PassTag: A Graphical-Textual Hybrid Fallback Authentication System

ABSTRACT

Designing a fallback authentication mechanism that is both memorable and strong is a challenging problem because of trade-off between usability and security. Security questions are popularly used as a fallback authentication method for password recovery. However, they are prone to guessing attacks by users' acquaintances and may be hard to recall. To overcome these limitations, we present PassTag, a hybrid password scheme that takes advantage of both graphical and textual password authentication methods. PassTag combines a user-provided image and a short personalized text description of the image, imagetag, as an authentication secret. Furthermore, PassTag incorporates decoy images to make it difficult to guess the user-provided pictures. We conducted three user studies with 161 participants for up to three months to evaluate the performance of PassTag against security questions. The evaluation results demonstrate that PassTag is significantly stronger against close adversaries and also highly memorable (92.6%–95.0%) after one, two, and three months, respectively. Our longitudinal study results show PassTag is a promising alternative for fallback authentication.

1 INTRODUCTION

The increasing number of passwords that users have to remember and their complicated password composition policies make it challenging for people to remember their passwords [35]. A recent study [15] demonstrated that 45% of participants experienced at least one account lockout in a year. When users are unable to recall their passwords, fallback authentication schemes are required for users to regain control of their accounts. Communication-based password resets or security questions are the most common approaches for fallback authentication. Communication-based password resets such as by email or mobile phone work well, but may not be appropriate in certain situations [17] (e.g., when users lost the password for the email service itself). Hence, security questions have been popularly used as an alternative that takes advantage of users' personal information, but they are not easy to achieve both in terms of security and usability [32, 34].

To overcome the limitations of existing fallback authentication mechanisms, we propose a novel fallback authentication method called PassTag (see Fig. 1) based on *picture superiority effect*[1] [28] and *levels-of-processing effect*[2] [5]. Our goal is to strengthen fallback authentication systems by taking advantages of both graphical and textual password authentication methods. We use user-chosen images (i.e., user-provided image secrets) as cues to improve memorability of textual passwords by creating an additional retrieval path to recall the textual passwords. Users will be able to remember the image secrets more easily without much practice or reinforcement than text passwords due to the *picture superiority effect*, while *levels-of-processing effect* would help users recall their user-provided image-tag secrets (personalized, unique, memorable, and difficult-to-guess tagging for their own images) using the images as cues. Moreover, user-provided image secrets can also be used as an additional security barrier to strengthen the security of PassTag, as it forces attackers to correctly choose the user-provided image secrets followed by the user-provided image-tag secrets.

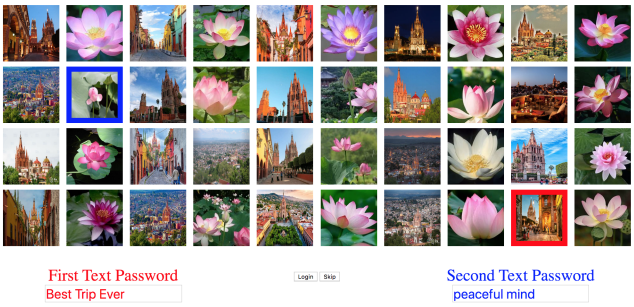

Figure 1: Screenshot of PassTag which uses user-provided images and user-provided image-tags as authentication secrets.

While security questions generally suffer from close adversary attacks (e.g., friends, colleagues, family members or acquaintances) [16, 33], PassTag was designed to make it difficult from such attacks by employing different security mechanisms in PassTag design. We specifically introduce the idea of generating a set of *decoy* images that share a common theme with user-provided image secrets so that even close adversaries cannot easily guess the victim's user-provided image secrets. In addition, we provide defense mechanisms against sophisticated human attackers as well as automated machine learning-based attacks, and demonstrate the effectiveness of our approach.

Our contributions are summarized as follows:

1) We explore how to construct images as gadgets or primitives not only to be used as passwords but also as cues to help recall textual passwords and discuss its security and memorability.

2) We design our graphical-textual hybrid fallback authentication system, PassTag, to enable such concepts, whilst simultaneously generating decoy images using adversarial samples similar to the user-provided images to guarantee that the images can be used as memory cues for legitimate users but cannot be leveraged by attackers to guess textual passwords.

3) We provide a proof-of-concept implementation of PassTag to evaluate the security and usability of PassTag. We conduct three user studies with 161 participants and find that the secrets in

[1] Graphical recognition is easier than textual recognition.
[2] More deeply encoded information becomes accessible to more cues at the time of recall.

PassTag are significantly more memorable and resilient to close adversaries than security questions.

Our IRB-approved user studies show that PassTag achieves (92.6%–95.0%) recall rates which are similar or higher than other fallback authentication systems [17, 18, 21, 34], with an average of (52.3–53.9 seconds) authentication time. In addition, we analyzed the correlation between each user's provided images and their corresponding textual image-tags using various machine learning APIs and found that users did not make guessable or obvious choices for their respected images and texts, which show that PassTag can be robust against different types of attacks. Based on these findings, we believe that PassTag can be used as a viable alternative for fallback authentication, striking a good balance between security and privacy.

The rest of this paper is organized as follows: Section 2 reviews related work. Section 3 describes the overall design of the user-chosen graphical-text hybrid fallback authentication system. Section 4 highlights the possible attack models for our system. Section 5 to Section 11 discuss the user-studies and provides the results. Section 12 offers discussion and Section 13 conclusion.

2 RELATED WORK

In this section, we present prior research that is directly relevant to fallback authentication mechanisms as well as textual and graphical passwords.

Fallback Authentication: Fallback authentication usually consists of two stages. In the first stage, users have to provide various information, such as email addresses, phone numbers or choose security questions and answers to the corresponding security questions. This information is used in the second stage to retrieve or reset forgotten passwords. In particular, communication-based multi-factor password resets relying on email and mobile phone are often used for fallback authentication. This approach works well, but may not be appropriate in certain situations [17] (e.g., when users lose the password for the email service itself or do not have access to their mobile phone). Garfinkel [10] identified that the email accounts used for fallback authentication can become a single point of failure or may be out of date and not be accessible anymore. Additionally, mobile phone numbers could sometimes be sensitive information that not every user would be comfortable sharing with their service providers [14]. Security questions are also popularly used as an alternative for communication-based password reset [9] due to the fact that security questions are basically a knowledge-based user authentication without requiring any communication with other component (e.g., server). Users have to answer a number of questions, which have to be recalled during fallback authentication. Most security questions are predefined by the service and are based on users' personal information (e.g. "What is your mother's maiden name?"). However, several previous studies have showed that security questions are neither proficient in usability nor security [13, 16, 20, 21].

Schechter et al. [34] evaluated security questions in popular webmail providers such as AOL, Google, Microsoft and Yahoo. According to their user study results, participants' acquaintances were able to guess 17% of the participants' answers within 5 guesses, demonstrating that personal knowledge questions (e.g., names of relatives, names of schools attended) are vulnerable to close adversaries. Moreover, 20% of the participants in the user study did not remember their answers within six months. To overcome the limitations of traditional security questions, Hang et al. [17] proposed dynamic security questions which were generated using mobile phone usage behavior (e.g., calls, text messages or app usage) of its users. However, it would be challenging to generate dynamic security questions that are sufficiently secure and usable. Security questions-based authentication schemes are inherently vulnerable to close adversaries. To overcome this security weakness, PassTag is designed to be secure against close adversaries by using decoy images, which are similar to the user-provided image secrets, in order to introduce confusions to attackers carrying out educated guessing attacks. In this work, we compare our approach with security questions as a baseline because it is a fallback authentication system that is widely utilized and well studied.

Cognitive Effects on Authentication: Humans have an exceptional ability to recognize images previously seen, even when the images are viewed very briefly due to the *picture superiority effect* [1]. *Picture superiority effect* refers to the phenomenon that for the human brain, recognition is an easier memory task than recall, and due to this it is easier for humans to recall graphical information as compared to textual information [6]. The most widely recognized explanation for the *picture superiority effect* is the *dual-coding theory* [29]. According to the *dual-coding theory*, graphical images are encoded in the human brain not only visually, but the images are also converted into a verbal form and remembered semantically. Biddle et al. [3] leveraged this phenomenon of human memory more effectively retrieving images than textual description to develop various recognition-based graphical passwords schemes. We explore and incorporate these effects into PassTag 's system design.

