On User Choice in Graphical Password Schemes

Darren Davis Fabian Monrose
Johns Hopkins University
{ddavis,fabian}@cs.jhu.edu

Michael K. Reiter
Carnegie Mellon University
reiter@cmu.edu

Abstract

Graphical password schemes have been proposed as an alternative to text passwords in applications that support graphics and mouse or stylus entry. In this paper we detail what is, to our knowledge, the largest published empirical evaluation of the effects of user choice on the security of graphical password schemes. We show that permitting user selection of passwords in two graphical password schemes, one based directly on an existing commercial product, can yield passwords with entropy far below the theoretical optimum and, in some cases, that are highly correlated with the race or gender of the user. For one scheme, this effect is so dramatic so as to render the scheme insecure. A conclusion of our work is that graphical password schemes of the type we study may generally require a different posture toward password selection than text passwords, where selection by the user remains the norm today.

1 Introduction

The ubiquity of graphical interfaces for applications, and input devices such as the mouse, stylus and touch-screen that permit other than typed input, has enabled the emergence of graphical user authentication techniques (e.g., [2, 8, 4, 24, 7, 30]). Graphical authentication techniques are particularly useful when such devices do not permit typewritten input. In addition, they offer the possibility of providing a form of authentication that is strictly stronger than text passwords. History has shown that the distribution of text passwords chosen by human users has entropy far lower than possible [22, 5, 9, 32], and this has remained a significant weakness of user authentication for over thirty years. Given the fact that pictures are generally more easily remembered than words [23, 14], it is conceivable that humans would be able to remember stronger passwords of a graphical nature.

In this paper we study a particular facet of graphical password schemes, namely the strength of graphical passwords chosen by users. We note that not all graphical password schemes prescribe user chosen passwords (e.g., [24]), though most do (e.g., [2, 8, 3, 4, 7]). However, all of these schemes can be implemented using either system-chosen or user-chosen passwords, just as text passwords can be user-chosen or system-chosen. As with text passwords, there is potentially a tradeoff in graphical passwords between security, which benefits by the system choosing the passwords, and usability and memorability, which benefit by permitting the user to choose the password.

Our evaluation here focuses on one end of this spectrum, namely user chosen graphical passwords. The graphical password schemes we evaluate are a scheme we call "Face" that is intentionally very closely modeled after the commercial Passfaces$^{\text{TM}}$ scheme [3, 24] and one of our own invention (to our knowledge) that we call the "Story" scheme. In the Face scheme, the password is a collection of k faces, each chosen from a distinct set of $n > 1$ faces, yielding n^k possible choices. In the Story scheme, a password is a sequence of k images selected by the user to make a "story", from a single set of $n > k$ images each drawn from a distinct category of image types (cars, landscapes, etc.); this yields $n!/(n-k)!$ choices. Obviously, the password spaces yielded by these schemes is exhaustively searchable by a computer for reasonable values of k and n (we use $k = 4$ and $n = 9$), and so it relies on the authentication server refusing to permit authentication to proceed after sufficiently many incorrect authentication attempts on an account. Nevertheless, an argument given to justify the presumed security of graphical passwords over text passwords in such environments is the lack of a predefined "dictionary" of "likely" choices, as an English dictionary provides for English text passwords, for example (c.f., [8, Section 3.3.3]).

For our study we utilize a dataset we collected during the fall semester of 2003, of graphical password usage by three separate computer engineering and computer science classes at two different universities, yielding a total of 154 subjects. Students used graphical passwords (from one of the two schemes above) to access their grades, homework, homework solutions, course reading materials, etc., in a manner that we describe in Section 3.2. At the end of the semester, we asked students to complete an exit survey in which they described why they picked the faces they did (for Face) or their chosen stories (for Story) and some demographic information about themselves.

Using this dataset, in this paper we evaluate the Face and Story schemes to estimate the ability of an attacker to guess user-chosen passwords, possibly given knowledge of demographic information about the user. As we will show, our analysis suggests that the faces chosen by users in the Face scheme is highly affected by the race of the user, and that the gender and attractiveness of the faces also bias password choice. As to the latter, both male and female users select female faces far more often than male faces, and then select attractive ones more often than not. In the case of male users, we found this bias so severe that we do not believe it possible to make this scheme secure against an online attack by merely limiting the number of incorrect password guesses permitted. We also quantify the security of the passwords chosen in the Story scheme, which still demonstrates bias though less so, and make recommendations as to the number of incorrect password attempts that can be permitted in this scheme before it becomes insecure. Finally, we benchmark the memorability of Story passwords against those of the Face scheme, and identify a factor of the Story scheme that most likely contributes to its relative security but also impinges on its memorability.

On the whole, we believe that this study brings into question the argument that user-chosen graphical passwords of the type we consider here are likely to offer additional security over text passwords, unless users are somehow trained to choose better passwords, as they must be with text passwords today. Another alternative is to utilize only system-chosen passwords, though we might expect this would sacrifice some degree of memorability; we intend to evaluate this end of the spectrum in future work.

The rest of this paper is structured as follows. We describe related work in Section 2. In Section 3 we describe in more detail the graphical password schemes that we evaluate, and discuss our data sources and experimental setup. In Section 4 we introduce our chosen security measures, and present our results for them. In Section 5 we discuss issues and findings pertinent to the memorability of the two schemes. Finally, we conclude in Section 6.

