Mental Trapdoors for User Authentication on Small Mobile Devices

Eiji Hayashi, Nicolas Christin, Rachna Dhamija, Adrian Perrig

August 12, 2007
CMU-CyLab-07-011

CyLab
Carnegie Mellon University
Pittsburgh, PA 15213

Mental Trapdoors for User Authentication on Small Mobile Devices

Eiji Hayashi
CMU/CyLab Japan

Nicolas Christin
CMU/CyLab Japan

Rachna Dhamija
Harvard University

Adrian Perrig
CMU/CyLab

ABSTRACT

As small mobile devices such as mobile phones become increasingly sophisticated, they are beginning to be used for highly security-sensitive applications such as payment systems, stock trading, and access control systems. The increasing importance of mobile phones exposes the tremendous lack of access control systems that restrict access to the legitimate user. In fact, a lost mobile phone "delegates" all rights to its new owner. The main challenges in designing a secure user authentication system for small mobile devices are the miniaturization as well as the requirement for usability across a wide range of people.

In this paper, we propose and evaluate a novel mechanism for user authentication. The cognitive process we rely on is the human ability to recognize degraded images; degraded images are easily recognized by legitimate users who have been being exposed to the original picture. On the other hand, without knowledge of the original image, it is difficult to mentally "revert" from the degraded image to the original image, which provides a line of defense against guessing attacks.

We implement a prototype user authentication system in Nokia N70 cellular phones, and conduct a usability study of our scheme with 54 participants. We find that all users manage to authenticate, even after four weeks, which is a strong indication that the scheme is usable by a wide range of people, even on miniaturized portable devices.

We anticipate that this research will revitalize and encourage research in the important topic of portable device based user authentication.

1. INTRODUCTION

Secure user authentication continues to be the weakest link in today's security systems, especially in the case of small portable devices such as mobile phones. The compact form factor of such devices severely limits the amount of data that can be displayed, and makes data entry challenging. For instance, the fact that the keyboard is generally limited to a dozen keys (numeric pad and a few function keys) makes it impractical to enter long text messages. At the same time, the increased integration of electronic wallet functionality (e.g., [7]) in portable devices, or the use of cellular phones as banking terminals [1] all rely on the crucial assumption that the person using the device is the legitimate owner. Owner authentication is therefore of utmost importance.

Knowledge-based schemes, where the user authenticates herself by proving she *knows* an authentication token, remain, given current technology, the most likely candidate for authentication on small mobile devices. Indeed, alternative solutions (e.g., biometrics, which characterize who the user *is*), despite recent advances, are still costly to implement if one wants to avoid potentially high error rates and vulnerability to certain attacks [19].

This paper poses the question whether it is possible to devise authentication schemes that are usable by a wide range of people, while being more secure than PIN or password based authentication. To help us answer this question, cognitive psychology research has shown that images are significantly easier to remember than text [14, 23]. Thus, image-based authentication schemes (e.g., graphical passwords) have been proposed as an alternative to textual password authentication. [1] Unfortunately, existing graphical password schemes have met with limited adoption. An impediment to the deployment of image-based authentication schemes lies in their non-negligible error rates, be they failure for a legitimate user to authenticate herself, or success for an attacker in impersonating a legitimate user. In addition, some image-based authentication schemes require relatively good display capabilities, which is at odds with the limited graphical capabilities offered by portable devices. Even considering technological advances that allow small displays to have finer resolutions, limitations of the human eye may impede the usability of graphical authentication systems on small portable devices.

To arrive at a viable user authentication system for small mobile devices, we argue that we need to rely on a cognitive process that is different from that used by traditional authentication schemes.

The cognitive process we rely on is the human ability to recognize degraded images; people are able to capture the salient features from images they see, and they can easily recognize these features even when a considerably degraded version of the image is presented to them. It is on the other hand difficult to mentally "revert" from the degraded image to the original image, without being exposed to the original picture.

Using an analogy with cryptographic terminology, one can view the degradation of an image as a one-way function. By associating the original picture and the degraded version of that picture, the legitimate user can create a "mental trapdoor," which enables her to recognize the degraded version even when the original picture

Carnegie Mellon University CyLab technical report. August 2007.
.

[1] Monrose and Reiter provide an overview of graphical password schemes ([8], chap. 9).

is not shown.[2] Shortly stated, a legitimate user "sees" something that *is not there*, and this gives her an advantage over unauthorized parties.

We hypothesize that such mental trapdoors can be used to design authentication schemes that reconcile usability with strong security. This cognitive process, in itself, is not dependent on the use of portable devices, and can in fact be used in a wide range of contexts (computer systems, ATMs, etc). Using degraded images for authentication is nevertheless particularly useful in the context of portable devices. Compared to photographs, or even Random Art, degraded images are largely unaffected by a low graphical resolution, which makes an authentication scheme based on them well-suited to portable devices with limited graphical capabilities.

We make the following contributions in this paper. We introduce the concept of mental trapdoors, and substantiate their applicability to user authentication. We then focus on applying mental trapdoors to user authentication in small portable devices. To that effect, we design and analyze a novel authentication scheme, *Fata Morgana*. We present a proof-of-concept implementation of Fata Morgana in cell phones, and provide evidence of the viability of the scheme through usability testing.

The rest of this paper is organized as follows. Section 2 presents a problem definition in the form of security requirements, and describes how related proposals have attempted to meet them. Section 3 describes mental trapdoors. Section 4 discusses the Fata Morgana scheme. Section 5 describes a usability study and provides empirical measurements of the effectiveness of Fata Morgana. Section 6 lists the numerous challenges that we still face. Finally, we present our conclusions in Section 7.

2. REQUIREMENTS AND RELATED WORK

An ideal authentication system should provide strong security while maintaining very high usability – it should be usable everywhere, by everyone, without the need for any specific training. In this section, we formalize our desired system requirements, for both security and usability. We then discuss the limitations existing proposals face when attempting to address these requirements. Notably, we examine the interplay between human cognitive processes and the desired security requirements.

2.1 Requirements

Security requirements of an authentication system are relatively straightforward:

No unauthorized access. The probability that the authentication system grants access to anybody impersonating a legitimate user should not be higher than that of obtaining the correct selection of authentication tokens by choosing them at random. For instance, a four-digit PIN authentication system where a user has three chances to enter their PIN would satisfy these properties only if the probability an impostor gains access to the system does not exceed 3/10000.

The usability requirements an authentication system has to fulfill to be applicable to small mobile devices such as cellular phones are rather stringent:

Universal use. The authentication system should produce consistent results regardless of the physical environment in which it is used. For example, the system should not be affected by lighting conditions or surrounding noise.

Limited scale. The authentication system should be usable on typical cellular phones or devices of similar size and capabilities. To be more precise, the authentication system should work with small color displays of relatively low resolution (e.g., 350x400), and at most a dozen of keys (e.g., numeric pad and function keys). Processor and memory requirements should not exceed the resources available in medium-priced cellular phones.

Limited training. The authentication system should not require users to undergo a long or intensive training session prior to first use. Simply stated, ideally, the system should be usable in minutes by a new user with minimal external assistance, such as that available from a busy sales clerk.

Rapid setup. Creation of a new authentication token should be fast, so that users can easily renew authentication tokens.

Fast authentication process. A legitimate user should be able to authenticate herself to the system in seconds.

Low error rate. The number of occurrences where the authentication system denies access to a legitimate user should be held at low levels, comparable with that observed in PIN-based authentication systems where users have the ability to select the PINs at their entire discretion. This not only implies that the authentication system itself must have a negligible (or null) error rate, but also that legitimate users' mistakes in the authentication process should be highly infrequent. For a knowledge-based authentication system, this requirement demands that the authentication tokens be easily memorable.