In addition, Craik et al. studied how information associated with a stimuli is longer-lasting due to the *levels-of-processing effect* [5]. *Levels-of-processing effect* refers to the phenomenon that as number of connections between information increases, it produces more elaborate, longer-lasting, and stronger memory traces. Increasing the levels of memory establishes a ceiling of potential memory performance, and retrieval cues determine the extent to which that potential is utilized [27]. *Levels-of-processing effect* can be evident by observing how textual information is easier to recall for the human brain when associated with pictorial information than recalling text without any illustration [31]. Vu et al. [42] showed that increased depth of processing increases password recall. The improvement in recall is due to the amount of connections between memorable information, which has the possibility to create more retrieval paths for recalling the information.

Graphical-Textual Hybrid Passwords: Stubblefield and Simon [36] introduced a scheme called Inkblot using images as a cue for text password entry, where users are presented with a series of computer-generated "inkblots" and asked to type the first and last letter of the word or phrase that best describes each inkblot. These letter pairs are used as the password during log-in with presented inkblots as cues to remember the characters. GridWord [2] is another hybrid scheme where users select a set of three words. The system stores an one-to-one mapping of the words to cells on a 2D grid. During authentication, users enter their password by either

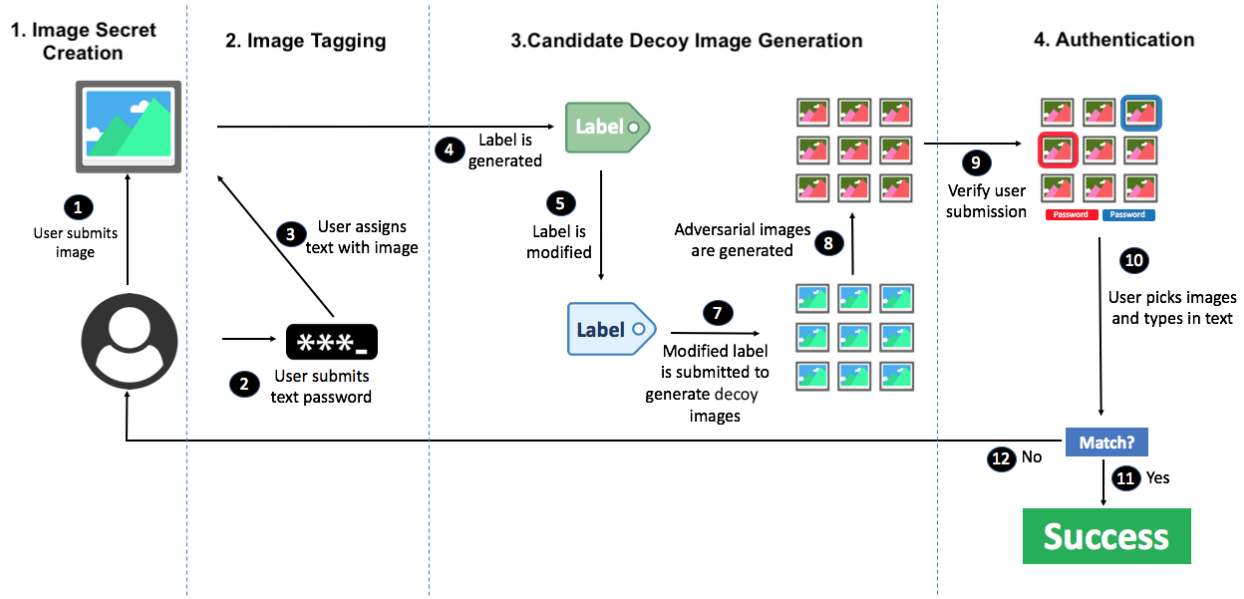

Figure 2: Overall process of PassTag.

selecting the three grid cells or the three words. GeoPassNotes [25] is an extension of the GeoPass [39], where users create a password by first selecting a location on the map as and then creating an annotation. For a log-in to be successful, both the same location and annotation must be re-entered. Macrae et al. [25] found GeoPassNotes to be highly memorable and the addition of annotations increased security with minimal usability impact.

In Deja' Vu [7], users select and memorize a subset of "random art" images from a large sample for their defined portfolio. For authentication, users must recognize images belonging to their pre-defined portfolio from a set of decoy images. In the test system, a screen of 25 images was displayed, in which 5 of the images was selected by the users for their portfolio. Users must correctly identify all images from their portfolio distinguishing that from the decoy images. Random art images are used to make it more difficult for users to write down their password or share it with others by describing their images. Marasim [22] is a jigsaw-based password scheme where users must recognize system-generated images that represent tags, which the users labeled from their user-submitted image during password creation. During password creation, users create textual tags for a image of their choice, and four random images are returned by Google for each textual tag. Users then choose one image per tag as their password. At the time of log-in a challenge set of 25 images are randomly placed and users must identify their images by entering the corresponding number on the images. Oorschot and Wan [41] proposed TwoStep to combine a textual password with graphical passwords – users must not only correctly enter a textual password but also choose pre-registered secret images from a portfolio of system-provided image.

At first glance, PassTag may seem nearly identical to TwoSteps because both systems ask users to enter textual passwords as well graphical passwords. However, the two schemes are fundamentally different in design and user behavior as PassTag starts from user-provided images as cues for textual passwords to improve both security and memorability, while TwoStep uses textual passwords and (system-provided) images independently to improve security.

3 PASSTAG DESIGN

In order to exploit the *picture superiority effect* [1] and *levels-of-processing effect* [5] in PassTag design, users must first provide their images and then input their text passwords sequentially. We ask users to supply their own images and memorable short texts along with those images because we surmise that both personalized and user-provided images and texts can increase memorability to a greater degree – users might easily recognize their own images (e.g., their own lotus image) from other (random) pictures and use the images as cues to further supply text passwords along with those images. By choosing the correct images and supplying texts, users can successfully authenticate.

PassTag consists of the following four steps: 1) Image Secret Creation, 2) Image Tagging, 3) Candidate Decoy Image Generation, and 4) Authentication. Fig. 2 pictorially describes the overall process of PassTagwith those steps.

Step 1. Image Secret Creation: In the first step, a user selects and uploads a pair of independent images that would be easily memorable and recognizable to the user, but would be difficult for others. When the user chooses and uploads his/her own images from their desktop or mobile phone as shown in the step 1 of Fig. 2, there must be some constraints on images. We specifically informed the user that these images were used for passwords. That is, user-provided images should not be easily guessable by others (e.g., self-portrait). Therefore, we advised the user not to use images that can easily be obtained from the public Internet, which lowers the security. In fact, PassTag utilizes the web to automatically query

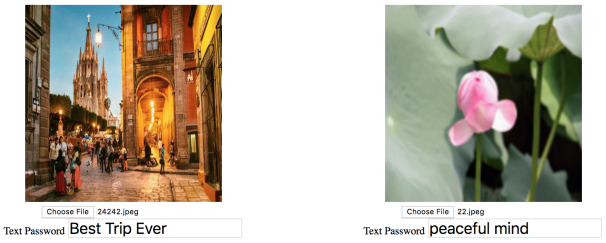

Figure 3: User-provided image secrets and image-tag secrets.

and check whether the uploaded images can be obtained from the public Internet in order to avoid the selection of weak images.

Step 2. Image Tagging: Next, the user enters his/her textual password consisting of at least two words for each image (i.e., user-provided image-tag secrets), and verifies a second time by entering the same textual password for the images as shown in the step 2 of Fig. 2. Fig. 3 shows an example of user-provided image secrets. Then, the user was additionally advised to create the corresponding textual password associated with the image (e.g., "peaceful mind" in Fig. 3). However, we reminded the user again that the provided texts would also be used for his/her password to encourage the user to choose the texts that can easily be guessable from the image (e.g., lotus and flower).

