch collective educated guess attacks against PassPoints.

Recognition-based schemes

There is another class of graphical authentication schemes that rely on *recognition*, the ability to judge whether a person has seen an item before or not. Past work in cognitive psychology has demonstrated that humans have an impressive ability to recognize pictures they have seen before [18–20]. This body of work has also shown that they recognize pictures better than they recognize words or sentences [18].

As such, recognition is used in many graphical password schemes by asking users to select previously chosen images from a larger subset of images [4, 5].

In selecting previously chosen images, people remember images more accurately when they are contextually meaning-

ful and when the images are generated by the individuals themselves [12]. Hence, graphical password images selected by users tend to be more memorable than assigned images. For that reason, many recognition-based graphical authentication schemes let users choose their authentication images.

For example, in PassFaces [4], users authenticate by identifying a set of faces they have previously chosen. An evaluation of PassFaces showed such authentication images are much more memorable than text-based passwords [4]. However, users are also more likely to choose faces of females belonging to their ethnic group [14], which makes authentication predictable and thus less secure. This bias toward specific faces allows an attacker to guess a user's authentication images based on the user's profile.

THWARTING EDUCATED GUESS ATTACKS

One countermeasure against educated guess attacks is to randomly assign authentication images to users, or to use abstract images [5], which are less predictable than real images. However, past work has shown that image scenes that are coherent and contextually meaningful are memorized more accurately than incoherent or abstract images [3, 8]. For instance, a picture taken on a vacation trip will be easier to memorize than randomly assigned scenery, or abstract images. As such, random assignment or use of abstract images can reduce the memorability of the authentication images.

Another alternative is to have users self-select images and assign distorted versions of these images to authenticate. Use Your Illusion [10] requires users to produce three images to authenticate. These images are assigned as authentication images after being distorted through an "oil-painting" filter, which preserves rough shape and colors, but eliminates most details. When users want to be authenticated, they are asked to identify their three authentication images from a set of 27 distorted images. Use Your Illusion relies on a finding from cognitive psychology that people's interpretations of visual input are affected by their expectations [6]. Users expect to find distorted images representing contextual information of their original photos; thus they interpret the distorted images as images with the contextual information. However, an attacker is far less likely to have these expectations, and therefore should have a much more difficult time interpreting the contextual information of the distorted images. The work [10] demonstrated that users could recognize their authentication images accurately even one month later. Nevertheless, the marginal security created by the distortion technique has not been investigated yet. As such, a secondary contribution of this paper is to evaluate the marginal security obtained by distorting images in quantitative ways.

HYPOTHESES

At a high level, we hypothesize that educated guess attacks will let an attacker guess user-chosen authentication images with a higher success rate than random guesses. In addition, we hypothesize that *Use Your Illusion* [10] will increase the security of the authentication images against educated guess attacks. In the rest of this section, we formally define five hypotheses that we tested in our user studies.

H1 (context effect): *If a recognition-based graphical authentication system lets users choose original, undistorted pictures as authentication images, an attacker can predict the images more accurately by using educated guesses than by guessing randomly.*

In H1, we hypothesize that an attacker can launch collective educated guess attacks against authentication images. We argue that original, undistorted pictures will improve an attackers ability to make guesses better than by chance, even without knowing anything about the specific user.

H2 (knowledge effect): *An attacker can make more accurate guesses about authentication images if the attacker possesses information about the user who chose them.*

In H2, we hypothesize that an attacker can launch individualized educated guess attacks against authentication images. Authentication images have more contextual information than text-based authentication images. The additional information helps users to memorize the authentication images; however, at the same time, the additional information also helps attackers guess the authentication images more accurately given user preferences. For instance, if the attacker knows that her target is a car fanatic and finds photos of cars in the challenge set, the attacker can infer that those photos have a higher likelihood to be authentication images.

H3 (distortion effect on educated guesses): *Distorted authentication images are significantly more resilient against educated guess attacks than original, undistorted authentication images.*

In H3, we hypothesize that distortion improves the security of authentication images against educated guess attacks. We argue that distorted authentication images are more difficult to interpret without knowing the original images, as distortion obfuscates the contextual information of authentication images. As a result, an attacker should have more difficulty in guessing authentication images based on the contextual information.

H4 (user biases): *Users tend to choose specific categories of images as their authentication images.*

In H4, we hypothesize that there are biases toward certain types of images among authentication images chosen by users. For instance, people could tend to choose pictures of faces because they are good at memorizing faces. Observing biases toward certain types of images among authentication images chosen by users, the attacker can make better guesses by choosing frequently chosen types of images. That is, the attacker can launch the collective educated guess attacks against authentication images. According to existing work [14, 23, 24], H4 is highly likely to be supported. However, we still want to test H4 as a baseline for the next hypothesis.