In essence, the above requirements attempt to characterize a system that 1) is well-suited to small mobile devices, and 2) has the same ease of use as a PIN-based authentication system, without suffering from the traditional weaknesses of password-based systems. In particular, the predictability of user-chosen PINs defeats the *no unauthorized access* requirement, as an attacker with some knowledge of the user's preferences may be able to greatly reduce the number of possible PINs. Conversely, the difficulty to memorize randomly assigned PINs may confuse legitimate users, and prevent them from authenticating successfully, thereby defeating the *low error rate* requirement. The low memorability of randomly assigned PINs and its effect on authentication failures is documented in [9].

We note that small portable devices, while presenting stringent usability requirements, also actually reduce some of the potential threats. Because of the compact form factor, observation attacks (e.g., shoulder surfing or recording of authentication sessions) are harder to conduct, compared to, say, what could happen at a bank ATM. We will discuss possible attacks in more details in Section 4.

2.2 Related work

Biometric authentication may satisfy the *limited training*, *rapid setup* and *fast authentication process* properties, but, when within the constraints posed by portable handsets, existing technology can generally not accommodate all of the *universal use*, *low error rate* and *no unauthorized access* properties simultaneously.[3] In the rest of this section, we therefore focus on knowledge-based authentication systems.

Knowledge-based authentication systems usually satisfy *limited training*, *fast authentication process*, *universal use*, with some exceptions noted below. The question then becomes: how good are the proposed schemes at avoiding unauthorized accesses and limiting mistaken access denials, especially while taking into account the compact cellular phone form factor?

The limits of human memory is a central factor that influences

[2] Despite similarities in names, the concept of mental trapdoor is therefore quite different from the cognitive trapdoors discussed in [21] as we detail later.

[3] For instance, voice authentication may have significant error rates in noisy environments, facial recognition schemes are sensitive to variations in lighting conditions, and fingerprint readers can be defeated by fake fingerprints [19].

the design of knowledge-based authentication systems. Traditional passwords and PINs depend on *recall*, the ability to remember items from memory without help, a skill that humans are not inherently good at. Even if it were possible to easily enter long strings of random symbols into cellular phones, it is not easy for people to remember them.

Based on the observation that humans are considerably better at remembering images than they are at remembering text [23], much research (e.g., [9, 17, 22]) has been devoted to graphical password authentication as one alternative to traditional (textual) passwords. Different graphical password systems attempt to address the limits of human memory by relying on different cognitive processes.

Some graphical password schemes also rely on recall. For example the Draw-A-Secret scheme requires users to draw an image on a grid in order to authenticate [17]. This is a recall task, because the user has to recollect the drawing from memory during each authentication. Thorpe et al. show that the actual Draw-A-Secret password space may be smaller than theoretically possible, because users have a tendency to choose mirror symmetric passwords with a small number of strokes [26, 27].

Thus, graphical passwords based on recall may have many of same problems of textual passwords based on recall, and they do not satisfy our *no unauthorized access* requirement. Furthermore, even though some smartphones have a stylus, most cell phones do not have pointing devices, so this scheme does not satisfy our *limited scale* requirement.

Other authentication systems rely on *cued recall*, where clues are provided to the user to aide the recollection task. Cues can be provided in the form of text or images. For example, some authentication schemes provide textual password hints to help users remember their password or challenge questions. Some graphical password schemes require users to authenticate by clicking on points, or selecting regions, of an image that were previously chosen by the user [3, 22, 30]. In this case, the image itself serves as a clue to the regions of the image that a user must recall. This scheme also produces passwords that are predictable; because users have to recall the points they have chosen, there tend to be "hot spots," or regions that are often selected because they are most memorable or most obvious [28, 30]. Such click-based schemes generally need a relatively wide display to provide a large enough key space, and requires a pointing device. As such, they do not fulfill our *limited scale* requirement.

It is known that *recognition*, the ability to remember items we have seen before, is an easier memory task than recall. In particular, humans have an impressive ability to recognize pictures they have seen before, and they recognize pictures better than they recognize words or sentences [23]. Research has shown that individuals can distinguish large sets of previously seen pictures from new "distractor" pictures at high levels of accuracy [23, 24]. Many graphical password schemes rely on this skill by asking users to select previously chosen images from a larger subset of images [5, 9].

Research shows that people remember images more accurately when they are semantically meaningful and when the images are generated by people themselves [18]. Therefore, graphical password images self-chosen by users may be more memorable. However, given some knowledge about the user, self-chosen images are also easier for an attacker to predict [20], which defeats our *no unauthorized access* requirement. One countermeasure is to assign pictures to users rather than allow them to choose images themselves. Another option is to use abstract images, which are less predictable than real images [9]. However, it has been shown that image scenes that are coherent and semantically meaningful are stored more accurately than incoherent or abstract images [4, 11]. A third alternative, which is closest to our proposal, is to assign blurry images [15]. Harada et al. [15] propose a graphical password scheme with obscured portfolio pictures to prevent shoulder surfing. The goal is to minimize the risk of unauthorized accesses by minimizing the information a portfolio picture shows to a user. While the approach is seemingly effective at preventing shoulder surfing, the obscured pictures are difficult to memorize, and authentication failure rates of legitimate users are expected to be high. Thus, it seems that there is a trade-off between the memorability and predictability of images.

Stated differently, previously proposed schemes seem to be unable to minimize both the risk of unauthorized accesses and authentication error rates simultaneously. A recently proposed graphical authentication scheme is based on computing a "path" of known pictures [29]. This approach seemed promising, with error rates in the order of 5%, and no unauthorized accesses, but has recently been shown to be vulnerable to an eavesdropping attack [12] leaving open the challenge of designing a cognitive authentication scheme resilient to eavesdropping attacks. In addition, the scheme proposed in [29] requires a large screen, authentication reportedly takes on the order of minutes, and users need to sustain significant training prior to use.

3. MENTAL TRAPDOORS

We design a graphical password scheme that allows the use of images that are self-chosen and familiar to the user, yet that are not easily predictable by an attacker. The goal of this section is to provide an overview of our approach and the cognitive process on which it relies.

In our system, we allow users to select their own graphical password images. In the context of mobile devices, the user can perform this selection by capturing images with a camera embedded to the mobile device, e.g., a cell phone camera. Psychology studies show that images that are self-generated are recognized better than those that are not [18]. In addition, users enjoy the ability to select and personalize their image portfolios, and that they tend to choose images that are semantically meaningful to them [9, 10, 20].

Next, we distort the images using a non-photorealistic rendering algorithm that eliminates most details in the image, while preserving some features such as color and rough shapes. Because information is lost in the rendering algorithm, it is impossible to mathematically revert back to the original image from its distorted version. As such, the distortion function is analogous to one-way functions used in cryptography.

We then prime the user during a training session to associate the distorted image with the original image and the meaning of that image. During the priming phase, we display the original and distorted pictures side-by-side, and ask the user to practice selecting their images from a set of "distractor images."