Step 3. Candidate Decoy Image Generation: It is important that not only the user can correctly choose his/her images but also it must be difficult for others to guess the images. Perhaps, close adversaries can guess the user-provided image secret with a high chance because they are likely to exploit the preferences of a victim user and his/her recent and/or important daily life events that can give a hint of the possible password. To induce confusions to such adversaries, we introduce a strategy to present a set of decoy images which are similar to the original user-provided image secrets $I_{similar}$ (e.g., pink lotus) with user-provided image secret I_{user} (e.g., lotus).

In our system, we use Imgur [19] to create a URL for each image. Then, image URLs are submitted to Google Cloud Vision API [12], which can detect and extract information about entities within an image with the API's label detection feature as shown in the step 4 in Fig 2. As a result, we can obtain the relevant text description or labels of user-provided images (e.g., sacred lotus).

However, there is a significant security issue in this step. If I_{user} is fixed, then most likely the same sets of decoy images I_{decoy} will be returned for the same input image I_{user}, because the queries are fixed. Therefore, an attacker can easily determine I_{user} and I_{decoy} by using the same queries with presented inputs and analyze the returned decoy images. In order to thwart this attack, we inject random words to perturb the original labels, and generate dynamically changed decoy images of the user-provided images to mitigate the risk of such automated attacks.

Pre-Processing for Modified Label Construction: To obtain dynamically changed I_{decoy} from I_{user}, we pre-process each label by adding a random word (e.g., adjective or adverb) as a perturbation from WordNet [26] to construct different label (e.g., gigantic sacred lotus) to as shown in the step 5 in Fig. 2. The main reason for adding

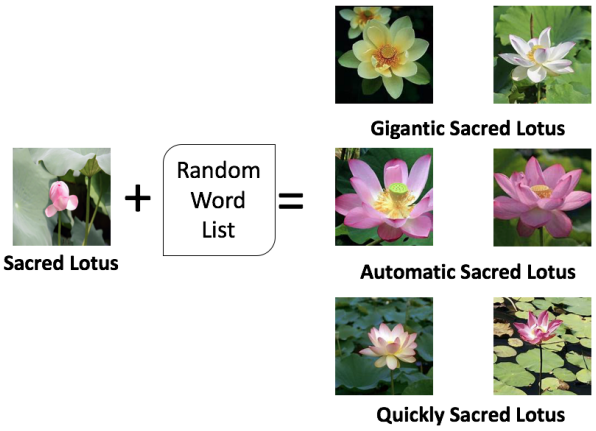

Figure 4: Pre-processing for variable label construction.

adjectives and adverbs is that it does not change the meaning greatly but retains an underlying original image class or category as shown in Fig. 4.

After pre-processing and generating a modified label, this generated modified label (e.g., gigantic sacred lotus) is submitted to a search engine as shown in the step 6 in Fig 2. Then, we use an API to search images of the given label from the Internet. For example, the Google Image Search API crawls the images similar to the ones the users uploaded. For each user-provided image, Nth number (e.g., 20) of similar images are obtained to be displayed alongside the user-provided images (e.g. 20 different lotus images). The number of similar images are less than certain threshold Nth per image, then we ask the user to upload a new image again so that we can have sufficient candidate images to prevent adversaries from guessing the user-provided images. We chose to obtain 20 similar images per user-provided image to display 40 images at a time during authentication step to be over the offline attack limit of 2^{14} by Florêncio et al. [8].

Adversarial Image Generation: Even with a modified label, decoy images are obtained from the Internet which can lead to reverse image search attacks where the attackers can filter out decoy images with user-provided images. To thwart this attack, we generate adversarial images I^{Adv}_{user} and $I^{Adv}_{similar}$ for I_{user} and I_{decoy}, respectively as shown in the step 7 of Fig. 2. The goal of an adversarial image is to make an image classifier to mis-classify the original input, where we formally define an adversarial example generation as follow: given a valid input image $\mathbb{I}$, and a target $t \neq C^*(\mathbb{I})$, it is possible to find a similar input $\mathbb{I}'$ such that $C^*(\mathbb{I}')=t$, yet $\mathbb{I}$ and $\mathbb{I}'$ are close according to some distance metric, which is an adversarial example [38]. In *Untargeted adversarial examples*, attackers only search for an input $\mathbb{I}'$ so that $C(\mathbb{I}) \neq C^*(\mathbb{I}')$ and $\mathbb{I}$ and $\mathbb{I}'$ are close. Then, finding adversarial examples can be formulated as follows [38, 43]:

$$\begin{aligned} &\min_{\mathbb{I}'} ||\mathbb{I}' - \mathbb{I}|| \\ s.t. \quad &C(\mathbb{I}) \neq C^*(\mathbb{I}'). \end{aligned} \tag{1}$$

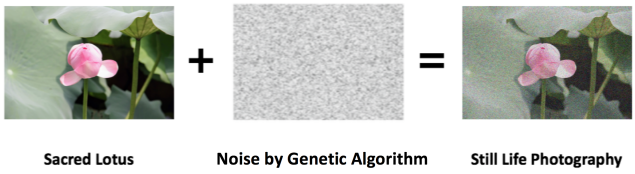

Figure 5: Adversarial image generation, where I_{user} is the user provided original image and I_{user}^{Adv} is the generated adversarial image.

The example of adversarial image generation process is depicted in Fig. 5, where I_{user} (i.e., sacred lotus) is the original input image and a noise is added according to adversarial image generation algorithm [23]. Then, the final I_{user}^{Adv} is generated. Humans can still recognize the added noise generated I_{user}^{Adv} and do not find much difference from the original image I_{user}. However, these images will be new such that they cannot be searched from the Internet. Moreover, these images cannot be recognized by conventional image recognition tools. Among the several adversarial image generation algorithms, we chose an approach by Kim and Woo [23] which does not require any knowledge of underlying machine learning models (whitebox) in commercial APIs and can generate adversarial images effectively.

Step 4. Authentication: The last step is *Authentication*, as shown in the step 8 of Fig. 2 in which the user verifies his/her submission by correctly selecting his/her own images from the decoy images and typing in the correct text password for each selection. An example authentication page is shown in Fig. 1. If the user is able to correctly select both images and type in text passwords, the information is saved. Otherwise, the user must go through three trials where each trial displays 40 images containing 0, 1, or 2 original images alongside the other decoy images in random order. If the user selects an image, he/she will be asked to type the corresponding text password associated with the image.

4 SECURITY CONSIDERATIONS FOR PASSTAG

Given a total of 40 images presented in three sessions and two text inputs a user needs to provide, we considered various theoretical as well as practical attack models. These were the key driving security requirement parameters to the design of our system.

Offline Brute-Force Random Guessing Attack: An attacker can randomly try to guess images and texts. The number of trials needed to guess correctly both two images and two texts are defined as T^{Images} and T^{Texts}, respectively. If an attacker randomly attempts to choose 2 user-provided images successfully out of 40 displayed in each image displayed over the 3 session, at the worst case, it would require the correct guesses up to $\binom{40}{2} \times \binom{40}{2} \times \binom{40}{2}$ ($> 2^{28}$). In addition to two images, two input texts have to be entered, correctly. Therefore, it is easily over the offline attack limit of 2^{14} guesses by Florêncio et al. [8]. However, this is the worst case scenario, and next we present more realistic attacks, which an attacker can exploit the correlations between images and texts.

Automated Images and Texts Correlation Attack: We assume attackers can use popular machine learning APIs to automatically find highly relevant texts from input images. For example, attackers can download the presented images and produced the relevant words for the presented images to guess text passwords, which is more efficient than the above brute-force attack. In the result section, we will evaluate the results of how user-provided texts and images are correlated. This result can show that the effectiveness of attacks using machine learning APIs that leverage the correlations between images and texts.

Automated Image Search Attack: Similarly, attackers can look for similarity and differences among the presented image set. For example, analyzing the distribution of presented images and category in which images are belong to (e.g., foods vs. clothes), an attacker can possibly narrow down the guess by first throwing out the different category of images. In order to prevent this attack, we pre-process the labels of images by adding a random word as seen in Fig. 4 to prevent identical images to appear when attackers choose to search for images using a keyword and uniformly choose candidate images from within their respective categories.