H5 (biases after distortion): *Even when authentication images are distorted, attackers still can make better guesses based on users' biases in choosing authentication images.*

In H5, we hypothesize that, even after distortion, attackers can launch collective educated guess attacks. This hypothesis is supported when, for instance, distorted images which seem like *people* are more likely to be authentication images than distractors.

USER STUDIES

We conducted two user studies to test the hypotheses defined in the previous section.

In the first user study, we tested H1 (context effect), H2 (knowledge effect), and H3 (distortion effect on educated guesses). We recruited pairs of friends as participants. We asked participants to guess their friends' authentication images. In other words, we asked participants to act as attackers and to guess authentication images of targets whom they know well. Based on the results, we evaluated whether knowing the target well helps attackers to make more accurate guesses about the user's authentication images.

In the second user study, we tested H4 (user biases) and H5 (biases after distortion). We used Amazon's Mechanical Turk to organize images into 12 categories according to their contextual information. Mechanical Turk is a web service that coordinates many users working on tasks that require human processing, such as image tagging tasks [13]. Our participants organized two sets of photos. The first set consisted of 120 photos randomly selected from Flickr.com. The other set consisted of 180 photos chosen as authentication images by the participants in the first user study. We asked the participants to categorize these two sets of photos and distorted versions of them. Based on the categorizations, we evaluated whether attackers could make collective educated guesses effectively.

USER STUDY #1: EVALUATION WITH PAIRS OF FRIENDS

For our first user study, we evaluated the security of authentication images against both individualized and collective educated guess attacks using *Use Your Illusion* as a platform. We asked participants to act as attackers and to guess their targets' authentication images. We evaluated the security based on the number of correct guesses that participants made.

Participants

We recruited pairs of friends as participants from our university. We defined "friend pair" as a pair of two persons which satisfied the following four conditions: 1) both of them had Facebook accounts; 2) both of them had "friended" each other in Facebook; 3) they met face-to-face at least twice a week; and 4) they had known each other for at least three months. In total, we recruited 30 university students (15 pairs). Nineteen participants were male and 11 participants were female. Their ages ranged from 20 to 28 with a mean age of 22.6 years old. According to a post experiment survey, the pairs had known each other for three months to six years with a mean length of 19.4 months. Also according to the post experiment survey, the pairs met each other 6.3 days a week on average.

We compensated the participants $10 USD for their participation. In addition, we paid an additional $5 USD for each correct guess they made about their target's authentication image. This additional payment gave participants an incentive to make their best guesses about their target's authentication images.

Our participants did not have deep knowledge of authentication systems. However, *Use Your Illusion* and its distortion technique are simple enough for them to understand. At first glance, it appears that the system should be more vulnerable to security experts with knowledge of the system's internals. However, for the educated guess attacks, which are essentially variants of social engineering attacks, we contend that knowledge of the target is more critical to the attack's success than technical expertise. For this reason, we asked friends rather than security experts to pose as attackers in this study.

Variables

For this user study, we had two independent variables: type of image and relationship between the attacker and target. The type of image variable was either original photo and distorted photo. In the rest of this paper, we refer to these conditions as *original* condition and *distorted* condition. Relationship between attacker and target was either friend or stranger. We refer to them as *friend* condition and *stranger* condition respectively. In the friend condition, participants were asked to guess images chosen by their partners. While, in the stranger condition, participants were asked to guess images chosen by another set of 15 participants, living in a different city, and who were not part of this experiment. This selection allowed us to avoid attackers guessing authentication images by choosing images taken in local areas, such as neighboring landmarks.

We conducted this user study in a 2×2 mixed design (Figure 1). We used a within-subject design for type of images variable, i.e., all participants tested both image types, because the amount of knowledge that an attacker had about his target varies from attacker to attacker. If an attacker tested both image types, we could minimize the effect of the variation. For relationship, we used a between-subject design, i.e., in each pair, one of the pair was chosen randomly and assigned to a friend group and the other was assigned to a stranger group. If we used within-subject design for relationship as well, participants must have tested four conditions, which could cause ordering effects, such as fatigue. Thus, to minimize the ordering effects, we used between-subject design for relationship. In other words, participants in the friend group and in the stranger group only tested the friend condition or the stranger condition, but not both (Table 1). The dependent variable of this user study is the success rates of the educated guess attacks. Using this dependent variable, lower success rates mean higher resilience to educated guess attacks.