To authenticate, the user must choose her own distorted images from a set of distractor images. We rely on the fact that human perception is affected by what we know. Once a user knows that a image is a distorted version of particular subject (e.g., of her child), it becomes hard to perceive the image as simply a collection of colored blobs. When the meaning of the image is known, our brains impose that knowledge on our perception and it becomes hard to interpret the picture in another way [13]. Furthermore, there is evidence that the ability to recognize objects in degraded images increases dramatically with familiarity of the subject in the image (this effect is very strong in the case of faces [6, 16]). During the authentication task, the user can recognize the original objects in the image by using color and shape cues and by remembering the semantic meaning that she previously associated with the image. In

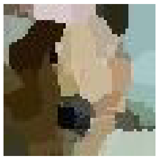


Figure 1: Example of the relationship between original and distorted pictures. While it is obvious that this figure is a distorted version of Figure 2 when seeing both pictures, it is difficult to infer Figure 2 from this figure without having seen Figure 2 before.

other words, the training phase allows the user to associate the distorted picture with its original, thereby creating a "mental trapdoor" that allows her to revert from the distorted picture to the original, even though such a transform is mathematically impossible.

Leaving mathematics aside, it is also difficult to mentally "revert" from the degraded image to the original image, without being exposed to the original picture. To illustrate this effect, try the following experiment. Consider Figure 1, which is a distortion of a photo. Try to guess what the original photo represents. Then, compare your guess to Figure 2, on the next page. To someone who has seen both pictures, it is clear that Figure 1 represents the object in Figure 2, and that seeing Figure 1 evokes the original object. However, without having been exposed to Figure 2, in Figure 1 does not carry much meaning. You can repeat the experiment with different people. You may hear amusing guesses, but it is unlikely that you will obtain a answer that is even approximately correct, unless the person has prior knowledge of the subject of the picture and makes an educated guess.

We point out that the cognitive process used is more general than the related proposal for "unclear images" put forth by Harada et al. [15]. Harada et al. rely on the faculty humans have to infer contours (i.e., edges) from noisy images to extract salient features and facilitate recognition of the picture subject, while the mental trapdoors we advocate rely on the more general ability to associate severely degraded pictures with their original picture after having been primed with the original image. Further, by adding noise to the original image to yield an unclear image, the unclear image may still contain significant amounts of information from the original picture, that might even be extracted automatically, for instance, by an "edge detect" filter. Returning to our cryptographic analogy, the specific one-way distortion function used in [15] may leak information, and could even be, to some extent, reversible.

We instead advocate to remove as much information as possible from the pictures used for authentication, so that only the user can revert back to the original picture through mental association. We hypothesize that such mental trapdoors can be used to design an authentication scheme that reconciles usability with strong security.

4. FATA MORGANA

We next describe how we use mental trapdoors to devise an authentication scheme that attempts to fulfill all the requirements of Section 2. While Fata Morgana can be implemented in any computing system equipped with a graphical display and an input device, our focus in this paper is to take into account the specifics of mobile devices such as cellular phone in the design, especially the limitations due to miniaturization.

We first discuss the overall architecture of Fata Morgana, before elaborating on the threat model and possible attacks against the authentication scheme. We then provide details about our prototype implementation for Nokia N70 smartphones.

4.1 System architecture

Fata Morgana is a knowledge-based authentication system, which uses images as authentication tokens. During setup (e.g., immediately after purchase of the cell phone), the user selects a personal portfolio of p pictures. To authenticate, we present the user with a challenge set of n pictures, where $n > p$. The user has to correctly identify the p pictures within the challenge set that belongs to her portfolio.

Fata Morgana consists of three phases: portfolio creation, practice and authentication.

Portfolio creation phase To create an image portfolio, the user first has to create a set of p images. To enhance memorability, users are expected to build their portfolio from scratch, rather than using default pictures. Here, we ask the users to take p photos they want to use for authentication. The photos should be taken in as secure an environment as possible; in particular, these photos should not be transferred to the authentication device using an insecure channel. Ideally, the user should be able to use the authentication device itself to capture the images. For instance, when Fata Morgana is deployed on a mobile phone, the mobile phone camera is the best option.[4]

Once the photos are taken and passed to the authentication device (e.g., the cell phone itself), they are distorted using a lossy filter. Using a lossy filter ensures that it is mathematically impossible to revert from the distorted image to the original image. The transform is performed on the authentication device.

Fata Morgana does not mandate a specific type of filter. The selection of the "best" filter possible is an area that warrants further investigation, but cognitive research can give us some heuristics: Rough shape and colors of the original picture should be preserved through the distortion process to make the distorted picture more memorable and facilitate the creation of mental trapdoors. As we will discuss in the presentation of our prototype, we currently use an oil-painting filter, which satisfies these characteristics. Moreover, when considering small portable devices like cellular phones or PDAs, the constraints imposed by the size and quality of the display require that the distortion filter can work with low resolutions. In other words, a distortion filter that yields a high resolution output is probably unsuitable.

The resulting set of p distorted pictures is assigned to the user as her image portfolio. The original pictures and distorted pictures are shown simultaneously to the user, so that the user can mentally associate the distorted pictures with their original meaning, thereby creating the mental trapdoor needed to invert the one-way lossy filter. Figure 3 is a screen shot of the image portfolio assignment. The top row contains the three photos taken by the user. The bottom row contains the distorted versions of the photos, which will be used in the authentication phase.

Concomitant with the portfolio creation phase, the authentication device selects $(n-p)$ pictures to be used as decoys during the authentication phase. The $(n-p)$ decoys *do not change* until the portfolio images are recreated, to prevent so-called intersection attacks. We elaborate on intersection (and other) attacks later.

Two approaches are possible to generate decoys. One approach is to generate "synthetic" decoys purely algorithmically, that is, without using any original photo. Another approach is to generate decoys by applying the lossy filter to a set of existing photos

[4] While there are still some mobile phone models that do not have an embedded camera, these models are becoming quickly obsolete, and generally do not support applications (e.g., electronic wallet) that require strong user authentication.

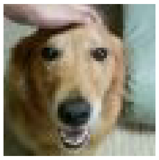
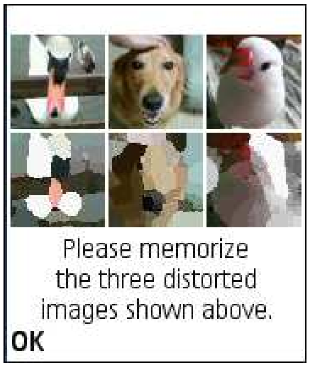


Figure 2: Example of the relationship between original and distorted pictures. This figure represents the original picture, before being distorted to obtain Figure 1.



Please memorize the three distorted images shown above.

Figure 3: Image portfolio assignment. In this example, the user has been asked to take $p = 3$ pictures and is subsequently presented with the three pairs of original images-distorted images. The distorted images will be used as authentication tokens in the authentication phase.

that were not chosen by the user. Generating convincing synthetic decoys remains an open problem. Indeed, we found that synthetic decoys look very different from distorted images and could easily be detected, thereby immediately revealing the pictures chosen by the user. Hence, we suggest using distorted versions of photos the user did not choose. In a cell phone, the manufacturer could store a number $N \gg n$ of decoys to secure memory at manufacturing time. As long as the decoys stored in the cell phone during the manufacturing process are taken at random from a database of $N' \gg N$ images that evolves over time, decoys should be fairly hard to identify.

Training phase After the portfolio creation phase, the system conducts a short training phase to improve memorability of the portfolio images. In this phase, the challenge set is presented to the user. The user can practice selecting their image portfolio, and the system provides immediate feedback as each image is selected on whether the choice is correct. In addition, the user can at any time access original-distorted image pairs as shown in Figure 3.

The training phase should be conducted in as secure an environment as possible. After the training phase is complete and the user is confident that she remembers her images, the original pictures used in the generation of the distorted pictures should be removed from the cell phone memory.