Candidate Image (Decoy) Exclusion Attack: Even though we generate candidate decoy images with other images not available from the Internet, it is possible for attackers to match and guess candidate images from the Internet. Therefore, an attacker can remove and exclude those images as possible users' answers. An attacker can leverage again image search APIs to find returned similar images for each candidate image. In order to defend against this attack, we pre-process all our images by adding noise and create adversarial images to thwart the machine learning attack. In this way, image search APIs cannot produce any meaningful output or return relevant image result. We carefully designed and pre-processed each image to deceive Microsoft and Google machine learning APIs for all images we present to users. Hence, this attacker will not be effective.

Shoulder-Surfing Attack: Attackers may gain information about victims' passwords by direct observation or external recording device. Modern cameras and cellphones with high resolution lenses make shoulder-surfing [24] a real concern if attackers target specific users using passwords in public environment. To mitigate shoulder-surfing attack, PassTag introduce decoy images that can be helpful to induce confusions to attackers. Also, the user-entered text in the password field can be shown as an asterisk to reduce the risk of shoulder-surfing attack.

Close Adversary Attack: Close adversaries have the advantage to know the users well and thus, make educated guesses to find the correct answers. The threat can be significantly be increased, when these close adversaries use additional tools such as social networks or search engines for searching images. This kind of attack can be considered as one of the worst case scenarios for security questions, and authentication schemes, where users create authentication secrets in which they provide their own images. Threats by close adversaries were shown to be very likely and thus, interesting to consider [32]. To mitigate threats by close adversaries, PassTag generates decoy images that are similar to the user-provided images. We hypothesize that if all images closely similar to each other, it will be significantly more difficult for close adversaries to discern the user-provided images compared to a scheme which generates random

images alongside the image secrets. In this work, we specifically recruit close friends or family members to evaluate PassTag against close adversaries.

5 STUDY PROCEDURE

We conducted three different user studies to evaluate the performance of PassTagand to evaluate the different aspects of our approach. Table 1 shows the demographics the participants in all three user studies. The objective of study 1 was to evaluate the effectiveness PassTag's design of generating similar decoy images by evaluating its security against close adversaries with an near-identical graphical-textual hybrid scheme that generated random decoy images (RandomTag). The goal of study 2 was to evaluate the usability of PassTag and evaluate the types of authentication secrets and errors users made using PassTag. Lastly, we conducted study 3 to evaluate the memorability and security of PassTag against a comparable baseline, security questions for up to three months.

For statistical testing we performed the pairwise Fisher's Exact Test (FET), which yields more accurate confidence for relatively smaller sample size and the *t*-test for creation and authentication time.

Table 1: Demographics of participants (N = 161).

Gender		
Male	101	(62.7%)
Female	60	(37.3%)
Age group		
Under 20	33	(20.5%)
20–29	105	(65.2%)
30–39	11	(6.8%)
40–49	8	(5.0%)
50+	4	(2.5%)

6 STUDY 1: CLOSE ADVERSARY GUESSING ON PASSTAG VS. RANDOMTAG

We first conducted a between-subject user study to test our hypothesis that it would be much more difficult for close adversaries to correctly guess secrets when similar images are used as decoy images than when random images are used as decoy images. We designed another fallback authentication system RandomTag with mostly identical functionalities as PassTag. However, RandomTag displayed random decoy images instead of similar decoy images during the authentication process. Participants were required to enroll into the user study with one other close friend or acquaintance. Each pair of participants were instructed to create an authentication secret on either PassTag or RandomTag. The two schemes were randomly assigned for each pair of participants.

Each participant created authentication by following the same methodology we employed in the first session of study 1. Once participants completed creating their authentication secrets, we asked each pair of participants to play as close adversaries of each other and were given five attempts to correctly choose the correct image secrets along with the corresponding image-tags. Participants were allowed to use the Internet and their smartphones for research.

7 RESULTS OF STUDY 1

Demographics: We recruited 30 volunteers (18 men, 12 women) from our campus. The age group of 20 to 29 was most represented with 27 participants. 11 pair of participants were close friend, 2 were their partner, and 2 were coworkers. Each participant received $10 gift vouchers as incentives.

Close Adversary Guessing on PassTag vs. RandomTag: We tested the hypothesis that PassTag with its similar decoy images should be more resilient attacks from close adversaries than RandomTag. We found strong evidence in support of this hypothesis. None (0/15) of the close adversaries were able to correctly authenticate PassTag compared to 6.7% (1/15) who were able to correctly authenticate in RandomTag.

It is important to note that although the results between PassTag and RandomTag appear similar, close adversaries were able to correctly choose the image secret of RandomTag 28.0% (42/150) much more often than PassTag 7.3% (11/150), showing significant statisitcal difference (FET with one-tailed, $p \ll 0.00001$). The additional authentication step of providing text tags for image secrets prevented all but one close adversary to correctly authenticate in RandomTag, and none in PassTag. Therefore, our clear recommendation is using decoy images generated by PassTag rather than random images with respect to security.

8 STUDY 2: EVALUATING USABILITY OF PASSTAG

First Session: First, we asked the participants to select and upload two sets of images that would be easily memorable to distinguishable for the participant, but be difficult for others to guess. Next, users were asked to create corresponding text password for each image, which is a personalized, memorable unique, and not easily guessable textual password (tagging) for each image.

For each user uploaded image, PassTag generated 20 similar decoy images and users were then asked to verify their submission by correctly selecting their own images among 40 other images and type in the correct textual password for each selection. If the user was unable to correctly select both images and type in the text passwords, which they had provided, then they were given another attempt. If they failed the second attempt, they were asked to resubmit more memorable images and textual passwords. If the user was able to correctly select both images and type in the correct text password, then the registration session was successfully completed.

Second Session: Participants returned one week after registration and tried to log-in to the website with the account they had created. To measure memorability, we had users go through the log-in procedure three consecutive times. Each log-in procedure was divided into a three-part trial. In each trial users were presented with 40 images that may contain 0, 1, or 2 of their original images alongside displayed decoy images. Users went through each trial to correctly select their own images from the decoy images and type in the correct textual passwords for each selection. Every input by the users and the time it took to finish the authentication was recorded. Users were not be notified if they correctly completed the authentication process. Once the authentication was over, participants were asked to provide feedback on their experience using

PassTag, such as what they liked or disliked about it and whether it was easy or difficult to use. The exit survey questions are provided in Table 2 to measure users' sentiment, using System Usability Scale (SUS) [37].

Table 2: Survey questions based on SUS.

Survey Questions
Q1. Authentication time for PassTag was reasonable/adequate.
Q2. Successfully authenticating using PassTag was reasonable/adequate.
Q3. It was easy to use PassTag.
Q4. It was easier to remember the images than the text passwords.
Q5. It was easier to remember the text passwords than the images.
Q6. The images that were generated by PassTag were similar to my images.
Q7. Looking at the images helped me remember my password.

9 RESULTS OF STUDY 2

We measure the recall after one week, creation and authentication time, user sentiment, and strength. In addition, we analyze the images and textual passwords users chose. Lastly, we measure the distance between image-to-image and image-to-text to analyze the correlations.

Demographics: We recruited 51 volunteers (30 men, 21 women) from our campus. The age of the participants varied between 18 and over 65. The age group of 20 to 29 was most represented with 35 participants. Among 51 participants who created the password, all of them came back for authentication and received $10 gift vouchers as incentives.

One Week Recall: We present the one week recall results in Table 3. Among 51 participants, 92.6% of the participants were successful within 3 trials after one week. 88.9% of the participants were able to successfully select and recall both images and matching textual passwords for all of their trials without any kind of errors. PassTag achieved overall authentication success rate of 92.6%. Although most participants were able to successfully authenticate using PassTag, some were unable to successfully authentication due to partial recollection of images and texts, which tributes to 7.4% of participants.

Table 3: Average, median, and std. of creation and completion time in secs and average auth. success rate per trial.