Hypotheses

We tested H1 (context effect), H2 (knowledge effect) and H3 (distortion effect on educated guesses). We conducted

Image Type \ Relationships	Friend	Stranger
Original	2× 2 mixed design	
Distorted		

Table 1. The first user study was conducted in a 2×2 mixed design. We assigned participants to either the friend group or the stranger group. Among four combinations of conditions, participants in the friend condition tested both friend-original condition and friend-distorted condition. Likewise, participants in the stranger group tested both stranger-original condition and stranger-distorted condition.

the tests by comparing number of correct guesses, using a binomial test and a Fisher's exact test.

For H1 (context effect), we compared the number of correct guesses in the original condition and the expected number of correct guesses, assuming that a participant made guesses randomly. H1 was supported if there was a statistically significant difference between these two numbers

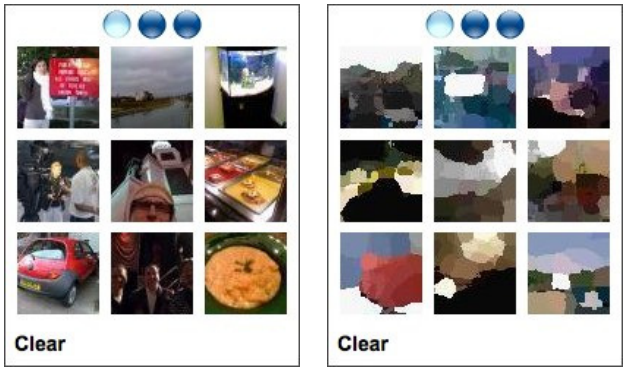
For H2 (knowledge effect), we compared the number of correct guesses in the friend condition with that in the stranger condition. In the friend condition, participants had information about their targets, whereas in the stranger condition, participants did not have information about their target. H2 was supported if there was statistically significant difference between these two numbers.

Finally, for H3 (distortion effect on educated guesses), we compared the number of correct guesses in the original condition with that in the distorted condition. H3 was supported if there was a statistically significant difference between these two numbers.

Evaluation Environment

We emulated Use Your Illusion [10] on a desktop computer (Figure 3). Images used in the emulator had the same resolution as those used in Use Your Illusion. Use Your Illusion has a setup phase and a challenge phase. During setup, the user is asked to produce three pictures as his authentication images, Then, during authentication, the system presents the user with a challenge set consisting of 27 images, including the three authentication images. The user is asked to correctly identify the three authentication images from the challenge set within three trials. The 27 images are shown as the three sets of nine images (Figure 2). However, it does not mean each set contain one authentication image. Each set can contain zero to three authentication images. When choosing the images, the user can go back and forth among the three sets.

In the original condition, all 27 images, including user-chosen images and distractors, were original, non-distorted images. In the distorted condition, all images were distorted using an oil-painting image processing filter. The filter distorts precise shape and minor colors while keeping rough shapes and major colors. Previous research showed that users can memorize the distorted images as accurately as they memorize original images for at least one month [10].



(a) Original (b) Distorted

Figure 2. Example of challenge sets. A challenge set consists of three screens showing nine images (27 images in total). The examples show one of the three screens in the original condition and the distorted condition. A user has to select 3 images correctly from the 27 images to be authenticated. Distractors are chosen from an image pool consisting of 144 images.

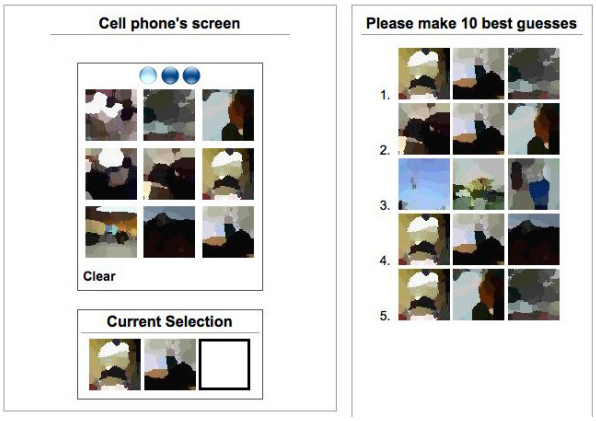


Figure 3. Use Your Illusion emulator. The top left part emulated Use Your Illusion. The right part showed combinations of images already chosen by a participant. The bottom left part showed images which a participant was currently choosing.

Procedure

This user study consisted of three phases: preparation, an experiment and a post experiment survey. The entire user study took about 40 minutes.

Preparation

Prior to our user study, we asked all participants to submit six digital photos taken by themselves. The participants could submit photos that they had already taken or took specifically for this study. In the request, we explicitly mentioned that the photos would be used as "passwords" in a graphical authentication system. Moreover, we asked participants not to talk with others about the photos they selected. Three of the photos were randomly selected and distorted using an oil-painting filter, for use in the distorted condition. The other three photos were used as is for the original condition.