Authentication phase During the authentication phase, the user must correctly select her p portfolio pictures from the challenge set. The decoy images themselves are produced using original pictures, and the distortion levels are high enough that most details of the original pictures are obscured. Therefore, an outsider will have a very hard time identifying which pictures belong to the portfolio, even if she possesses information about the user's personal preferences. Note that, contrary to the decoy database, the challenge set of n pictures (portfolio and chosen decoys) presented in the authentication phase does not need to be stored in secure memory.

4.2 Attacks and countermeasures

We next investigate how Fata Morgana addresses possible attacks aimed at impersonating a legitimate user.

Brute force attack The simplest attack is to try to randomly guess the correct portfolio. With a challenge set of n pictures, and a portfolio of p pictures, the probability that a single random guess succeeds is $1/\binom{n}{p}$.

Obviously, if we allow the impostor to try all possible $\binom{n}{p}$ combinations, then she will eventually manage to fraudulently authenticate. To prevent such an undesirable outcome, we use a reference counter that locks the device after t failed authentication attempts, similar to systems used in automated teller machines.

The probability that attacker can impersonate the user within t trials, using a succession of random guesses, is $t/\binom{n}{p}$. For example, for $(n,p,t) = (27,3,3)$, we get $3/\binom{27}{3} \approx 0.001$. We can adjust (n,p,t) according to the desired failure probability. For instance, increasing n to $n = 36$ yields a failure probability less than that of a random four-digit PIN-based system.

Educated guess In an educated guess attack, the impostor tries to guess the user's portfolio pictures based on previously obtained information about the user, e.g., through social engineering.

For instance, in an image-based authentication scheme that allows a user to take picture by herself and to use the picture as is as part of her portfolio, the following scenario is possible. Assume the attacker has previously learned that the user owns a white dog, Fifi. If the attacker finds a picture of a white dog in the challenge set, the attacker can guess the picture is actually included in the user's portfolio.

In Fata Morgana, similar to the above example, users create their own pictures. However, Fata Morgana never uses the original pictures beyond the practice phase (and the original pictures are, in fact, destroyed as soon as practice is completed). Because Fata Morgana only utilizes distorted images resulting from the lossy filter transform, an educated guess attack is less likely to succeed. In our above example, the distorted picture of Fifi, used in the user's portfolio, is nothing more than a predominantly white blob. Even if the tentative impostor knows of the existence of Fifi, figuring out that the white blob originally came from a picture of Fifi, and not from a picture of a chicken or a snowman, may be difficult. As such, the educated guess attack is more difficult to carry out in Fata Morgana than in some other image-based password authentication schemes. Of course, more difficult does not necessarily mean impossible: if familiar objects can be easily identified even after distortion, for instance, if Fifi is dyed in pink, the presence of a pink blob in the challenge set may yield a clue the image may be part of the user's portfolio.

Variants of the educated guess attack exist. A car fanatic may select a portfolio consisting solely of images of cars. For an outsider aware of the user's fascination of automobiles, the presence of p pictures that look like cars may allow her to reduce the number of possible portfolio pictures. Hence, it is desirable to recommend that the original pictures taken by the user present significant differences between each other.[5] We advised participants in our usability tests to select photos evoking different subjects.

Observer attack In an observer attack, the attacker identifies the

[5] This also enhances memorability of the portfolio as people generally assign categories to help memorize each given picture.

pictures in the user's portfolio by observing authentication procedures of the user. Observer (or shoulder surfing) attacks are currently one of the most significant threats to user authentication. Reports of ATM equipped with rogue video cameras recording authentication sessions have circulated for a while [2]. We propose two countermeasures to mitigate the threat of observer attacks.

First, we constantly change the respective positions of the decoys and portfolio pictures on the authentication screen(s), so that the authentication pattern cannot be inferred from observing which keys are pressed. This line of defense is particularly useful when combined with the use of optical filters that render the display difficult to observe from a distance.

Second, we avoid showing any hints regarding the picture selection. In particular, no feedback is given to the user when a picture is selected. In a cell phone environment, we have the added benefit that observing which keys are pressed from a distance is difficult, given the small size of the keyboard. Without any correlation between the keys pressed and the output on the display, an observer cannot identify the user's portfolio.

While we do believe these defenses will be adequate for small mobile devices, we can consider the following extension to Fata Morgana to make it resilient to observer attacks that can be conducted on larger devices such as ATM terminals. Rather than asking the user to select her pictures from a challenge set, we may simply ask "Is one or more of your portfolio pictures present on this screen?" and repeat the query a few times. In essence, this modified algorithm resembles closely a zero-knowledge proof, in a manner similar to the method Roth et al. proposed for PIN entry [21], and is more resilient to observer attacks. We defer a more thorough investigation of such a variant on Fata Morgana to future work, but note that this type of authentication primitive can make full use of mental trapdoors.

Intersection attack In an intersection attack, the impostor identifies the legitimate user's portfolio pictures by observing multiple authentication procedures; the intersection of all of the images that are observed will reveal the user's portfolio [9].

For simplicity, assume a challenge set of size $n = 3$, and a portfolio set of size $p = 1$, and denote the portfolio picture by X. On a first authentication attempt, the challenge set may consist of $\{A, B, X\}$. Now, if decoys are changed from one authentication session to the next, the challenge set proposed in a subsequent authentication may be $\{D, X, C\}$. An impostor only needs to look at both challenges to figure out the portfolio picture is $\{A, B, X\} \cap \{D, X, C\} = \{X\}$.

In Fata Morgana, we resist the intersection attack by always maintaining identical decoys in each authentication challenge. In the above example, the challenge set would always be $\{A, B, X\}$. A and B would only be replaced when the user decides to change her portfolio image from X to Y. Therefore, each authentication challenge is always the same, and the intersection of all challenges reveals nothing about the portfolio.

In summing, Fata Morgana appears resistant to brute force, guessing, prediction, and casual shoulder surfing. Fata Morgana is more vulnerable to prolonged observation and spyware, even though some of the countermeasures used against observer attacks increase the difficulty of carrying out such attacks. For instance, a simple keylogger would not work since the order of pictures in the challenge set changes from one authentication session to the next; it is also necessary for the spyware to simultaneously record the contents of the display.

A criticism of Fata Morgana is that, per the above discussion, it only provide marginally more resilience against observer attacks than PINs. We point out that the main objective of Fata Morgana is to thwart social engineering and guessing attacks, which are very easy to carry out with self-chosen passwords or PINs, while avoiding memorability problems encountered with randomly assigned PINs or passwords.

4.3 Prototype

We next describe the prototype we implemented as a proof-of-concept, and to conduct usability tests. The prototype is implemented in Java on Symbian OS, and has been tested on Nokia N70 cell phones. The reason behind the selection of the Nokia N70 handset lies in the fact its core features (display, camera, keypad, memory) are fairly common, while at the same time being of reasonably high quality. Both web-based and Windows versions of the prototype are publicly available at `http://arima.okoze.net/fyeo/`, and attempt to emulate the interface available on the Nokia N70.

In the prototype, a user is assigned three portfolio pictures ($p = 3$), derived from photos she has taken herself, as shown in Figure 3. In the authentication phase, three sets of nine pictures are presented to the user as a challenge set ($n = 27$). The user is given three attempts to login ($t = 3$). After three consecutive authentication failures, the phone is locked. The probability that a random guess results in a successful authentication is $3/\binom{27}{3} \approx 0.001$. We can lower that probability by increasing the number of portfolio pictures and/or decoys.