Attempt	Time (sec) Avg./Med./Std.	Auth Succ. Rate
Creation	51.7/50.0/16.7	–
Auth. Trial 1	66.3/60.0/27.8	88.9%
Auth. Trial 2	56.6/51.0/24/5	92.6%
Auth. Trial 3	52.3/46.0/22.9	92.6%
Average Auth.	58.4/52.0/25.1	92.6%

Next, we carefully measured the rate of correct selection of all users for each image and text type, not taking into account whether the authentication was a for success or failure. This allow us to measure, observe, and compare whether text passwords (tagging) are difficult to recall than images, etc.

Table 4: Success rate with image and/or text.

Authentication type	Trial 1 Selection	Trial 2 Selection	Trial 3 Selection	Average. Selection
Both image and text	91.2%	93.5%	95.4%	**93.3%**
Only image	94.9%	97.2%	99.1%	**97.1%**
Only text	94.9%	94.4%	94.9%	**94.7%**
Neither image nor text	0%	0.9%	3.7%	**1.5%**

Table 4 summarizes the percentage of correct recall of each password type. In Table 4, selections where only the image (only image) was chosen correctly amounted to overall 97.1%, while only the text passwords were correct amounted to 94.7%. Therefore, when we compare images and textual passwords success rate, we can observe the *picture superiority effect* by 2.4% (97.1% vs. 94.7%). We also asked participants about this question during the exit survey on what was easier to remember.

Lastly, about 1.5% (neither image nor text) of the selections were made, where the participants selected both the image and text passwords incorrectly, which is extremely small. Also, it is interesting to observe the combined success rate result of 95.4% at trial 3 slightly increased from the text success rate (94.9%) when combined with image (99.1%). This result can indirectly indicate that *levels-of-processing* effect would help remember users about their authentication secrets, where images helps remember textual passwords. We also asked this during the exit survey and obtained moderate agreement on this effect. Therefore, we demonstrate that our approach yields high recall compared to other competing approaches (e.g., 87% with dynamic security questions [17], 88% of location-based security questions of PCCP [4], and 86% from Story scheme [6]). For users who were unsuccessful, we analyzed the errors each user made in the later section.

Time for Creation and Authentication: We present the average, median, and std. for time it took for users to: (1) complete the creation and (2) complete the authentication during the user-study. On average, it took the participants to complete the creation trial was 51.7 seconds with standard deviation of 20.0 seconds, while the median creation time was 50.0 seconds. The average time it took the participants to complete the authentication trial was 58.4 seconds with standard deviation of 25.1 seconds as shown in Table 3. Hence, the overall creation time is slightly less than 1 minute and the authentication time was very similar on average as shown in Table 3.

Fig. 6 illustrates the times for each trial the participants required to complete authentication. The X-axis represents each individual participant and the Y-axis represents the amount of seconds it took the participant to complete the authentication. The time to complete from trial 1 to trial 3 were totaled for each participant and the totals were sorted in ascending order.

The longer authentication time was expected as PassTag requires a user to carefully observe 40 images to determine if their images exists and provide matching textual password input after selection if the image exists. However, as users progressed through each trial from trial 1 to trial 3, we can generally observe that the average authentication time decreases as trial increases from 66.3 seconds (trial 1) to 52.3 seconds (trial 3) on average as shown in Table 3. We believe that as users become more familiar with usingPassTag along

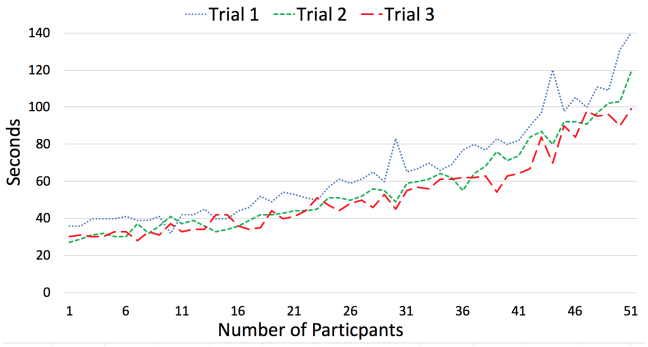

Figure 6: Distribution of time to complete authentication for each trial per user.

with the decoy images, users' average authentication time may continue to decrease.

Images and Texts Correlations: We also analyze the correlations on visual and semantic distance among the user-provided images, texts, and system generated decoy images. To measure the visual distance between images, we first calculated the color coherence vectors (CCV) of the images provided by Pass et al. [30] using Color Coherence Vectors (CCV) to measure image similarities. CCV is widely used for image retrieval and image content dissimilarity comparison, where the distance between two images is obtained by comparing the pixel color and coherence between 2 CCVs, where 0 means no correlation and 1 means the two are identical.

The four correlations are shown in Table 5 as follows: (1) img-to-img distance $dist(I_1, I_2)$, which is the visual distance between the first user-provided image I_1 and the second user-provided image I_2, (2) img-to-decoy image distance $dist(\{I_1, I_2\}, \{I_{Decoy}\})$, which is the visual distance between user-provided images I_1, I_2 and 40 system generated decoy images I_{Decoy}, (3) img-to-txt distance $dist(\{I_1, I_2\}, \{T_1, T_2\})$, which is the semantic distance between user-provided images I_1 and I_2 and text passwords T_1 and T_2, and (4) txt-to-txt distance $dist(T_1, T_2)$ which is the semantic distance between first text password and second text password.

To measure the correlation between user-submitted images to the text passwords, image-to-text, and the correlation between the two text passwords, we utilized word2vec [11], which is widely used to measure the semantic distance between words. The distance For image and text correlation, we inputted the text labels of the image generated by Google Cloud Vision API [12] and the corresponding text passwords submitted by the users.

We hypothesize that if the 1) visual distance of img-to-img distance, 3) img-to-txt dist, and 4) txt-to-img distance are large which show low correlations (close to 0), then it will be difficult for attackers to guess the correct images and texts as the images and texts do have low correlation. On the other hand, 2) img-to-decoy distance needs to be highly correlated (close to 1), which will result in the correct images and decoy images to be hardly indistinguishable to attackers. Therefore it is important to measure the correlation among the images and texts.

Table 5 shows that the average correlation between the two user-provided images that are used as passwords was 0.240, and the

Table 5: Various image-to-image and image-to-text distances.

Distance	Avg. (std.)
(1) $dist(I_1, I_2)$ (img-to-img)	0.240 (0.113)
(2) $dist(\{I_1, I_2\}, \{I_{Decoy}\})$ (img-to-decoy)	0.480 (0.021)
(3) $dist(\{I_1, I_2\}, \{I_{T_1,T_2}\})$ (img-to-txt)	0.000 (0.000)
(4) $dist(T_1, T_2)$ (txt-to-txt)	0.003 (0.003)

standard deviation was 0.113. Therefore, we can observe that users provided two images have fairly low correlation and users seem to submit dissimilar images. While correlations between images and decoy images are higher with 0.480. This shows that close to half of decoy images (about 20 images) are in the same category and make it difficult to guess.

The average distance between first user-provided image and first text password and second user-provided image and second text password was 0. Therefore, PassTag generates highly uncorrelated adversarial images, which can mitigate the machine learning attackers. Additionally, the average distance between the text passwords was 0.00 and the standard deviation was 0.003 as shown in Table 5. This clearly confirms that users do not provide the related textual passwords.

9.1 Error Analysis on User Inputs

Image Input Errors: The overall success rate of image selection was 97.1% as shown in Table. 4. To understand the reasons for failures to recall images, Fig. 7 shows the percentage of the types of image input errors made during authentication, where we show them as red bars. 72% of the errors ("Left Empty") were due to users not selecting an image when their image was displayed. 21% of the errors ("Incorrect Image") were due to users selecting an incorrect image when their image was displayed alongside it. Lastly, 7% of the errors ("Unnecessary Selection") were due to users selecting an incorrect (decoy) image when no correct image was displayed. We can conclude from the results that valid users rarely select the decoy images set by PassTag. Majority of the errors were made when a correct image was presented to the users (94%). We observed that users were able to more accurately discern decoy images than recognizing their correct image. We also noticed that users made more errors during the 1st and 2nd attempts which may be due to being initially unfamiliar with the system design. Additional reason for the high error rate of users not selecting an image when it was displayed could be due to the display scheme of PassTag which presents 40 images in 150x150 pixels. A different quantity and quality of images and sizes in which they are displayed could allow users to better recognize they images and improve the error rate. We leave this study for future work.