2 Related Work

This work, and in particular our investigation of the Face scheme, was motivated in part by scientific literature in psychology and perception. Two results documented in the psychological literature that motivated our study are:

- Studies show that people tend to agree about the attractiveness of both adults and children, even across cultures. (Interested readers are referred to [10] for a comprehensive literature review on attractiveness.) In other words, the adage that "beauty is in the eye of the beholder," which suggests that each individual has a different notion of what is attractive, is largely false. For graphical password schemes like Face, this raises the question of what influence general perceptions of beauty (e.g, facial symmetry, youthfulness, averageness) [1, 6] might have on an individual's graphical password choices. In particular, given these a priori perceptions, are users more inclined to chose the most attractive images when constructing their passwords?
- Studies show that individuals are better able to recognize faces of people from their own race than faces of people from other races [31, 20, 11, 29]. The most straightforward account of the own-race effect is that people tend to have more exposure to members of their own racial group relative to other-race contact [31]. As such, they are better able to recognize intra-racial distinctive characteristics which leads to better recall. This so-called "race-effect" [13, 15] raises the question of whether users would favor members of their own race when selecting images to construct their passwords.

To the best of our knowledge, there has been no prior study structured to quantify the influence of the various factors that we evaluate here, including those above, on user *choice* of graphical passwords, particularly with respect to security. However, prior reports on graphical passwords have suggested the possibility of bias, or anecdotally noted apparent bias, in the selection or recognition of passwords. For example, a document [24] published by the corporation that markets Passfaces^{TM} makes reference to the race-effect, though stops short of indicating any effect it might have on password choice. In a study of twenty users of a graphical password system much like the Story scheme, except in which the password is a set of images as opposed to a sequence, several users reported that they did *not* select photographs of people because they did not feel they could relate personally to the image [4]. The same study also observed two instances in which users selected photographs of people of the same race as themselves, leading to a conjecture that this could play a role in password selection.

The Face scheme we consider here, and minor variants, have been the topic of several user studies focused on evaluating memorability (e.g., [34, 27, 28, 3]). These studies generally support the hypothesis that the Face scheme and variants thereof offer better memorability than text passwords. For instance, in [3], the authors report results of a three month trial investigation with 34 students that shows that fewer login errors were made when using Passfaces^{TM} (compared to textual passwords), even given significant periods of inactivity between logins.

Other studies, e.g., [34, 4], have explored memorability of other types of graphical passwords. We emphasize, however, that memorability is a secondary consideration for our purposes. Our primary goal is to quantify the effect of user choice on the *security* of passwords chosen.

3 Graphical Password Schemes

As mentioned earlier, our evaluation is based on two graphical schemes. In the Face scheme, the password is a collection of k faces, each selected from a distinct set of $n > 1$ faces. Each of the n faces are chosen uniformly at random from a set of faces classified as belonging to either a "typical" Asian, black or white male or female, or an Asian, black or white male or female model. This categorization is further discussed in Section 3.1. For our evaluation we choose $k = 4$ and $n = 9$. So, while choosing her password, the user is shown four successive 3×3 grids containing randomly chosen images (see Figure 1, for example), and for each, she selects one image from that grid as an element of her password. Images are unique and do not appear more than once for a given user. During the authentication phase, the same sets of images are shown to the user, but with the images randomly permuted.

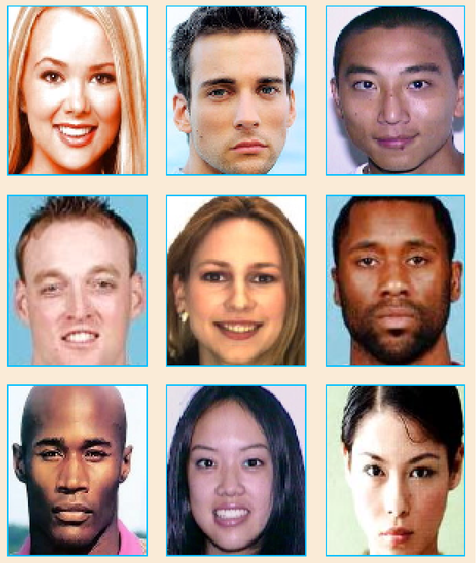

Figure 1: In the Face scheme, a user's password is a sequence of k faces, each chosen from a distinct set of $n > 1$ faces like the one above. Here, $n = 9$, and images are placed randomly in a 3×3 grid.

In the Story scheme, a password is a sequence of k unique images selected by the user to make a "story", from a single set of $n > k$ images, each derived from a distinct category of image types. The images are drawn from categories that depict everyday objects, food, automobiles, animals, children, sports, scenic locations, and male and female models. A sample set of images for the story scheme is shown in Figure 2.