Figure 4 represents the challenge set for the user who had selected the portfolio shown in Figure 3. The challenge set is divided in three authentication screens containing nine pictures each. On a given authentication screen, the top indicator denotes which authentication screen the user is currently seeing. As long as they are not authenticated, users can go forward or backward between the different authentication screens, using two function keys present on the handset. The nine challenge set pictures in each screen are arranged in a 3x3 grid, and each of the nine pictures is associated with a numeric key. Some screens may only contain decoys, as is the case of the leftmost screen in this example, while some screens may contain more than one portfolio image, as in the case of the middle screen here. Last, the "Clear" button, accessible by a third function key on the handset, allows users to reset their current selection and start the authentication process afresh.

In the case of Figure 4, based on the portfolio of Figure 3, a correct authentication sequence is "NEXT," (since the first screen only contains decoys), followed by 7, 8, which are the two portfolio pictures on the second page, "NEXT," and 5, which is the only portfolio picture present on the third page. Fata Morgana does not impose any specific order in the inputs. For instance, "NEXT, NEXT, 5, PREV, 8, 7," which allows to first peruse the entire challenge set, authenticates as well.

As discussed above, to avoid intersection attacks, the same challenge set will be presented as long as the portfolio remains unchanged. On the other hand, the order in which pictures are presented will differ from one authentication session to the next. For instance, in the next session, the portfolio could map to key 1 on each of the three screens, while using the same decoys. Any permutation on the order of the 27 pictures composing the challenge set is possible.

Image processing filter In our prototype, we select an *oil-painting filter* as the lossy filter used to generate portfolio images from the original photos input by the user. A oil-painting filter blurs the edges and the colors of the processed picture while preserving the main shapes and colors of the original picture. The oil painting filter can be viewed as a function parameterized by two integers: the number of "bins" ν, and the brush size β. Shortly stated, these two integers determine how aggressively the filter will attempt to dis-

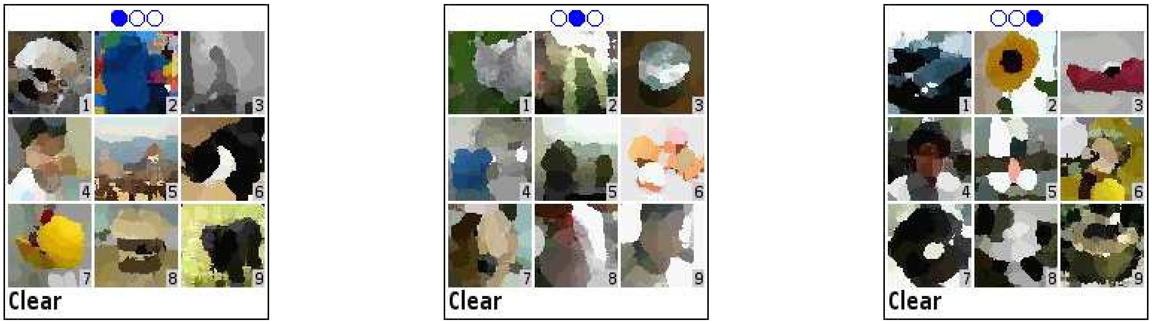


Figure 4: Example of challenge set. This series of screenshots represents a challenge set as presented to the user. Each picture maps to a given key on the handset, and navigation between the different screens is possible through function keys.

tort the original picture. (The precise meaning of these parameters is described in Appendix A.)

Properly tuning ν and β is an important and non-trivial task: If the distortion level is too high, a legitimate user cannot easily recognize a distorted picture even when that distorted picture has been created from a self-chosen original picture. If the distortion level is too low, an attacker may be able to guess the subject of a distorted picture, without knowing the original picture. In order to determine suitable values for the parameters ν and β, we conducted low-fidelity tests, which we discuss in the next section.

5. USABILITY TEST

We conducted usability tests to help in the design and evaluate the effectiveness of the proposed scheme. We began by conducting an informal, *low fidelity* test to determine the optimal parameters of the oil painting filter used in our prototype. Low-fidelity testing is a technique that is used to obtain feedback on mock prototypes early in the design process.

We then conducted a formal usability experiment, using a between-subjects design. Participants were divided into three groups to test one of three portfolio types: self-chosen images, self-chosen images that were distorted, and distorted images that were assigned to participants. We asked participants to login during four sessions that spanned four weeks. During each session, we measured the number of successful logins and the login time for each authentication. We also collected qualitative data that was provided by participants in questionnaires after the study.

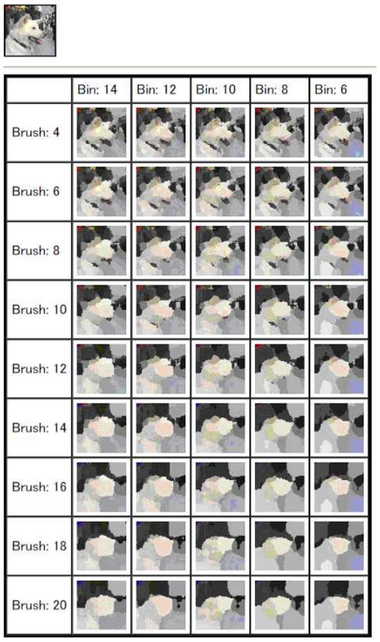


Figure 5: Low-fidelity test image sheet. The sheet was used to help determine how to select the oil-painting filter parameters.

5.1 Low-fidelity testing

We used a low-fidelity test with six participants, in order to determine the optimal parameters for the oil painting filter used in our prototype. We showed each user a sheet of paper that contained an original picture and a set of distorted pictures, as shown in Figure 5. Each picture is a 56x56 pixels image.

The two configuration parameters used in the oil-painting filter, number of bins and brush size, are varied along the horizontal and vertical axes, respectively. The pictures progress from least distorted (top left corner) to most distorted (bottom right corner).

We conducted the test in two phases; we asked participants to recognized distorted images without knowledge of the original image in the first phase, and with knowledge of the original image in the second phase.

We began by completely covering the sheet of distorted images with another piece of paper. We uncovered the bottom row, and asked the participant if they could recognize the image. We then proceeded to uncover each row one at a time, until the participant could recognize the subject in the image. This procedure allowed us to determine the lower bound of distortion that prevents users from recognizing a distorted image, when they have no knowledge of the original image.

Next, we showed participants an original image. Starting from the top left distorted picture, we asked the participant when they could no longer recognize the image. Based on this test, we could obtain a rough upper bound on the level of distortion up to which users with knowledge of the original can still recognize a distorted image.

We repeated this procedure for five different sets of pictures. We observed the following:

- The number of bins did not affect the recognition task.
- Participants could not recognize distorted pictures when they did not know the original pictures if the brush size was greater than 4.
- When they knew the original picture, participants could not recognize distorted pictures if the brush size was greater than 8.

Higher brush sizes reflect greater distortion. We discovered that prior knowledge of the image increased the level of distortion that we could apply before the image became unrecognizable. According to the findings, we determined that the optimal brush size should be a value between 6 and 8. In this range, those who know the original picture can recognize its distorted picture and those who do not know it cannot recognize it. Based on this result, we set the number of bins to 16 and the brush size to 8 in the prototype.

5.2 Fata Morgana evaluation

We next conducted a formal usability test to evaluate the Fata Morgana prototype. For this experiment, we recruited participants by posting flyers in four universities. 54 people participated in our study. Of these, 50 were students and 4 were university staff. Nine participants were female and 45 participants were male. All of the students majored in electrical engineering, computer science or another scientific subject. Their ages range from 18 to 29 with a mean age of 23. None of these participants had taken part in our low-fidelity test.