In our post-survey, we asked when the participants took the photos. One participant responded he took his photos after

we started this study. Other participants answered that they took their photos at least one month before we started this study. This indicated that most participants submitted photos that they had already taken.

Experiment

We conducted this experiment in an isolated room under the supervision of an experimenter. Each participant used a Use Your Illusion emulator installed on the experimenter's laptop to guess his target's authentication images.

When a participant tried to guess his target's (i.e., his friend's or stranger's) authentication images, two sets of images were given, one at a time. One of them consisted of original photos (i.e., original condition) and the other consisted of distorted photos (i.e., distorted condition). Each set consisted of three authentication images and 24 distractors. The distractors were chosen from a pool consisting of 144 images. The images in the pool were randomly chosen from photos under Creative Commons License in Flicker.com. All of the sets were manually composed ahead of time to achieve balance, and to make the two sets given to a participants mutually exclusive. Thus, each individual saw different distractors in the two conditions. The images were resized to 56×56 pixels to be used in Use Your Illusion emulator.

In the original photo set, photos taken by the target were used as authentication images after resizing. In the distorted photo set, photos taken by the target were used after resizing and distortion. The photos used as authentication images or source of authentication images were mutually exclusive.

We randomly chose half of the participants and let them start with a set of original photos, while the other half of the participants started with a set of distorted photos. After we gave one of the sets, we asked the participant to make their 10 best guesses (i.e., to choose 10 sets of three photos.) The experimenter did not give feedback whether their guesses were correct or not. After 10 guesses were made in the first set of photos, we gave another set of 27 images. Then, participants were asked to make 10 best guesses for this set.

Post Experiment Survey

After the experiment, we asked participants to answer a post experiment survey. The survey polled participants' profile, intimacy with their friends, their photo sharing through web sites and their strategy of making guesses. The survey consisted of 70 questions and took 15 minutes to complete.

Results

We considered attackers to be successful if they could correctly guess their targets' authentication images within ten trials. In Use Your Illusion, the number of trials was limited to three. However, considering the difficulty of the guessing task, we allowed attackers to choose ten different combinations of three images, so as to have more statistical power. Among these ten guesses, an attacker could not choose the exact same combination twice. Assuming that an attacker chose images completely randomly, the probability that the correct combination was included in the ten selected combinations was $P = 10/\begin{pmatrix} 27 \\ 3 \end{pmatrix} \approx 0.003$. We used the success rate of random guess denoted by P as a baseline in the following analysis.

Table 2 shows the success rates of the attackers in each condition. Asterisks "*" in Table 2 denote statistical significance when the success rate is greater than the baseline P at significance level $p = 0.01$ based on a binomial test. The numbers in parentheses are the number of attackers who successfully guessed their targets' authentication images.

	Friend	Stranger
Original	0.53* (8/15)	0.20* (3/15)
Distorted	0.067 (1/15)	0.00 (0/15)

Table 2. Success Rates of Educated Guesses. Asterisks "*" denote that the success rates were higher (statistically significant) than success rate of random guesses. The success rate of random guesses is 0.003. Numbers in parentheses denote the number of attackers (out of 15) who made correct guesses.

As Table 2 shows, in the original conditions, the success rates were higher than the baseline ($p < 0.01$). Therefore, H1 (context effect) was supported.

These results showed that participants could guess better than chance even without any knowledge about their targets. This finding suggested that contextual information in the undistorted images themselves could help an attacker make better guesses.

In the friend-original condition, eight attackers out of 15 correctly guessed targets' authentication images. In addition, three attackers chose correct combinations in the first attempt. One of the eight attackers answered that his strategy was "[to] try to think of places my friend has been and guess those pictures." Other attackers also used similar strategies to make their guesses. These results strongly suggested that using self-selected original photos for authentication was highly vulnerable to individualized educated guess attacks. Thus, H2 (knowledge effect) was supported.

In the stranger-original condition, where attackers did not know their targets, three attackers correctly guessed their targets' authentication images. According to post surveys, these attackers chose images that had common properties. One of the three attackers said that "[he] selected the 3 photos depicting the similar content." His target's authentication images consisted of two photos of a beer can and one photo of a bottle of alcohol.

In contrast, in the distorted conditions, there was no statistical significance between the success rates of attackers and random guesses ($p > 0.01$). Moreover, in the post-survey, most of the attackers answered that they relied on random guess. Moreover, using Fisher's exact test, the success rate in friend-distorted condition was lower than in friend-original condition ($p = 0.007 < 0.01$). Thus, H3 (distortion effect) was supported.