3.1 Images

As indicated above, the images in each scheme were classified into non-overlapping categories. In Face, there were twelve categories: typical Asian males,

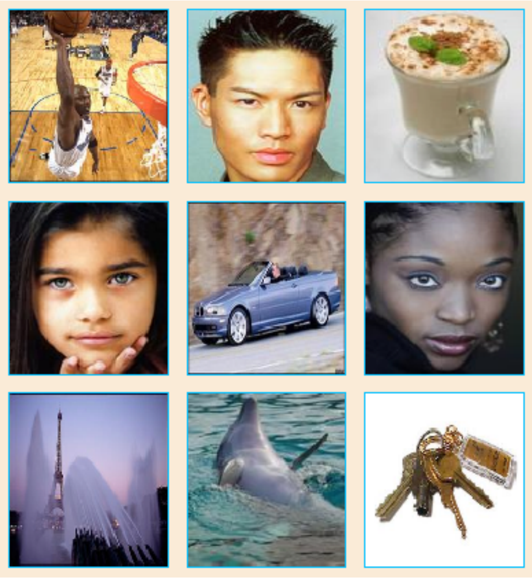

Figure 2: In the Story scheme, a user's password is sequence of k unique images selected from one set of n images, shown above, to depict a "story". Here, $n = 9$, and images are placed randomly in a 3×3 grid.

typical Asian females, typical black males, typical black females, typical white males, typical white females, Asian male models, Asian female models, black male models, black female models, white male models and white female models. In the Story scheme, there were nine categories: animals, cars, women, food, children, men, objects, nature, and sports.

The images used for each category were carefully selected from a number of sources. "Typical male" and "typical female" subjects include faces selected from (*i*) the Asian face database [26] which contains color frontal face images of 103 people and (*ii*) the AR Face database [17] which contains well over 4000 color images corresponding to 126 people. For the AR database we used images in angle 2 only, i.e, frontal images in the *smile* position. These databases were collected under controlled conditions and are made public primarily for use in evaluating face recognition technologies. For the most part, the subjects in these databases are students, and we believe provide a good representative population for our study. Additional images for typical male subjects were derived from a random sampling of images from the Sports Illustrated$^{\text{TM}}$NBA gallery. Images of "female models" were gathered from a myriad of pageant sites including Miss USA$^{\text{TM}}$, Miss Universe$^{\text{TM}}$, Miss NY Chinese, and fashion modeling sites. Images of "male models" were gathered from various online modeling sources including `FordModels.com` and `StormModels.com`.

For the Story scheme, the "men" and "women" categories were the same as the male and female models in our Face experiment. All other images were chosen from `PicturesOf.NET` and span the previously mentioned categories.

To lessen the effect that an image's intensity, hue, and background color may have on influencing a user choice, we used the `ImageMagick` library (see `www.imagemagick.org`) to set image backgrounds to a light pastel color at reduced intensity. Additionally, images with bright or distracting backgrounds, or of low quality, were deleted. All remaining images were resized to have similar aspect ratios. Of course, it is always possible that differences in such secondary factors influenced the results of our experiment, though we went to significant effort to avoid this and have found little to support a hypothesis of such influence.

3.2 Experiment

For our empirical evaluation we analyze observations collected during the fall semester (roughly the four month period of late-August through early-December) of 2003, of graphical password usage by three separate computer engineering and computer science classes at two different universities, yielding a total of 154 subjects. Each student was randomly assigned to one of the two graphical schemes. Each student then used the graphical password scheme for access to published content including his or her grades, homework, homework solutions, course reading materials, etc., via standard Java enabled browsers. Our system was designed so that instructors would not post documents on the login server, but rather that this server was merely used to encrypt and decrypt documents for posting or retrieval elsewhere. As such, from a student's perspective, the login server provided the means to decrypt documents retrieved from their usual course web pages.

Since there was no requirement for users to change their passwords, most users kept one password for the entire semester. However, a total of 174 pass-

Population		Scheme	
Gender	Race	Face	Story
any	*any*	79	95
Male	*any*	55	77
Female	*any*	20	13
Male	Asian	24	27
Female	Asian	12	8
Male	Black	3	-
Female	Black	-	-
Male	Hispanic	-	2
Female	Hispanic	-	-
Male	White	27	48
Female	White	8	4

Table 1: Population breakdown (in passwords).

words were chosen during the semester, implying that a few users changed their password at least once. During the evaluation period there were a total of 2648 login attempts, of which 2271 (85.76%) were successful. Toward the end of the semester, students were asked to complete an exit survey in which they described why they picked the faces they did (for Face) or their chosen stories (for Story) and provide some demographic information about themselves. This information was used to validate some of our findings which we discuss shortly. Table 1 summarizes the demographic information for our users. A gender or race of *any* includes those for which the user did not specify their gender or race. Such users account for differences between the sum of numbers of passwords for individual populations and populations permitting a race or gender of *any*.

The students participating in this study did so voluntarily and with the knowledge they were participating in a study, as required by the Institutional Review Boards of the participating universities. However, they were not instructed as to the particular factors being studied and, in particular, that the passwords they selected were of primary interest. Nor were they informed of the questions they would be asked at the end of the study. As such, we do not believe that knowledge of our study influenced their password choices. In addition, since personal information such as their individual grades were protected using their passwords, we have reason to believe that they did not choose them intentionally to be easily guessable.

4 Security evaluation

Recall that in both the Face and Story schemes, images are grouped into non-overlapping categories. In our derivations below, we make the simplifying assumption that images in a category are equivalent, that is, the specific images in a category that are available do not significantly influence a user's choice in picking a specific category.