We distributed participants evenly across the three groups in case factors such as memory, eyesight and/or familiarity with cell phones influenced their behavior. That is, we ensured that different groups presented similar age, major, and origin (i.e., university) distributions.

In our study, we aim to study two design parameters of Fata Morgana: 1) the effect of self-chosen portfolio images compared to assigned portfolio images, and 2) the effect of distorted images compared to non-distorted images. If we find that adding image distortion has no effect on the authentication success rate, then Fata Morgana can reap the memorability benefits of previously proposed graphical password schemes [9, 20], while being significantly more resilient to guessing attacks.

To investigate how success rate and login time are affected by the type of images used, we assigned each group a different type of portfolio as follows:

Group 1: We asked participants in this group to take three pictures using a cell phone camera. We then assigned the three pictures, *as is*, as their image portfolio. We also assigned 24 decoy photographic images that were identical for each participant.

Group 2: Group 2 was the "Fata Morgana group." We asked participants in this group to choose three of their own pictures. They could select the three pictures from pictures they had already taken or they could capture new pictures for this usability test. We distorted each of the pictures using oil-painting filter. We then assigned the distorted pictures as the image portfolio. We also assigned 24 decoy photographic images that were distorted using the same filter settings. The decoys were the same for each participant.

Group 3: We assigned participants in this group the same portfolio images that were assigned to participants in Group 2. The only difference was that participants in this group did not take the pictures by themselves, and they did not see the original pictures. The decoy images were the same for each participant. The decoy images we used in Group 3 were identical to those used in Group 2.

Procedure Our usability test spanned four weeks and consisted of four sessions. All sessions were conducted in a university class room. Each of the participants took part in the test individually using the same phone and same version of the prototype. Table 1 shows when each session occurred and the tasks that participants completed in each session. In the first session, we assigned three portfolio images to participants according to group to which they were assigned. Participants then completed the training phase described in Section 4, which lasted up to five minutes.

Session #	Date	Tasks
1	First day	Memorize portfolio Training Authenticate
2	Two days later	Authenticate
3	One week later	Authenticate
4	Four weeks later	Authenticate Fill out questionnaire

Table 1: Tasks completed in each session of the usability test.

After the training phase, we instructed participants to login by selecting their portfolio pictures from a challenge set. As discussed in Section 4, the prototype requires participants to authenticate by selecting their three portfolio images out of a set of 27 images.

In the second, third and fourth sessions, we asked participants to authenticate again by selecting their portfolio. At the end of the fourth session, we asked participants to fill out questionnaires for the sake of qualitative analysis.

Success rate We considered an authentication to be successful if the participant could correct identify their own portfolio images. We gave participants a maximum of three attempts to login.

Table 2 shows success rates measured in the sessions. All participants in Group 1 and Group 2 succeeded in all sessions. However, after four weeks, the success rate of Group 3 decreased to 89%. Participants assigned to use distorted image portfolios, without knowledge of the original image, had lower success rates than the other two groups. These results suggest that the graphical passwords with distorted image portfolios may achieve equivalent error rates to those that use traditional images, but only when the user knows the original image.

	The 1st day	2 days later	1 week later	4 weeks later
Group 1	100%(18)	100%(18)	100%(18)	100%(18)
Group 2	100%(18)	100%(18)	100%(18)	100%(18)
Group 3	100%(18)	89%(16)	94%(17)	89%(16)

Table 2: Percentage of successful logins. The number of participants with successful logins is included in parenthesis– all groups had 18 participants.

Table 3 shows the number of participants who successfully authenticated during the fourth testing session, which occurred four weeks after creating the portfolio. The table shows how many attempts it took to succeed; if a participant could not login within three attempts, they are considered to have failed the authentication task. All participants in Group 1 were able to authenticate on the first attempt. In Group 2, 17 participants succeeded in one attempt, and one succeeded in the second attempt. In Group 3, 13 participants succeeded in the first attempt, three participants succeeded in the second attempt and two participants failed to authenticate. These results suggest that randomly assigned abstract images are difficult for some user to remember, even after a few attempts. However, users may be able to recover better from errors when the original image is known and when the image is self-selected.

	Attempt 1	Attempt 2	Attempt 3	Failed
Group 1	18	-	-	-
Group 2	17	1	-	-
Group 3	13	3	-	2

Table 3: Number of successful authentications after four weeks. The table presents the number of participants that succeeded in each attempt during the fourth session (held four weeks after portfolio assignment).

Login time Table 4 shows the means and medians of the login times in seconds. Left and right values in the cells stand for mean and median respectively. We do not include the results for the participants who could not login at all. The difference in mean login times between groups is not statistically significant. However, we observed that participants in Group 2 required a longer time to login that participants in Group 1, and they required less time to login than participants in Group 3.

There was a considerable gap of 3 weeks between the third session and the fourth session. The login times for Group 1 stay fairly consistent over the course of 4 weeks. The login time for Group 2 increases slightly after the course of a week and then rises by approximately 5 seconds after the three week gap. However, the login time for Group 3 increases by more than 8 seconds after the first day and stays consistently at that level after 4 weeks.

	1st day	2 days later	1 week later	4 weeks later
Group 1	11.5 (9.9)	12.3 (12.3)	12.7 (11.9)	12.5 (12.8)
Group 2	12.4 (11.2)	16.4 (15.9)	14.3 (13.4)	17.9 (16.5)
Group 3	16.7 (14.1)	25.8 (19.0)	25.1 (17.6)	24.7 (16.7)

Table 4: Login time. The table gives the mean time, in seconds, users needed to authenticate. Median login times are given in parentheses.

Table 5 shows the mean number of times participants switched between pages in each test session. As described in Section 4, the 27 portfolio images are displayed on the cellular phone display 9 at a time; the participant must switch to the next page to see the next 9 images. When a participant can easily remember her portfolio, she can quickly choose her image and advance to the next page. As a result, the number of page advances would be at most two. On the other hand, when a participant can not recall her images well, she may have to switch back and forth several times. Obviously, page switching also increases the amount of time required to login. Our results show that on average, after 4 weeks, participants in Group 1 required one more page switch than those in Group 1, while participants in Group 3 required approximately 3 more page switches than those in Group 1.

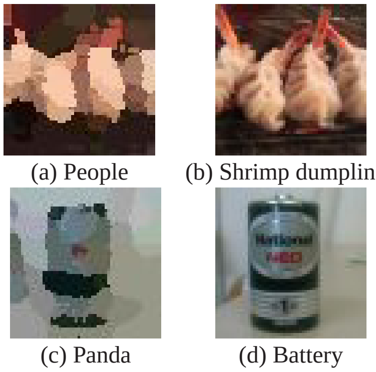
Overall, our results indicate that it is considerably more difficult to remember *assigned* distorted images than self-selected original images. On the other hand, self-selected original images are not made much more difficult to remember by adding distortion.

	The 1st day	2 days later	1 week later	4 weeks later
Group 1	2.28	1.83	2.17	2.22
Group 2	2.00	3.05	2.33	3.28
Group 3	2.94	4.89	4.39	5.11

Table 5: Mean number of page switches for all participants during the fourth session

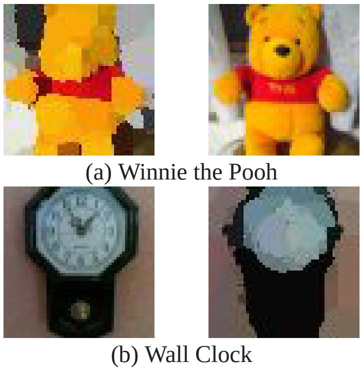
Qualitative results In the questionnaires, we asked all participants to evaluate the ease of recognizing their portfolio images on a scale from 1 to 5, where 1 was “very easy” and 5 was “very difficult”. The average difficulty scores were 1.63 for Group 1, 1.59 for Group 2 and 2.17 for Group 3.