Textual Password Errors: We initially categorized textual errors into two categories as easily fixable errors (e.g. spacing) shown in blue bars and complex errors (e.g. unnecessary input) in purple bars as shown in Fig. 7. The "complex error" criteria is for when users' inputs are completely different from their original passwords (colored in purple bar), and the "fixable error" means it matches the entire strings but misses capitalization, spacing between words or

similar words (colored in blue bar). As shown in Fig. 7, only 35% of textual errors were complex errors. The 65% fixable errors rate leads us to believe that users were able to get an idea of what the text password was but not the exact text itself. We further categorized the types of textual errors users made in PassTag as shown in as shown in Fig. 7 we categorized the errors into 7 categories: 1) Spacing (e.g., favoritememory vs. favorite memory); 2) totally incorrect submission; 3) similar word was provided (e.g., hungry4food vs. starving4food); 4) capitalization errors (e.g., 9312happytimes vs. 9312Happytimes); 5) unnecessary input as wrong image to text combination ($I_1 + T_1$) vs. ($I_1 + T_2$); 6) spelling errors (e.g., b4nn4nap0p vs bannanapop); or 7) additional words were added (e.g., workSunday vs myworkSunday). Our analysis showed that the types of textual errors users made by the users appear to be minor and show that majority of the user seem to recall the gist of what their original password was.

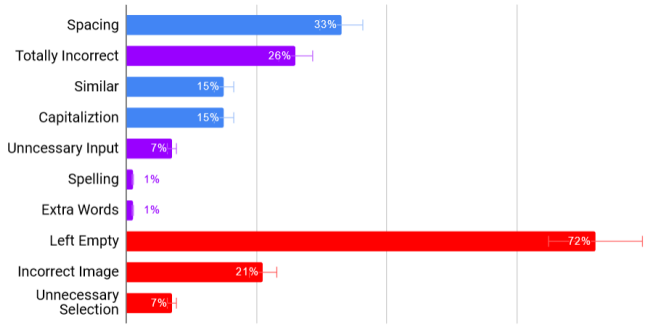

Figure 7: Histogram of textual and image input errors.

User Sentiment: At the end of the authentication study, we asked participants to provide feedback through a survey created about the usability and memorability of PassTag. We asked each participant to answer questions shown in Table 2 regarding their experience in 1 to 7 scale. The score 1 being strongly disagree, 4 being neither agree or disagree, and 7 being strongly agree. The results are summarized in Fig. 8.

For Q1. "whether the solving time was reasonable", the average score was 6.11, median was 6. And for Q2, "if successfully logging in with the authentication system was reasonable", the average score was 6.24, median was 6, and standard deviation was .95. Lastly for Q3, "if it was easy to use the authentication system", the average score was 6.09, median was 6, and standard deviation was 1.04. Therefore, most participants agreed that PassTag was quite usable and found the time it took to complete registration and authentication to be reasonable.

For memorability, when Q4 was asked "if it was easier to remember the images than the text passwords", the average score was 6.17, median was 6.15. Conversely, when asked question Q5, "if it was easier to remember the text passwords than the images", the average score was 3.19, and median was 3. Hence, we confirm the effectiveness of the picture superiority as users found images to be easier to recall than textual passwords. For question Q6, "if looking at the images helped users remember the text passwords", the average score was 5.35, and median was 6. Therefore, we can demonstrate

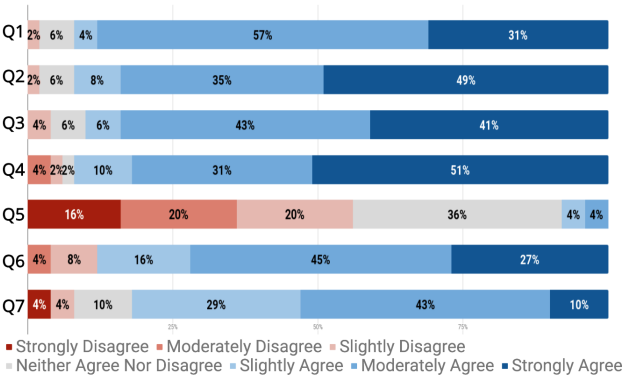

Figure 8: Participant ratings (Q1. Authentication time was reasonable, Q2. Successfully authenticating using PassTag was reasonable, Q3. It was easy to use PassTag, Q4. It was easier to remember the images than the text passwords, Q5. It was easier to remember the text passwords than the images, Q6. The images that were generated by PassTag were similar to my images, and Q7. Looking at the images helped me remember my password in Table 2).

the effectiveness of and level-of-processing effect, where the images are viable to be used as cues to remember text passwords. Lastly, when asked question Q7, "if the images that were generated by the authentication system was similar to the user-provided images", the average score was 5.76, and median was 6. This suggests that the decoy images generated by PassTag are similar to user-provided images. This allows for an even and consistent distribution of images to be displayed making it difficult for attackers to specifically distinguish a specific single image.

10 STUDY 3: LONG TERM MEMORABILITY COMPARISON WITH SECURITY QUESTIONS

Our initial motivation for designing PassTag with decoy images was to create a memorable and secure fallback authentication scheme. To evaluate PassTag against this motivation, we experimentally evaluated PassTag against another comparable baseline security questions, for three key performance measures: 1) long-term memorability with infrequent authentication, 2) resilience to close adversary attacks, and 3) authentication speed. Long-term memorability after extended disuse was our key measure of interest, because PassTag is designed for fallback authentication, where secrets should be easily memorable for long-term without a lot of repetitions. Next, we tested for resilience against close adversaries such as close friends, sibling, and spouse to measure the strength of decoy images and as security questions come with shortcomings in this aspect [32]. Lastly, we selected authentication speed because of its importance in usability and deployability.

First Session: We conducted a between-subject user study with 80 participants in our lab. While users (40 participants) had to partake in all four time-separated sessions in creating and authenticating with one of the three authentication scheme, close adversaries

(40 participants) only had to come for the first session. 40 participants were instructed to create an authentication secret on one of the authentication scheme (20 participants per scheme). Each scheme was randomly assigned for each participant and the complexity of the secrets were selected to be very similar on average – our security questions scheme used questions which are provided by popular webmail providers [34] and contained identical set of questions in relation to the images deployed by PassTag. The security questions and participant selections are provided in Table 6.

Table 6: Security Questions used by the top four webmail service providers

Security Question	Selected by User
What is your favorite food?	7.5%
What is your pet's name?	5%
Where were you born?	10%
What is your favorite restaurant?	17.5%
What is the name of your school?	0%
Who is your favorite singer?	0%
What is your favorite song?	10%
What is your favorite film?	15%
What is your favorite book?	0%
What was your first job?	5%
What was your first teacher's name?	12.5%
What is your first phone number?	0%
Where was your mother's birthplace?	7.5%
Who was your best childhood friend?	5%
Who was your favorite teacher?	0%
Who is your favorite historical person?	0%
What was the name of your first school?	0%
Who was your childhood hero?	5%
What is your favorite sports team?	0%
What is your father's middle name?	0%

Close adversaries were asked to leave the lab and wait, while users created authentication secrets for their assigned authentication scheme. Identical to creation in PassTag, users in all authentication schemes were given three attempts to correctly select and create authentication secrets. Once users completed creating their authentication secrets, we asked users to leave the lab and invited the close adversaries in. Close adversaries had five attempts to correctly choose the images and the security questions, along with the corresponding answers. Close adversaries were allowed to use the Internet and their phones for research.