First we introduce some notation. An ℓ-element tuple x is denoted $x^{(\ell)}$. If $\mathcal{S}$ is either the Face or Story scheme, then the expression $x^{(\ell)} \leftarrow \mathcal{S}$ denotes the selection of an ℓ-tuple $x^{(\ell)}$ (a password or password prefix, consisting of ℓ image categories) according to $\mathcal{S}$, involving both user choices and random algorithm choices.

4.1 Password distribution

In this section we describe how we approximately compute $\Pr\left[p^{(k)} \leftarrow \mathcal{S}\right]$ for any $p^{(k)}$, i.e., the probability that the scheme yields the password $p^{(k)}$. This probability is taken with respect to both random choices by the password selection algorithm and user choices.

We compute this probability inductively as follows. Suppose $p^{(\ell+1)} = q^{(\ell)} r^{(1)}$. Then

$$\begin{aligned}
&\Pr\left[p^{(\ell+1)} \leftarrow \mathcal{S}\right] \\
&\quad = \Pr\left[q^{(\ell)} \leftarrow \mathcal{S}\right] \cdot \\
&\qquad \Pr\left[q^{(\ell)} r^{(1)} \leftarrow \mathcal{S} \mid q^{(\ell)} \leftarrow \mathcal{S}\right]
\end{aligned} \tag{1}$$

if $p^{(\ell+1)}$ is valid for $\mathcal{S}$ and zero otherwise, where $\Pr\left[q^{(0)} \leftarrow \mathcal{S}\right] \stackrel{\text{def}}{=} 1$. Here, $p^{(\ell+1)}$ is *valid* iff $\ell < k$ and, for the Story scheme, $p^{(\ell+1)}$ does not contain any category more than once. The second factor $\Pr\left[q^{(\ell)} r^{(1)} \leftarrow \mathcal{S} \mid q^{(\ell)} \leftarrow \mathcal{S}\right]$ should be understood to mean the probability that the user selects $r^{(1)}$ after having already selected $q^{(\ell)}$ according to scheme $\mathcal{S}$. If the dataset contains sufficiently many observations, then this can be approximated by

$$\Pr\left[q^{(\ell)} r^{(1)} \leftarrow \mathcal{S} \mid q^{(\ell)} \leftarrow \mathcal{S}\right] \approx \frac{\#\left[q^{(\ell)} r^{(1)} \leftarrow \mathcal{S}\right]}{\#\left[q^{(\ell)} \leftarrow \mathcal{S}\right]}, \tag{2}$$

i.e., using the maximum likelihood estimation, where $\#\left[x^{(\ell)} \leftarrow \mathcal{S}\right]$ denotes the number of occurrences of $x^{(\ell)} \leftarrow \mathcal{S}$ in our dataset, and where

$\#\left[x^{(0)} \leftarrow \mathcal{S}\right]$ is defined to be the number of passwords for scheme $\mathcal{S}$ in our dataset.

A necessary condition for the denominator of (2) to be nonzero for every possible $q^{(k-1)}$ is that the dataset contain N^{k-1} samples for scheme $\mathcal{S}$ where $N \geq n$ denotes the number of image categories for $\mathcal{S}$. ($N = 12$ in Face, and $N = 9$ in Story.) N^{k-1} is over 1700 in the Face scheme, for example. And, of course, to use (2) directly to perform a meaningful approximation, significantly more samples would be required. Thus, we introduce a simplifying, Markov assumption: a user's next decision is influenced only by her immediately prior decision(s) (e.g., see [16]). In other words, rather than condition on all of the previous choices made in a password ($q^{(\ell)}$), only the last few choices are taken into account. Let $\ldots x^{(\ell)} \leftarrow \mathcal{S}$ denote the selection of an ℓ'-tuple, $\ell' \geq \ell$, for which the most recent ℓ selections are $x^{(\ell)}$.

Assumption 4.1 *There exists a constant $\hat{\ell} \geq 0$ such that if $\ell \geq \hat{\ell}$ then*

$$\begin{aligned}&\Pr\left[q^{(\ell)}r^{(1)} \leftarrow \mathcal{S} \mid q^{(\ell)} \leftarrow \mathcal{S}\right]\\ &\quad\approx \Pr\left[\ldots s^{(\hat{\ell})}r^{(1)} \leftarrow \mathcal{S} \mid \ldots s^{(\hat{\ell})} \leftarrow \mathcal{S}\right] \qquad (3)\end{aligned}$$

where $s^{(\hat{\ell})}$ is the $\hat{\ell}$-length suffix of $q^{(\ell)}$. We denote probabilities under this assumption by $\Pr_{\hat{\ell}}[\cdot]$.

In other words, we assume that if $\ell \geq \hat{\ell}$, then the user's next selection $r^{(1)}$ is influenced only by her last $\hat{\ell}$ choices. This appears to be a reasonable assumption, which is anecdotally supported by certain survey answers, such as the following from a user of the Face scheme.

> "To start, I chose a face that stood out from the group, and then I picked the closest face that seemed to match."

While this user's intention may have been to choose a selection similar to the first image she selected, we conjecture that the most recent image she selected, being most freshly on her mind, influenced her next choice at least as much as the first one did. Assumption 4.1 also seems reasonable for the Story scheme on the whole, since users who selected passwords by choosing a story were presumably trying to continue a story based on what they previously selected.