Figure 4 shows distributions of partially correct guesses, which

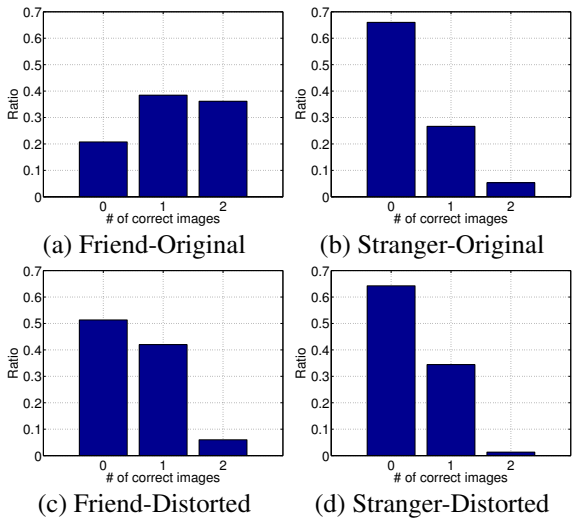


Figure 4. Partially correct guesses, showing number of correctly guessed authentication images. For example, in Friend-Original (a), 21% of all guesses by all participants in that condition had zero of the three correct images, 38% had only one of the three images, and about 36% had two images (i.e., a very close guess). The distributions were skewed to the right if participants had more accurate partially correct guesses.

included zero, one or two authentication images. Distributions skewed to right mean that attacker made more accurate guesses. The ratios were calculated by dividing the number of partially correct guesses by the total number of guesses. As described above, in the friend-original condition, there were two attackers who chose correct combinations of three images in their first choice. According to the post survey, they were completely sure that they chose the correct combinations and that they chose images randomly for their other nine guesses to save their effort. Thus, in the analysis of partial hits, we discarded their data. Differences among the distributions in Figure 4 were statistically significant except for the difference between the stranger-original condition and the stranger-distorted condition (Table 3).

In the friend-original condition (Figure 4 (a)), the distribution was skewed to the right. This illustrates that the ratios of partial hits, which included more authentication images, are higher than other conditions. It indicates that using original images as authentication images was more vulnerable to the educated guess attack. In friend-distorted condition (see Figure 4 (c)), the ratio of guesses including one authentication image was higher than that in stranger conditions (see Figure 4 (b) and (d)). This may indicate that, even in the case where an authentication system uses distorted images, attackers have an advantage if they make guesses based on the knowledge they have about their targets. However, the effect would be small because we did not observe statistical evidence that attackers could predict authentication image better than random guesses in friend-distorted condition.

USER STUDY #2: CATEGORIZATION

In the second user study, we further evaluated the risk of the collective educated guess attacks against authentication images by testing H4 (user biases) and H5 (biases after distortion) using Amazon Mechanical Turk. We asked Amazon Mechanical Turk users to organize images into categories. Then, we compared distribution of the categorization between images chosen by users as their authentication images in the previous study, and the images randomly chosen from Flickr. If the distributions were different, attackers could launch collective educated guess attacks with knowing the distributions. If they were similar, collective educated guess attacks were not feasible even knowing the distributions.

This study consisted of two parts. In the calibration study, we asked the Mechanical Turk users to label images based on their contextual meanings in an open format. The purpose of this part was to investigate what labels they used to describe images. Then, we coded the labels to categories used in the actual study. In the actual study, we asked the Mechanical Turk users to choose one of the categories for each image based on its contextual meaning.

Calibration Study

We randomly collected 50 photos under Creative Commons license from Flickr. Then, we asked Amazon Mechanical Turk users to label each photo with one tag describing the color in the photo, and three tags describing the content in that photo. We gave participants four guidelines: 1) the color tag must denote the dominant color in the photo; 2) the content tags must describe the contents of the photo or some relevant context; 3) the content tags should include at least one noun; and 4) no tag can be longer than 25 characters. For each photo, we asked three users to label it. As a result, we obtained 450 content tags in total for 50 photos. We asked users to add the color tag to prevent the users from writing names of colors in content tags. We paid each user $0.03 USD for tagging a photo. The task was completed in five hours by 40 unique users.

	Friend-distorted	Stranger-original	Stranger-distorted
Friend-original	$p < 0.01*$ $\chi^2(2, 273) = 49.4$	$p < 0.01*$ $\chi^2(2, 271) = 68.5$	$p < 0.01*$ $\chi^2(2, 275) = 79.0$
Friend-distorted		$p = 0.02 < 0.05*$ $\chi^2(2, 276) = 7.9$	$p = 0.02 < 0.05*$ $\chi^2(2, 300) = 7.8$
Stranger-original			$p = 0.07 > 0.05$ $\chi^2(2, 298) = 5.1$

Table 3. Results of χ^2 test among distributions of partial hit rates in the four conditions. Except for a combination of the stranger-original condition and the stranger-distorted condition, the differences are statistically significant.