We also asked the participants what technique they used to remember their portfolio images. 14 out of 18 participants in Group 2 replied that that they could recognize their original images within the distorted image and that they memorized the meaning of the image. Four participants said that they simply memorized characteristic colors and shapes within the images. Interestingly, 12 our of 18 participants in Group 3 also said that they assigned a semantic meaning to their portfolio images in order to memorize them, which mirrors observations from Stubblefield and Simon [25] on “inkblot authentication.” Of these, seven participants incorrectly guessed the meaning of the image. Four participants guessed the subject correctly for one out of three portfolio images. Only one participant guessed correctly for all three of her portfolio images. As shown in Figures 6 and 7 incorrect guesses included mistaking a battery for a panda, and shrimp dumplings for people and a battery for a panda. Some of the correct guesses included pictures of Winnie the Pooh and a wall clock.[6] We discuss in more details the factors that can influence the correctness of a guess and its impact on security in the next section.


(a) People (b) Shrimp dumplings

(c) Panda (d) Battery

Figure 6: Examples of incorrect semantic meanings. Users unfamiliar with the original picture assigned incorrect semantic meanings to the distorted pictures.


(a) Winnie the Pooh

(b) Wall Clock

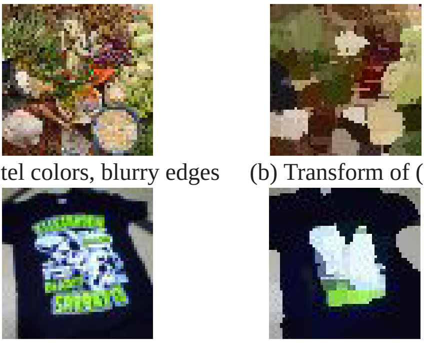
Figure 7: Examples of correct semantic meanings assigned by participants in Group 3. Some users were able to recognize Winnie the Pooh and the wall clock, even though they were not exposed to the original picture.

We found that 90% of the participants chose pictures whose subjects were things they saw regularly in their daily life, such as their pets, cars, and (more surprisingly) coffee mugs. Finally, many participants in Group 2 indicated that they considered Fata Morgana more as a game than as an authentication system in that it was fun and quite enjoyable to use.

6. DISCUSSION

While the usability tests presented above appear promising, Fata Morgana poses a number of interesting questions, which warrant further investigation. As discussed in Section 4, Fata Morgana does not mandate a given image processing filter: our choice of an oil

[6] Many other incorrect guesses, e.g., mistaking a duck for a dog, or correct guesses, e.g., a colorful T-shirt, are not represented here.



(a) Pastel colors, blurry edges (b) Transform of (a)

(c) Bright colors, sharp edges (d) Transform of (c)

Figure 8: Comparison of the filtering effects between different original images. Filtering parameters are held constant over both transformations. The transform (b) of an image (a) with pastel colors and slightly blurry shapes, conveys little information about the original picture. On the other hand, the transform (d) of an image (c) with bright colors and sharp edges, maintains most of the information of the original image.

painting filter has been driven mostly by heuristic considerations. Our usability and low-fidelity tests indicate that our current design choice is a viable one, but it is not yet clear it is the best one possible.

Further experimentation is also needed to better evidence the resilience of the scheme to some types of attacks; the low fidelity test we conducted to determine optimal parameter selection, while highly encouraging, needs to be expanded to provide stronger statistical evidence that attackers are not easily able to "revert" a distorted image back to its original meaning.

While investigating how best to tune our filter during the course of our prototype design and implementation, we discovered that finding an optimal parameter set point for our lossy filter depends on the picture to be transformed. For instance, holding filtering parameters constants, distorting an image with bright colors and clear shapes preserves more information about the original image than distorting an image with blurrier shapes and less contrast.

Figure 8 shows two original-distorted image pairs, obtained using identical filtering parameters for both transforms. In Figure 8(a), the original image consists of pastel colors, and slightly blurry shapes. The resulting distorted image, Figure 8(b) does not convey much information from the original picture. Conversely, distorting an original image with bright colors and sharp edges, as in Figure 8(c) results in a picture very similar to the original image, as shown by Figure 8(d).

One can argue that neither Figure 8(a) or Figure 8(c) is an adequate picture for Fata Morgana: (a) results in a distorted image that is not memorable, while (c) results in a distorted image that reveals too much information about the original picture. In the case of (a), a user can notice the problem and discard the picture during the training phase, so that we do not expect this problem to be a significant issue. The case illustrated in (c) is a little thornier in that it may be difficult to require users to voluntarily discard a portfolio image that is easy to remember.

Thus, it would be desirable to have a technical countermeasure to detect and adjust to such corner cases. For instance, if we had a metric that allows us to objectively measure the distortion level of a given picture, the filtering parameters could be adjusted for each original picture to obtain the maximum distortion level that still results in a memorable distorted picture. Finding such an objective "distortion metric" is an open problem.

Optimal parameter configuration does not only concern filter parameters, but also authentication parameters. The near-perfect recall rate after four weeks tends to indicate that the choice of $p = 3$ portfolio pictures, $n = 27$ challenge set pictures may have been a bit too conservative. One could want to increase either p or n to strengthen the resilience of the authentication scheme against brute force attacks. Further study is necessary to better characterize the relationship between n, p, and success rates.

Another point we think warrants further investigation, is that of authentication frequency. If the user only has to authenticate themselves when they turn on their cell phone, then a lost or stolen cell phone essentially "delegates" all the rights of the original owner to the person who finds/steals it, as long as that new owner ensures the phone is never turned off. A simple countermeasure against such a delegation problem is to periodically require the user to authenticate. The authentication frequency may be application-dependent. For instance, if the cell phone is to be used for business transactions, e.g., stock trading, access control primitives should be in place to ensure authentication is requested before money changes hands.

Finally, while Fata Morgana could be used for each single authentication instance, the 18 seconds taken on average to authenticate may be too high in some specific scenarios. Using mobile phones as railway passes, as can be seen in Japan, probably requires to authenticate in less than a second. To address such cases, one could envision a hierarchical authentication scheme, where Fata Morgana is at the top of the hierarchy, and is used for all essential services. For services where, from the user's perspective, strong security is less of a stringent requirement than having a fast authentication process, one could use a simple PIN (that could only be changed after successfully authenticating in Fata Morgana), or even no authentication at all, provided a successful authentication was performed within a reasonable timeframe. Identifying the most stringent requirement(s) for each service, and accordingly choosing the best authentication method to use, far exceeds the scope of the present paper.

7. CONCLUSION

We investigated a cognitive process relying on mental trapdoors, and showed that it can be used to maintain the usability of traditional graphical password authentication schemes, while making them more resilient against social engineering or guessing attacks. We demonstrated the benefits of mental trapdoors by designing, implementing, and testing a proof-of-concept authentication authentication system for small mobile devices, Fata Morgana. We have made available to the public web-based and Windows versions of our prototype at `http://arima.okoze.net/fyeo/`.

The results we obtained through our usability studies demonstrate that users are extremely skilled at recognizing degraded versions of self-chosen images, even after long time spans (one month). This illustrates the power of mental trapdoors: legitimate users who have been exposed to the original picture can easily mentally revert a highly lossy (one-way) transform on the picture, even though the transform is not mathematically reversible and the transformed picture conveys limited information. Furthermore, our interviews with participants indicate that the authentication process is more enjoyable, and "game-like," compared to current alternatives.