Assumption 4.1 permits us to replace (2) by

$$\begin{aligned}&\Pr_{\hat{\ell}}\left[q^{(\ell)}r^{(1)} \leftarrow \mathcal{S} \mid q^{(\ell)} \leftarrow \mathcal{S}\right]\\ &\quad\approx \frac{\#\left[\ldots s^{(\hat{\ell})}r^{(1)} \leftarrow \mathcal{S}\right]}{\#\left[\ldots s^{(\hat{\ell})} \leftarrow \mathcal{S}\right]} \qquad (4)\end{aligned}$$

where $s^{(\hat{\ell})}$ is the $\hat{\ell}$-length suffix of $q^{(\ell)}$ and we define $\#\left[\ldots s^{(0)} \leftarrow \mathcal{S}\right]$ to be the total number of category choices (k times the number of passwords) in our dataset for scheme $\mathcal{S}$. Here, the necessary condition for the denominator of (4) to be nonzero for each $s^{(\hat{\ell})}$ is that the dataset for $\mathcal{S}$ contain $N^{\hat{\ell}}$ samples, e.g., in the Face scheme, twelve for $\hat{\ell} = 1$, and so on.

We further augment the above approach with smoothing in order to compensate for gaps in the data (c.f., [16]). Specifically, we replace (4) with

$$\begin{aligned}&\Pr_{\hat{\ell}}\left[q^{(\ell)}r^{(1)} \leftarrow \mathcal{S} \mid q^{(\ell)} \leftarrow \mathcal{S}\right]\\ &\quad\approx \frac{\#\left[\ldots s^{(\hat{\ell})}r^{(1)} \leftarrow \mathcal{S}\right] + \lambda_{\hat{\ell}} \cdot \Psi_{\hat{\ell}-1}}{\#\left[\ldots s^{(\hat{\ell})} \leftarrow \mathcal{S}\right] + \lambda_{\hat{\ell}}} \qquad (5)\end{aligned}$$

where $s^{(\hat{\ell})}$ is the $\hat{\ell}$-length suffix of $q^{(\ell)}$; $\lambda_{\hat{\ell}} > 0$ is a real-valued parameter; and where if $\hat{\ell} > 0$ then

$$\Psi_{\hat{\ell}-1} = \Pr_{\hat{\ell}-1}\left[q^{(\ell)}r^{(1)} \leftarrow \mathcal{S} \mid q^{(\ell)} \leftarrow \mathcal{S}\right]$$

and $\Psi_{\hat{\ell}-1} = 1/N$ otherwise. Note that as $\lambda_{\hat{\ell}}$ is reduced toward 0, (5) converges toward (4). And, as $\lambda_{\hat{\ell}}$ is increased, (5) converges toward $\Psi_{\hat{\ell}-1}$, i.e., a probability under Assumption 4.1 for $\hat{\ell} - 1$, a stronger assumption. So, with sufficient data, we can use a small $\lambda_{\hat{\ell}}$ and thus a weaker assumption. Otherwise, using a small $\lambda_{\hat{\ell}}$ risks relying too heavily on a small number of occurrences of $\ldots s^{(\hat{\ell})} \leftarrow \mathcal{S}$, and so we use a large $\lambda_{\hat{\ell}}$ and thus the stronger assumption.

4.2 Measures

We are primarily concerned with measuring the ability of an attacker to guess the password of a user. Given accurate values for $\Pr\left[p^{(k)} \leftarrow \mathcal{S}\right]$ for each $p^{(k)}$, a measure that indicates this ability is the "guessing entropy" [18] of passwords. Informally, guessing entropy measures the expected number of guesses an attacker with perfect knowledge of the probability distribution on passwords would need in order to guess a password chosen from that distribution. If we enumerate passwords ${p_1}^{(k)}$, ${p_2}^{(k)}$, ... in non-increasing order of $\Pr\left[{p_i}^{(k)} \leftarrow \mathcal{S}\right]$, then the guessing entropy is simply

$$\sum_{i>0} i \cdot \Pr\left[{p_i}^{(k)} \leftarrow \mathcal{S}\right] \qquad (6)$$

Guessing entropy is closely related to Shannon entropy, and relations between the two are known.[1] Since guessing entropy intuitively corresponds more closely to the attacker's task in which we are interested (guessing a password), we will mainly consider measures motivated by the guessing entropy.

The direct use of (6) to compute guessing entropy using the probabilities in (5) is problematic for two reasons. First, an attacker guessing passwords will be offered additional information when performing a guess, such as the set of available categories from which the next image can be chosen. For example, in Face, each image choice is taken from nine images that represent nine categories of images, chosen uniformly at random from the twelve categories. This additional information constrains the set of possible passwords, and the attacker would have this information when performing a guess in many scenarios. Second, we have found that the absolute probabilities yielded by (5) can be somewhat sensitive to the choice of $\lambda_{\hat{\ell}}$, which introduces uncertainty into calculations that utilize these probabilities numerically.