The content tags were very diverse; thus, we coded them into more abstract concepts to make the categorization task in the actual study easier. Finally, as a result of the calibration, we defined 12 concepts, which were used as categories in the following study. (Table 4).

#	Category	#	Category
1	Human	7	Plant
2	Transportation	8	Building
3	Animal	9	Food
4	Insect	10	Clothing
5	Interior	11	Object
6	Landscape	12	None of them

Table 4. Category tags. One of the author coded 450 content tags collected in the calibration study.

Method

To test H4 (user biases), we compared two sets of original images. The first set is a set of images chosen by participants as authentication images in the first user study. The second set is a set of images randomly chosen from Flickr. We collected another 120 photos from Flickr using the Perl script. Because we could not prevent users who participated in the calibration study from participating the actual study, we used another set of photos to avoid possible contamination. Ideally, as the second set, we wanted to use a set of all the pictures which people in general took. However, such a set of images was not available. Therefore, we approximated the set by choosing images randomly from Flickr. By comparing distributions of categories in these two sets of images, we could test whether users were likely to choose specific categories of images as their authentication images.

To test H5, we tested distorted versions of the two sets of original images. We evaluated how distortion affects categorization of images by comparing categorization of the original and distorted images. Consequently, we used four sets of images in total:

- original baseline: 120 images randomly chosen from Flickr.com
- original user chosen images: 180 images chosen as authentication images by participants of the first user study
- distorted baseline: 120 distorted images generated from *original randomly chosen images*
- distorted user chosen images: 180 distorted images generated from *original user chosen images*

For each of these 600 images, we asked five Mechanical Turk users to choose one of the 12 categories shown in Table 4. For the distorted images, using examples, we explained that the distorted images were generated from original photos using an image processing filter. Then, we asked the users to guess the original photos of the given distorted images, and to choose one category according to their guesses. As a result, we obtained 3000 categorizations (five categorizations for each of the 600 images). This categorization task was completed by 288 unique Mechanical Turk users.

Results

Table 5 shows the ratios of category tags given to the original images. The ratio $p_i, (i = 1, 2, 3..., 12)$ is calculated as $p_i = c_i/N$ where c_i and N denote the number of ith category's tags and the total number of tags. For instance, the number in the top left cell is calculated as $153/600 = 0.26$.

Category	**Baseline**		**Users' choice**	
Human	0.26	(153)	0.26	(251)
Transportation	0.06	(38)	0.02	(21)
Animal	0.07	(39)	0.05	(38)
Insect	0.00	(0)	0.00	(1)
Interior	0.10	(60)	0.04	(35)
Landscape	0.10	(58)	0.15	(127)
Plant	0.01	(3)	0.03	(24)
Building	0.11	(68)	0.25	(218)
Food	0.09	(51)	0.04	(34)
Clothing	0.00	(2)	0.00	(1)
Object	0.17	(101)	0.13	(122)
None of them	0.04	(27)	0.03	(28)
Total	1.00	(600)	1.00	(900)

Table 5. Distribution of categories given to the original images. The numbers in parentheses stand for actual number of images in each category. The difference between the distributions of categories in the baseline and the users' choice was significant, $\chi^2(11, 1500) = 106.64, p < 0.01$.

In Table 5, the numbers in the baseline column shows categorization of photos people *take*. On the other hand, the numbers in the users' choice column shows categorization of photos participants *choose* as their authentication images. The differences between numbers in these two columns indicates participants' tendency choose or not to choose categories of images as their authentication images. Attackers can launch collective educated guess attacks by choosing the categories where the numbers for the users' choice are greater than the numbers for the baseline.

In Table 5, the difference between the distributions of categories in the baseline and the users' choice was significant, $\chi^2(11, 1500) = 106.64, p < 0.01$. Table 5 shows that the participants were less likely to choose photos of transportation and food as their authentication images. On the other hand, they were more likely to choose photos of landscapes, plants and buildings as their authentication images. This results indicates that there are biases in the user-chosen authentication images. Thus, H4 (user biases) was supported.

Participants showed stronger biases toward landscapes and buildings. One possible interpretation of these biases was that participants thought that landscape and building photos would be appropriate for their authentication images because anyone could take these photos and they seemed to be less related to their identities.