While this paper focuses on addressing the critical need for strong user authentication on small portable devices, the results obtained with Fata Morgana encourage us to consider using mental trapdoors to a much wider range of applications, such as spyware-resilient authentication primitives relying on zero-knowledge proofs.

8. REFERENCES

[1] Phoney finance. The Economist. October 26, 2006. Available online at http://www.economist.com/finance/displaystory.cfm?story_id=8089667.

[2] R. Anderson. Why cryptosystems fail. In *ACM Conference on Computer and Communications Security*, pages 215–227, Fairfax, VA, November 1993.

[3] G. Blonder. United states patent, 1996. United States Patent 5559961.

[4] G. H. Bower, M. B. Karlin, and A. Dueck. Comprehension and memory for pictures. *Memory and Cognition*, 2:216–220, 1975.

[5] S. Brostoff and M. Sasse. Are passfaces more usable than passwords? A field trial investigation. In *Proceedings of HCI 2000*, pages 405–424, Sunderland, UK, September 2000.

[6] M. Burton, S. Wilson, M. Cowan, and V. Bruce. Face recognition in poor quality video: Evidence from security surveillance. *Psychological Science*, 10:243–248, 1999.

[7] Sony Corporation. Overview of FeliCa. http://www.sony.net/Products/felica/abt/dvs.html.

[8] L. Cranor and S. Garfinkel, editors. *Security and Usability: Designing Secure Systems That People Can Use*. O'Reilly Media, Sebastopol, CA, 2005.

[9] Rachna Dhamija and Adrian Perrig. Déjà vu: A user study, using images for authentication. In *Proceedings of the 9th USENIX Security Symposium*, August 2000.

[10] Rachna Dhamija and J. D. Tygar. The battle against phishing: Dynamic security skins. In *SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security*, New York, NY, USA, 2005. ACM Press.

[11] A. Goldstein and J. E. Chance. Visual recognition memory for complex configurations. *Perception and Psychophysics*, 9:237–241, 1970.

[12] P. Golle and D. Wagner. Cryptanalysis of a cognitive authentication scheme. In *Proc. of the 2007 IEEE Symposium on Security and Privacy*, Oakland, CA, 2007. To appear.

[13] R. L. Gregory. *The Intelligent Eye*. 1970.

[14] R. Haber. How we remember what we see. *Scientific American*, 222(5):104–112, May 1970.

[15] A. Harada, T. Isarida, T. Mizuno, and M. Nishigaki. A user authentication system using schema of visual memory. In *Proceedings of BioADIT'06*, pages 338–345, Osaka, Japan, January 2006.

[16] Z. Henderson, V. Bruce, and M. Burton. Matching the faces of robbers captured on video. *Applied Cognitive Psychology*, 15:445–464, 2001.

[17] Ian Jermyn, Alain Mayer, Fabian Monrose, Michael K. Reiter, and Aviel D. Rubin. The design and analysis of graphical passwords. In *Proceedings of the 8th USENIX Security Symposium*, August 1999.

[18] H. Kinjo and J. G. Snodgrass. Does the generation effect occur for pictures? *The American journal of psychology*, 6:156–163, 2000.

[19] T. Matsumoto, H. Matsumoto, K. Yamada, and S. Hoshino. Impact of artificial gummy fingers on fingerprint systems. In *Proceedings of SPIE: Optical Security and Counterfeit Deterrence Techniques IV, 2002*, volume 4677, pages 275–289, San Jose, CA, January 2002.

[20] F. Monrose, D. Davis, and M. Reiter. On user choice to graphical password schemes. In *Proceedings of the 13th USENIX Security Symposium*, pages 151–164, San Diego, CA, August 2004.

[21] V. Roth, K. Fischer, and R. Freidinger. A PIN entry method resilient against shoulder surfing. In *Proceedings of ACM CCS'04*, Washington, DC, October 2004.

[22] J.C. Birget S. Wiedenbeck, J. Waters and N. Memon A. Brodskiy. Authentication using graphical passwords: Basic results. In *Human-Computer Interaction International*, Las Vegas, NJ, July 2005.

[23] RN Shepard. Recognition memory for words, sentences and pictures. *Journal of Verbal Learning and Verbal Behavior*, 113(1):95–121, 1967.

[24] 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.

[25] A. Stubblefield and D. Simon. Inkblot authentication. Technical Report MSR-TR-2004-85, August 2004.

[26] Julie Thorpe and Paul van Oorschot. Graphical dictionaries and the memorable space of graphical passwords. In *Proceedings of the 13th USENIX Security Symposium*, August 2004.

[27] Julie Thorpe and Paul van Oorschot. Towards secure design choices for implementing graphical passwords. In *Proceedings of the 20th Annual Computer Security Applications Conference (ACSAC)*, December 2004.

[28] Julie Thorpe and Paul van Oorschot. Human-seeded attacks and exploiting hot-spots in graphical passwords. In *Proceedings of USENIX Security 2007*, August 2007.

[29] D. Weinshall. Cognitive authentication schemes safe against spyware. In *Proceedings of the IEEE Symposium on Security and Privacy*, May 2006.

[30] Susan Wiedenbeck, Jim Waters, Jean-Camille Birget, Alex Brodskiy, and Nasir Memon. Authentication using graphical passwords: effects of tolerance and image choice. In *SOUPS '05: Proceedings of the 2005 symposium on Usable privacy and security*, pages 1–12, New York, NY, USA, 2005. ACM Press.

APPENDIX

A. OIL PAINTING FILTER

The oil painting filter is a lossy image processing filter Φ which produces an "oil-painted" image Q from an original picture P. Because the filter is lossy, the transform $\Phi: P \rightarrow Q$ is not mathematically reversible. That is, it is impossible to recreate P from Q. Additionally, there may exist P_1, P_2 such that $\Phi(P_1) = \Phi(P_2)$ while $P_1 \neq P_2$.

The oil painting filter processes the picture P by first assigning a set of Cartesian coordinates (X,Y) to each pixel in P. Then, for each pixel with coordinates (X,Y), using a given brush size β ($\beta \in \mathbb{N}$ is a configurable parameter), the filter extracts a range of values (x,y) as follows:

$$\begin{array}{rcl} X-\beta \le & x & \le X+\beta \\ Y-\beta \le & y & \le Y+\beta \end{array}$$

Then, for each of the $4\beta^2$ pixels (x,y) under consideration, the filter the filter classifies the RGB value of the pixel located at (x,y) in one of ν "bins", where $\nu \in \mathbb{N}$ ($\nu > 0$) is a second configurable filter parameter. For an RGB value $c(x,y)$, the bin $i(x,y)$ is given by

$$\frac{256(i(x,y)-1)}{\nu} \le c(x,y) \le \frac{256i(x,y)}{\nu} \quad (i(x,y) = 1,2,\ldots,\nu)\ ,$$

The filter then counts the number of pixels (x,y) present in each bin over the whole range of values (x,y) under consideration,

$$\begin{aligned} j(i) = \quad & \{(x,y) : i(x,y) = i\} \\ & (i \in \{1,\ldots,\nu\},\ X-\beta \leq x \leq X+\beta,\ Y-\beta \leq y \leq Y+\beta)\ , \end{aligned}$$

determines which bin j has the highest frequency of occurrence over the whole range,

$$j = \max_{1 \leq i \leq \nu} \{j(i)\}\ ,$$

and finally assigns $256(j-0.5)/\nu$ as the RGB value of the pixel located at (X,Y).