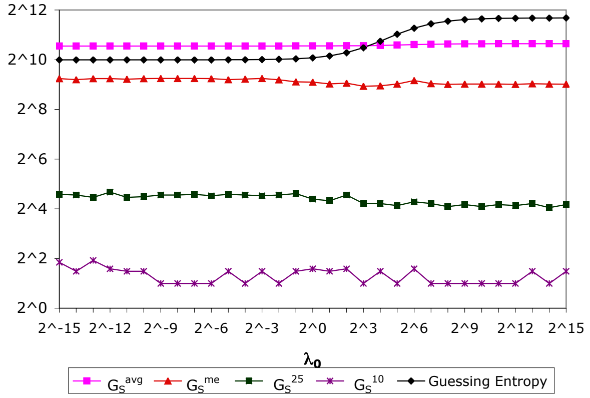

Figure 3: Measures versus λ_0 for Face

To account for the second of these issues, we use the probabilities computed with (5) only to determine an enumeration $\Pi = ({p_1}^{(k)}, {p_2}^{(k)}, \ldots)$ of passwords in non-increasing order of probability (as computed with (5)). This enumeration is far less sensitive to variations in $\lambda_{\hat{\ell}}$ than the numeric probabilities are, and so we believe this to be a more robust use of (5). We use this sequence to conduct tests with our dataset in which we randomly select a small set of "test" passwords from our dataset (20% of the dataset), and use the remainder of the data to compute the enumeration Π.

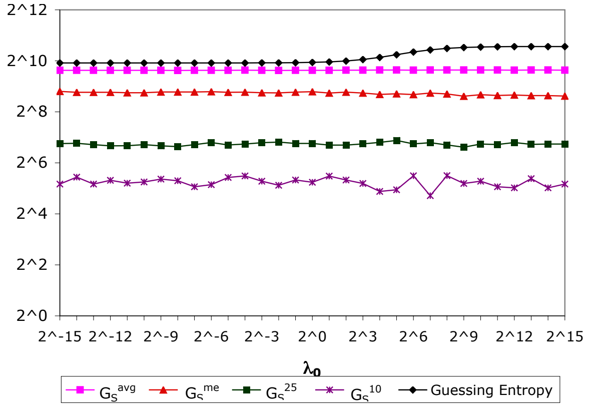

Figure 4: Measures versus λ_0 for Story

We then guess passwords in order of Π until each test password is guessed. To account for the first issue identified above, namely the set of available categories during password selection, we first filter from Π the passwords that would have been invalid given the available categories when the test password was chosen, and obviously do not guess them. By repeating this test with non-overlapping test sets of passwords, we obtain a number of guesses per test password. We use $G^{\mathrm{avg}}_{\mathcal{S}}$ to denote the average over all test passwords, and $G^{\mathrm{med}}_{\mathcal{S}}$ to denote the median over all test passwords. Finally, we use $G^{x}_{\mathcal{S}}$ for $0 < x \leq 100$ to denote the number of guesses sufficient to guess x percent of the test passwords. For example, if 25% of the test passwords could be guessed in 6 or fewer guesses, then $G^{25}_{\mathcal{S}} = 6$.

We emphasize that by computing our measures in this fashion, they are intrinsically conservative given our dataset. That is, an attacker who was given 80% of our dataset and challenged to guess the remaining 20% would do at least as well as our measures suggest.

4.3 Empirical results

To affirm our methodology of using $G^{\mathrm{avg}}_{\mathcal{S}}$, $G^{\mathrm{med}}_{\mathcal{S}}$, and $G^{x}_{\mathcal{S}}$ as mostly stable measures of password quality, we first plot these measures under various instances of Assumption 4.1, i.e., for various values of $\hat{\ell}$ and, for each, a range of values for $\lambda_{\hat{\ell}}$. For example, in the case of $\hat{\ell} = 0$, Figures 3 and 4 show measures $G_{\mathcal{S}}^{\text{avg}}$, $G_{\mathcal{S}}^{\text{med}}$, $G_{\mathcal{S}}^{25}$ and $G_{\mathcal{S}}^{10}$, as well as the guessing entropy as computed in (6), for various values of λ_0. Figure 3 is for the Face scheme, and Figures 4 is for the Story scheme.

The key point to notice is that each of $G_{\mathcal{S}}^{\text{avg}}$, $G_{\mathcal{S}}^{\text{med}}$, $G_{\mathcal{S}}^{25}$ and $G_{\mathcal{S}}^{10}$ is very stable as a function of λ_0, whereas guessing entropy varies more (particularly for Face). We highlight this fact to reiterate our reasons for adopting $G_{\mathcal{S}}^{\text{avg}}$, $G_{\mathcal{S}}^{\text{med}}$, and $G_{\mathcal{S}}^{x}$ as our measures of security, and to set aside concerns over whether particular choices of λ_0 have heavily influenced our results. Indeed, even for $\hat{\ell} = 1$ (with some degree of back-off to $\hat{\ell} = 0$ as prescribed by (5)), values of λ_0 and λ_1 do not greatly impact our measures. For example, Figures 5 and 6 show $G_{\mathcal{S}}^{\text{avg}}$ and $G_{\mathcal{S}}^{25}$ for Face. While these surfaces may suggest more variation, we draw the reader's attention to the small range on the vertical axis in Figure 5; in fact, the variation is between only 1361 and 1574. This is in contrast to guessing entropy as computed with (6), which varies between 252 and 3191 when λ_0 and λ_1 are varied (not shown). Similarly, while $G_{\mathcal{S}}^{25}$ varies between 24 and 72 (Figure 6), the analogous computation using (5) more directly—i.e., computing the smallest j such that $\sum_{i=1}^{j} \Pr\left[p_i{}^{(k)} \leftarrow \mathcal{S}\right] \geq .25$—varies between 27 and 1531. In the remainder of the paper, the numbers we report for $G_{\mathcal{S}}^{\text{avg}}$, $G_{\mathcal{S}}^{\text{med}}$, and $G_{\mathcal{S}}^{x}$ reflect values of λ_0 and λ_1 that simultaneously minimize these values to the extent possible.