Second and Third Session: One month after the first session, we invited users back to perform another memorability test. Again, users had to answer their three questions from the first session within three attempts. Another memorability test was conducted in a third session that took place one month after the second one (i.e. two months after the first session). Users were instructed to perform the same tasks as they had for the second session.

Final Session: One month after the third session, we invited users to a long-term evaluation in attempt to simulate a realistic fallback authentication scenario in which length time between enrollment and required fallback authentication had passed. Users performed the same procedure as similar to the second and third session. At the end of the sessions, we asked users to provide feedback through a survey created about the usability and memorability of each authentication scheme.

11 RESULTS OF STUDY 3

Demographics: We recruited 80 participants (53 men 27 women) from our campus. The age of the participants varied between 18 and over 65. The age group of 20 to 29 was most represented with 58 participants. The relationship between the users and their close adversaries was as follows: 16 participants brought their close friend, 7 brought their partner, 5 brought their spouse, and 2 brought their sibling. They received $10 gift vouchers for their participation.

Memorability: We first tested the hypothesis that graphical secrets in PassTag should be more memorable than security questions. We found strong evidence in support of this hypothesis. Table 7 shows the percentage of participants successfully authenticated in each scheme within three attempts. Overall, 95.0% of our participants (19/20) could recall the secrets in PassTag within three attempts even after the third month, compared to just 55.0% (11/20) who could recall the security questions, showing statistical significance (FET test with one tailed, p=0.04 < 0.05).

Table 7: Participants that successfully remembered their authentication secrets.

Scheme	One Month	Two Month	Three Month
PassTag	95.0%	95.0%	95.0%
Security Questions	65.0%	60.0%	55.0%

Resilience to Close Adversaries: We next tested the hypothesis that PassTag should be more resilient to shoulder surfing attack than security questions authentication. We also found strong evidence in support of this hypothesis. None (0/20) of the close adversaries were able to correctly authenticate PassTag compared to 30.0% (6/20) who were able to correctly answer the security questions. This result also shows statistical significance (FET test with one-tailed, p=0.04 < 0.05), demonstrating it is significantly harder for close adversaries to guess PassTags than security questions.

Authentication Speed: Lastly, we measured the authentication speeds of PassTag and security questions. Fig.9 shows the distributions of how long it took participants to authenticate in PassTag, and security questions. while the average authentication time of PassTag was 53.9 secs on average users were able to authenticate quicker for security questions with 38.2 secs on average (t-test, $p \ll 0.001$).

In sum, security questions are faster than the graphical hybrid schemes, and although the increased authentication time of PassTag may not be desirable for frequent use cases (e.g., day-to-day access), it can be justifiable (under 1 min.) for infrequent use cases such as for fallback authentication given the memorability and security improvements of PassTag. Unlike security questions, PassTag is more memorable without much practice or reinforcement.

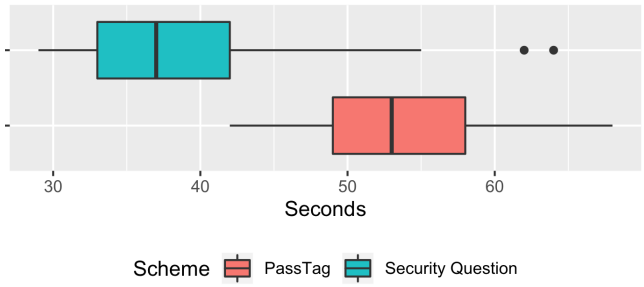

Figure 9: Box plot of the authentication time distributions for PassTag and security questions.

12 DISCUSSION

12.1 Advantages of PassTag

Our security and memorability analysis results demonstrate that participants who used PassTag can create user secrets that are not only secure against close adversaries but also memorable to them over a long time period (e.g., three months) compared with those who used security questions. We believe that combining graphical and textual passwords would be helpful to enhance both security and memorability. The advantages of PassTag can be summarized as follows.

PassTag is Memorable: The overall successful recall rate of images (97.1%) and input of images-tags (94.7%) after one week in the second user study indicate that PassTag is memorable. The high memorability is significant as the decoy images generated by PassTag were similar to the user-provided images. This shows that most users have an ability to successfully distinguish their chosen specific images from a cluster of similar images due to the *picture superiority effect* [1]. The memorability of image-tags was also high indicating that it could be easier for users to recall text after presented with images as cues to the corresponding image-tags due to the *levels-of-processing effect* [5]. Additionally, we found that users who used PassTag maintained high long-term memorability (95.0% after three months) of both their graphical and textual passwords with minimal reinforcement and infrequent authentication compared to the widely used security questions. We hypothesize that both graphical and textual passwords act as cues of one another to help during recall.

Users can Create Secure PassTag Secrets: For PassTag to be secure, the correlation between: (1) two user-provided images; (2) user-provided image label and text password; and (3) two text passwords; needs to be small. Surprisingly, our results demonstrate that users chose non-similar images and inputted non-obvious text passwords with the images. The diversity and combination of image and text passwords can be helpful to enhance the security of the fallback authentication system. Additionally, although the system was tasked to find 20 similar images for each user-provided image, the correlation between user-provided image and system generated image needs to be high and PassTag was able to generate similar images to the user-provided images.

Our approach showed promising results in terms of security as the guessing attacks from close adversaries performed poorly according to our study results. This was mainly because the decoy images used in PassTag would be effective in increasing the difficulty of guessing the correct image secret. We found that strangers failed to guess the image secrets at most times through user studies. Even close adversaries who have correctly guessed the image secret finally failed to provide the additional corresponding correct textual password. In comparison to the analysis by Rabkin [32] where 12% of the security question samples could be attacked through user studies, our close adversaries had 0% success rate attacking PassTag. Those results show that the superiority of PassTag over security questions in terms of security.

PassTag is Usable: Examining the results of the user studies, PassTag is both secure and memorable. However, this advantage comes at the cost of usability as our user-study results show that it took longer for users to create (average 51.7 seconds) and authenticate passwords (average 52.3 secs). The requirement for users to examine and discern passwords from 40 images could be the contributing factor to the extended creation and authentication time. Yet, given the high memorability and infrequent use for fallback authentication systems, we believe we can justify the longer authentication time. One possibility to reduce creation and authentication time is to reduce the total number of images that are displayed by PassTag. This could increase usability but at the cost of security. For future work, we plan to find the optimal amount of decoy images to generate, while also preserving high security PassTag provides.

12.2 Limitations and Future Work

The key limitations of our work are the following: (1) we did not have users create multiple accounts with PassTag; and, (2) our sample size was small and the majority of the participants were quite mostly young memory-strong population. Additionally, ecological validity is a limitation in our study designs: we evaluated PassTag in a controlled lab setting, and participants were not using these secrets to secure a device or account they personally used. These are all limitations that will need to be addressed before PassTag can be adopted into real world use-cases.

A concern about the current PassTag design is that users should provide secure and memorable images and texts. However, similar to conventional textual passwords, it is sometimes challenging for users to choose secure and usable secrets as password. For example, Ur et al. [40] demonstrated that users often have wrong perception of what constitutes a secure password. Therefore, one future direction to strengthen the security of PassTag is to explore the possibility of system-assigned image-tags instead of user-provided image-tags. The state-of-the-art machine learning technology might be helpful to automatically generate secure and memorable image-tags. Evaluating the usability and security of the PassTag in real-world settings is an area we plan to explore in the future as well. Lastly, though we were able to test recall rate up three months, we plan to continue to evaluate memorability of PassTag after three months.

13 CONCLUSION

We proposed PassTag, a fallback authentication system, which takes advantage of graphical and textual authentication methods. PassTag strikes a good balance between memorability and security.

Our user study results demonstrate that PassTag improved the success rate of authentication compared with conventional fallback authentication schemes such as security questions. While it may take more time for users to register and log-in using PassTag, we believe the benefits of PassTag outweigh the cost. In future work we plan to conduct large scale user studies to evaluate PassTag for longer periods of disuse. Furthermore, we plan to implement different number of images presented and selected by users in the authentication phase to find the optimal number of iterations that will reduce authentication time, while keeping it secure. In sum, PassTag can be a promising option for fallback authentication systems. Therefore, we need to focus on further understanding the images/tags users provide and the policies, guidance, and feedback that can help them choose more secure and memorable images and tags for PassTag.