Although more studies using different sets of images would be necessary, this observation implies that an attacker can launch collective educated guess attacks based on the knowledge that users are likely to choose specific types of authentication images. In this specific case, if an attacker chooses

landscape and building photos, the attacker would have better chance of guessing a user's authentication images than a random guess.

In contrast, when images were distorted (Table 6), there was no statistically significant difference between the distributions of categories in the baseline and the user chosen images, $\chi^2(11, 1500) = 18.00, p = 0.08$. Therefore, H5 (biases after distortion) was not supported. This indicates that after distortion, attackers do no have a statistically significant advantage in the collective educated guess attack even with knowing the biases shown in Table 6.

Category	Baseline		Users' choice	
Human	0.20	(120)	0.18	(166)
Transportation	0.04	(24)	0.02	(20)
Animal	0.10	(58)	0.09	(83)
Insect	0.00	(1)	0.00	(3)
Interior	0.10	(57)	0.10	(96)
Landscape	0.36	(216)	0.36	(316)
Plant	0.04	(21)	0.03	(31)
Building	0.06	(33)	0.09	(84)
Food	0.00	(4)	0.01	(14)
Clothing	0.01	(5)	0.01	(12)
Object	0.09	(54)	0.07	(61)
None of them	0.01	(7)	0.01	(14)
Total	1.00	(600)	1.00	(900)

Table 6. Distribution of categories given to distorted images. The numbers in parentheses stand for actual number of category tags given. There was no statistically significant differences between distribution of categories in the baseline and the users' choice.

DISCUSSION

Our user studies support H1 to H4, and reject H5. The first study shows that an attacker can make guesses better than random guess when authentication images are not distorted (H1: context effect), especially when the attacker knows about a target user (H2: knowledge effect). Furthermore, the study shows that using distorted images prevents these two attacks (H3: distortion effect on educated guesses). The second study illustrated that there are biased in choosing authentication images (H4: user biases), and that, with distortion, attackers have little advantage even if they know the biases (H5: biases after distortion).

These results yield a number of interesting implications. First, contextual information helps attackers as well as users. Balancing the two competing goals of resilience against attacks with ease of memorization is a critical design goal for any image-based authentication system. Our work in this paper helps quantify how well educated guess attacks can work, and offer a baseline for other image-based authentication systems.

Second, our results show that distortion is an effective technique for mitigating both individualized and collective educated guess attacks. Thus, it may be possible to apply the distortion technique to other recognition-based graphical authentication schemes. If an authentication scheme uses limited kinds of images (e.g., faces in PassFaces), users would have difficulty in recognizing their authentication images after distortion; however, if a scheme uses a wide variety of images, we could distort the images to improve their resilience against individualized educated guess attack and collective educated guess attack.

Third, we found that individualized educated guess attacks against undistorted authentication images were considerably more successful than collective educated guess attacks. To launch individualized educated guess attacks, attackers must collect adequate amount of information about their targets. However, recently, many people make personal information, such as preferences, recent activities, and personal histories, public through social networks, such as Facebook [15, 17]. As a result, it is becoming easier for attackers to collect such information. Thus, individualized educated guess attack can be one of the major threat against image-based authentication systems.

LIMITATIONS

This paper is not, and does not aim to be, a comprehensive security analysis of graphical password schemes. Instead we evaluated how a specific countermeasure (image distortion) provides resilience to a specific family of social engineering attacks (educated guess attacks). Studying resilience to other attacks, such as shoulder surfing, would require further work. Likewise, we focused on a specific authentication scheme (Use Your Illusion), and we need to caution against directly extrapolating to other graphical password schemes. Formally generalizing our results would likely require additional experimentation with a myriad of other graphical password schemes. We nevertheless believe our results are a useful contribution, in that neither security research nor cognitive science work has investigated the impact of distortion of image authentication security.

Our user studies also had limitations. As discussed in the first user study, our participants were not skilled attackers, but instead people with intimate knowledge of their targets; while we believe that the amount of knowledge about a target is the most important factor in the educated guess attacks, we need to point out that skilled attackers could launch sophisticated educated guess attacks. For instance, they could collect many photos about a target user from web services, such as Flickr, then, distort all the photos to make better guesses about the user's authentication images. Whether such attacks are practical remains an open problem, that this paper does not attempt to address.

In the second study, we used a set of photos randomly chosen from a pool of photos shared on Flickr under the Creative Commons license. We assumed that this random selection was a representative sample of pictures that people take in general. However, there would be some bias in the selected sample. For instance, private photos would not be shared under the Creative Commons license. Another possibility is that online users with knowledge of the Creative Commons license are likely to be more computer-literate than the general population. These biases are not straightforward to measure, and could have affected our analysis.