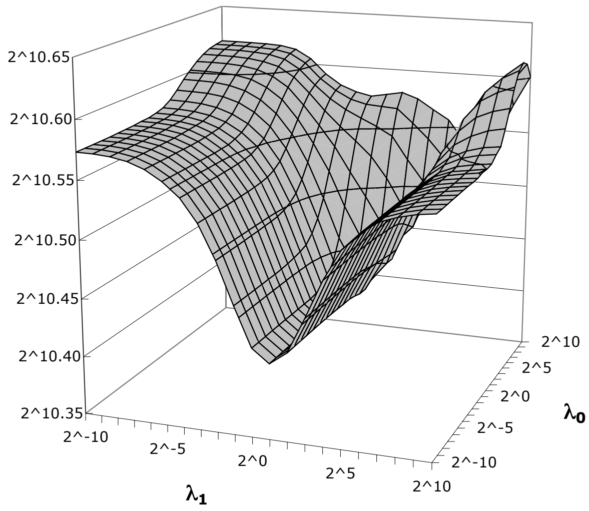

Figure 5: $G_{\mathcal{S}}^{\text{avg}}$ versus λ_0, λ_1 for Face

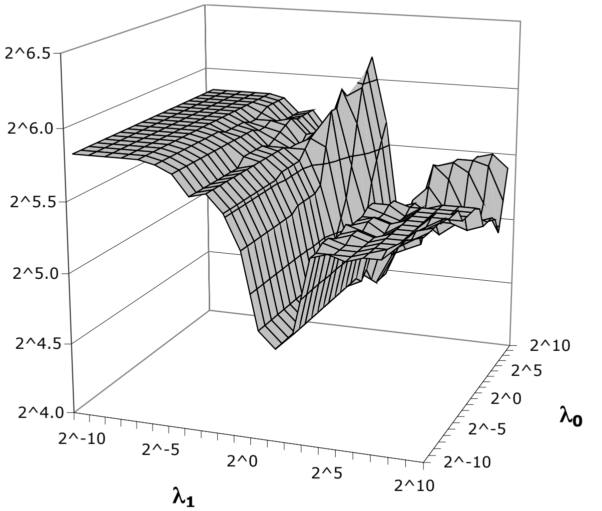

Figure 6: $G_{\mathcal{S}}^{25}$ versus λ_0, λ_1 for Face

Population	$G_{\mathcal{S}}^{\text{avg}}$	$G_{\mathcal{S}}^{\text{med}}$	$G_{\mathcal{S}}^{25}$	$G_{\mathcal{S}}^{10}$
Overall	790	428	112	35
Male	826	404	87	53
Female	989	723	125	98
White Male	844	394	146	76
Asian Male	877	589	155	20

Table 2: Results for Story, $\lambda_0 = 2^{-2}$

Tables 2 and 3 present results for the Story scheme and the Face scheme, respectively. Populations with less than ten passwords are excluded from these tables. These numbers were computed under Assumption 4.1 for $\hat{\ell} = 0$ in the case of Story and for $\hat{\ell} = 1$ in the case of Face. λ_0 and λ_1 were tuned as indicated in the table captions. These choices were dictated by our goal of minimizing the various measures we consider ($G_{\mathcal{S}}^{\text{avg}}$, $G_{\mathcal{S}}^{\text{med}}$, $G_{\mathcal{S}}^{25}$ and $G_{\mathcal{S}}^{10}$), though as already demonstrated, these values are generally not particularly sensitive to choices of λ_0 and λ_1.

The numbers in these tables should be considered in light of the number of available passwords. Story

Population	$G_{\mathcal{S}}^{\text{avg}}$	$G_{\mathcal{S}}^{\text{med}}$	$G_{\mathcal{S}}^{25}$	$G_{\mathcal{S}}^{10}$
Overall	1374	469	13	2
Male	1234	218	8	2
Female	2051	1454	255	12
Asian Male	1084	257	21	5.5
Asian Female	973	445	19	5.2
White Male	1260	81	8	1.6

Table 3: Results for Face, $\lambda_0 = 2^{-2}, \lambda_1 = 2^2$

has $9 \times 8 \times 7 \times 6 = 3024$ possible passwords, yielding a maximum possible guessing entropy of 1513. Face, on the other hand, has $9^4 = 6561$ possible passwords (for fixed sets of available images), for a maximum guessing entropy of 3281.