REFERENCES

[1] John R Anderson and Gordon H Bower. 1972. Recognition and retrieval processes in free recall. *Psychological review* 79, 2 (1972), 97.

[2] Kemal Bicakci and Paul C Van Oorschot. 2011. A multi-word password proposal (gridWord) and exploring questions about science in security research and usable security evaluation. In *Proceedings of the 2011 New security paradigms workshop*. ACM, 25–36.

[3] Robert Biddle, Sonia Chiasson, and Paul C Van Oorschot. 2012. Graphical passwords: Learning from the first twelve years. *ACM Computing Surveys (CSUR)* 44, 4 (2012), 19.

[4] Sonia Chiasson, Elizabeth Stobert, Alain Forget, Robert Biddle, and Paul C Van Oorschot. 2012. Persuasive cued click-points: Design, implementation, and evaluation of a knowledge-based authentication mechanism. *IEEE Transactions on Dependable and Secure Computing* 9, 2 (2012), 222–235.

[5] Fergus IM Craik and Robert S Lockhart. 1972. Levels of processing: A framework for memory research. *Journal of verbal learning and verbal behavior* 11, 6 (1972), 671–684.

[6] Darren Davis, Fabian Monrose, and Michael K Reiter. 2004. On User Choice in Graphical Password Schemes.. In *USENIX Security Symposium*, Vol. 13. 11–11.

[7] Rachna Dhamija, Adrian Perrig, et al. 2000. Deja Vu-A User Study: Using Images for Authentication.. In *USENIX Security Symposium*. 4–4.

[8] Dinei Florêncio, Cormac Herley, and Paul C Van Oorschot. 2016. Pushing on string: The'don't care'region of password strength. *Commun. ACM* 59, 11 (2016), 66–74.

[9] Steven Furnell. 2007. An assessment of website password practices. *Computers & Security* 26, 7-8 (2007), 445–451.

[10] Simson L Garfinkel. 2003. Email-based identification and authentication: An alternative to PKI? *IEEE security & privacy* (2003), 20–26.

[11] Yoav Goldberg and Omer Levy. 2014. word2vec Explained: deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722* (2014).

[12] Google Cloud Platform. 2018. Cloud Vision Api. https://cloud.google.com/vision/, Last accessed on 2018-11-30.

[13] Virgil Griffith and Markus Jakobsson. 2005. Messin' with Texas Deriving Mother's Maiden Names Using Public Records. In *International Conference on Applied Cryptography and Network Security*. Springer, 91–103.

[14] Ralph Gross and Alessandro Acquisti. 2005. Information revelation and privacy in online social networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM, 71–80.

[15] Hana Habib, Pardis Emami Naeini, Summer Devlin, Maggie Oates, Chelse Swoopes, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2018. User Behaviors and Attitudes Under Password Expiration Policies. In *Fourteenth Symposium on Usable Privacy and Security (SOUPS 2018)*. USENIX Association, Baltimore, MD, 13–30. https://www.usenix.org/conference/soups2018/presentation/habib-password

[16] William J Haga and Moshe Zviran. 1991. Question-and-answer passwords: an empirical evaluation. *Information systems* 16, 3 (1991), 335–343.

[17] Alina Hang, Alexander De Luca, and Heinrich Hussmann. 2015. I know what you did last week! do you?: Dynamic security questions for fallback authentication on smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. ACM, 1383–1392.

[18] Alina Hang, Alexander De Luca, Matthew Smith, Michael Richter, and Heinrich Hussmann. 2015. Where have you been? using location-based security questions for fallback authentication. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*. 169–183.

[19] Imgur. 2018. Imgur Api Version 3. https://api.imgur.com, Last accessed on 2018-11-30.

[20] Mike Just. 2004. Designing and evaluating challenge-question systems. *IEEE Security & Privacy* 2, 5 (2004), 32–39.

[21] Mike Just and David Aspinall. 2009. Personal choice and challenge questions: a security and usability assessment. In *Proceedings of the 5th Symposium on Usable Privacy and Security*. ACM, 8.

[22] Rohit Ashok Khot, Kannan Srinathan, and Ponnurangam Kumaraguru. 2011. Marasim: a novel jigsaw based authentication scheme using tagging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 2605–2614.

[23] Keeyoung Kim and Simon S Woo. 2018. When George Clooney Is Not George Clooney: Using GenAttack to Deceive Amazon's and Naver's Celebrity Recognition APIs. In *IFIP International Conference on ICT Systems Security and Privacy Protection*. Springer, 355–369.

[24] Benjamin Laxton, Kai Wang, and Stefan Savage. 2008. Reconsidering physical key secrecy: Teleduplication via optical decoding. In *Proceedings of the 15th ACM conference on Computer and communications security*. ACM, 469–478.

[25] Brent MacRae, Amirali Salehi-Abari, and Julie Thorpe. 2016. An exploration of geographic authentication schemes. *IEEE Transactions on Information Forensics and Security* 11, 9 (2016), 1997–2012.

[26] George A Miller. 1995. WordNet: a lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

[27] Morris Moscovitch and Fergus IM Craik. 1976. Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior* 15, 4 (1976), 447–458.

[28] Douglas L Nelson, Valerie S Reed, and John R Walling. 1976. Pictorial superiority effect. *Journal of Experimental Psychology: Human Learning and Memory* 2, 5 (1976), 523.

[29] Allan Paivio. 2006. Mind and its evolution; A dual coding theoretical interpretation. *Mahwah, NJ* (2006).

[30] Greg Pass, Ramin Zabih, and Justin Miller. 1997. Comparing images using color coherence vectors. In *Proceedings of the fourth ACM international conference on Multimedia*. ACM, 65–73.

[31] J Peeck. 1974. Retention of pictorial and verbal content of a text with illustrations. *Journal of Educational Psychology* 66, 6 (1974), 880.

[32] Ariel Rabkin. 2008. Personal knowledge questions for fallback authentication: Security questions in the era of Facebook. In *Proceedings of the 4th symposium on Usable privacy and security*. ACM, 13–23.

[33] Stuart Schechter, AJ Bernheim Brush, and Serge Egelman. 2009. It's no secret. measuring the security and reliability of authentication via "secret" questions. In *2009 30th IEEE Symposium on Security and Privacy*. IEEE, 375–390.

[34] Stuart Schechter, Serge Egelman, and Robert W Reeder. 2009. It's not what you know, but who you know: a social approach to last-resort authentication. In *Proceedings of the sigchi conference on human factors in computing systems*. ACM, 1983–1992.

[35] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2010. Encountering stronger password requirements: user attitudes and behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*. ACM, 2.

[36] Adam Stubblefield and Dan Simon. 2004. Inkblot authentication. *Microsoft Research* (2004).

[37] System Usability Scale. 2018. SUS. https://www.usability.gov/, Last accessed on 2018-11-30.

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[39] Julie Thorpe, Brent MacRae, and Amirali Salehi-Abari. 2013. Usability and security evaluation of GeoPass: a geographic location-password scheme. In *Proceedings of the Ninth symposium on usable privacy and security*. ACM, 14.

[40] Blase Ur, Jonathan Bees, Sean M Segreti, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Do users' perceptions of password security match reality?. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 3748–3760.

[41] Paul C Van Oorschot and Tao Wan. 2009. TwoStep: An authentication method combining text and graphical passwords. In *International Conference on E-Technologies*. Springer, 233–239.

[42] Kim-Phuong L Vu, Robert W Proctor, Abhilasha Bhargav-Spantzel, Bik-Lam Belin Tai, Joshua Cook, and E Eugene Schultz. 2007. Improving password security and memorability to protect personal and organizational information. *International Journal of Human-Computer Studies* 65, 8 (2007), 744–757.

[43] Xiaoyong Yuan, Pan He, Qile Zhu, Rajendra Rana Bhat, and Xiaolin Li. 2017. Adversarial Examples: Attacks and Defenses for Deep Learning. *arXiv preprint arXiv:1712.07107* (2017).