CONCLUSION

In this paper, we evaluated the security of authentication images against educated guess attacks. Our first user study showed that the original photos taken by users are vulnerable to individualized educated guess attacks if attackers have a good amount of information about the users. Moreover, even in the case when an attacker does not have any information about users, the attacker can make better guesses than random guesses based on the contextual information of the original photos. In contrast, when distorted photos are used as authentication images, attackers cannot make better guesses than random guesses even with a good amount of knowledge about target users. In our second user study, we showed that the distortion technique could mitigate the risk of the collective educated guess attacks using the biases in users' choices of authentication images.

These findings suggests that, while keeping their memorability as high as that of original photos [10], distortion technique makes *Use Your Illusion*, and potentially other recognition-based graphical authentication schemes, more secure against educated guess attacks.

ACKNOWLEDGEMENTS

This research was supported by CyLab at Carnegie Mellon under grants DAAD19-02-1-0389 and W911NF-09-1-0273 from the Army Research Office.

REFERENCES

1. A. Adams and M. Sasse. Users are not the enemy. *Communications of the ACM*, Jan 1999.
2. G. Blonder. United states patent, 1996. United States Patent 5559961.
3. G. H. Bower, M. B. Karlin, and A. Dueck. Comprehension and memory for pictures. *Memory and Cognition*, 2:216–220, 1975.
4. S. Brostoff and M. Sasse. Are passfaces more usable than passwords? a field trial investigation. In *Proc. of HCI 2000*, pages 405–424, 2000.
5. R. Dhamija and A. Perrig. Déjà vu: A user study, using images for authentication. In *Proc. of the 9th USENIX Security Symposium*, 2000.
6. W. Epstein and I. Rock. Perceptual set as an artifact of recency. *The American Journal of Psychology*, 73(2):214–228, 1960.
7. S. Gaw and E. Felten. Password management strategies for online accounts. In *Proc. of SOUPS*, 2006.
8. A. Goldstein and J. E. Chance. Visual recognition memory for complex configurations. *Perception and Psychophysics*, 9:237–241, 1970.
9. R. Haber. How we remember what we see. *Scientific American*, 222(5):104–112, May 1970.
10. E. Hayashi, R. Dhamija, N. Christin, and A. Perrig. Use your illusion: secure authentication usable anywhere. In *Proc. of SOUPS*, 2008.
11. I. Jermyn, A. Mayer, F. Monrose, M. K. Reiter, and A. D. Rubin. The design and analysis of graphical passwords. In *Proc. of the 8th USENIX Security Symposium*, 1999.
12. H. Kinjo and J. G. Snodgrass. Does the generation effect occur for pictures? *The American journal of psychology*, 6:156–163, 2000.
13. A. Kittur, E. Chi, and B. Suh. Crowdsourcing user studies with mechanical turk. In *Proc. of SIGCHI*, 2008.
14. F. Monrose, D. Davis, and M. Reiter. On user choice to graphical password schemes. In *Proc. of the 13th USENIX Security Symposium*, pages 151–164, 2004.
15. A. Rabkin. Personal knowledge questions for fallback authentication: security questions in the era of facebook. In *Proc. of SOUPS*, 2008.
16. J. B. S. Wiedenbeck, J. Waters and N. M. A. Brodskiy. Authentication using graphical passwords: Basic results. In *Human-Computer Interaction International*, 2005.
17. S. Schechter, A. J. B. Brush, and S. Egelman. It's no secret: Measuring the security and reliability of authentication via 'secret' questions. In *Proc. of the IEEE Symposium on Security and Privacy*, 2009.
18. R. Shepard. Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior*, 113(1):95–121, 1967.
19. L. Standing. Learning 10,000 pictures. *The Quarterly Journal of Experimental Psychology*, 25(2):207–222, 1973.
20. L. Standing, J. Conezio, and R. N. Haber. Perception and memory for pictures: single trial learning of 2,500 visual stimuli. *Psychonomic Science*, 19(2):73–74, 1970.
21. J. Thorpe and P. van Oorschot. Graphical dictionaries and the memorable space of graphical passwords. In *Proc. of the USENIX Security Symposium*, 2004.
22. J. Thorpe and P. van Oorschot. Towards secure design choices for implementing graphical passwords. In *Proc. of the 20th Annual Computer Security Applications Conference*, 2004.
23. J. Thorpe and P. van Oorschot. Human-seeded attacks and exploiting hot-spots in graphical passwords. In *Proc. of USENIX Security Symposium*, 2007.
24. S. Wiedenbeck, J. Waters, J. C. Birget, A. Brodskiy, and N. Memon. Authentication using graphical passwords: effects of tolerance and image choice. In *Proc. of SOUPS*, 2005.