Our results show that for Face, if the user is known to be a male, then the worst 10% of passwords can be easily guessed on the first or second attempt. This observation is sufficiently surprising as to warrant restatement: An online dictionary attack of passwords will succeed in merely **two guesses** for 10% of male users. Similarly, if the user is Asian and his/her gender is known, then the worst 10% of passwords can be guessed within the first six tries.

It is interesting to note that $G_{\mathcal{S}}^{\mathrm{avg}}$ is always higher than $G_{\mathcal{S}}^{\mathrm{med}}$. This implies that for both schemes, there are several good passwords chosen that significantly increase the average number of guesses an attacker would need to perform, but do not affect the median. The most dramatic example of this is for white males using the Face scheme, where $G_{\mathcal{S}}^{\mathrm{avg}} = 1260$ whereas $G_{\mathcal{S}}^{\mathrm{med}} = 81$.

These results raise the question of what different populations tend to choose as their passwords. Insight into this for the Face scheme is shown in Tables 4 and 5, which characterize selections by gender and race, respectively. As can be seen in Table 4, both males and females chose females in Face significantly more often than males (over 68% for females and over 75% for males), and when males chose females, they almost always chose models (roughly 80% of the time). These observations are also widely supported by users' remarks in the exit survey, e.g.:

> "I chose the images of the ladies which appealed the most."

> "I simply picked the best lookin girl on each page."

> "In order to remember all the pictures for my login (after forgetting my 'password' 4 times in a row) I needed to pick pictures I could EASILY remember - kind of the same pitfalls when picking a lettered password. So I chose all pictures of beautiful women. The other option I would have chosen was handsome men, but the women are much more pleasing to look at :)"

> "Best looking person among the choices."

Moreover, there was also significant correlation among members of the same race. As shown in Table 5, Asian females and white females chose from within their race roughly 50% of the time; white males chose whites over 60% of the time, and black males chose blacks roughly 90% of the time (though the reader should be warned that there were only three black males in the study, thus this number requires greater validation). Again, a number of exit surveys confirmed this correlation, e.g.:

> "I picked her because she was female and Asian and being female and Asian, I thought I could remember that."

> "I started by deciding to choose faces of people in my own race ... specifically, people that looked at least a little like me. The hope was that knowing this general piece of information about all of the images in my password would make the individual faces easier to remember."

> "... Plus he is African-American like me."

Pop.	Female Model	Male Model	Typical Female	Typical Male
Female	40.0%	20.0%	28.8%	11.3%
Male	63.2%	10.0%	12.7%	14.0%

Table 4: Gender and attractiveness selection in Face.

Insight into what categories of images different genders and races chose in the Story scheme are shown in Tables 6 and 7. The most significant deviations between males and females (Table 6) is that females chose animals twice as often as males did, and males chose women twice as often as females did. Less pronounced differences are that males tended to select nature and sports images somewhat more than females did, while females tended to select food images more often. However, since these differences

Pop.	Asian	Black	White
Asian Female	52.1%	16.7%	31.3%
Asian Male	34.4%	21.9%	43.8%
Black Male	8.3%	91.7%	0.0%
White Female	18.8%	31.3%	50.0%
White Male	17.6%	20.4%	62.0%

Table 5: Race selection in Face.

were all within four percentage points, it is not clear how significant they are. Little emerges as definitive trends by race in the Story scheme (Table 7), particularly considering that the Hispanic data reflects only two users and so should be discounted.

5 Memorability evaluation

In this section we briefly evaluate the memorability of the schemes we considered. As described in Section 2, there have been many usability studies performed for various graphical password schemes, including for variants of the Face scheme. As such, our goal in this section is not to exhaustively evaluate memorability for Face, but rather to simply benchmark the memorability of the Story scheme against that of Face to provide a qualitative and relative comparison between the two.

Figure 7 shows the percentage of successful logins versus the amount of time since the password was initially established, and Figure 8 shows the percentage of successful logins versus the time since that user's last login attempt. Each figure includes one plot for Face and one plot for Story. A trend that emerges is that while memorability of both schemes is strong, Story passwords appear to be somewhat harder to remember than Face. We do not find this to be surprising, since previous studies have shown Face to have a high degree of memorability.

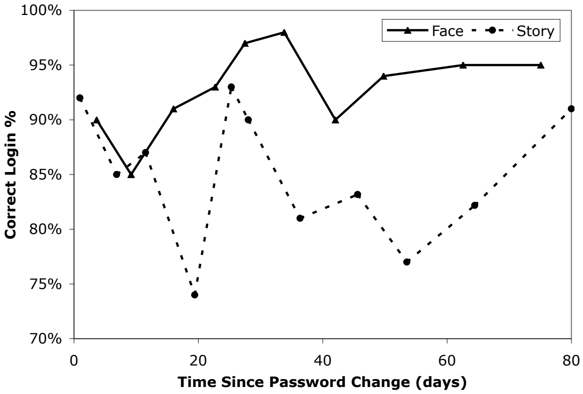

Figure 7: Memorability versus time since password change. Each data point represents the average of 100 login attempts.

One potential reason for users' relative difficulty in remembering their Story passwords is that apparently few of them actually chose stories, despite our suggestion to do so. Nearly 50% of Story users reported choosing no story whatsoever in their exit surveys. Rather, these users employed a variety of alternative strategies, such as picking four pleasing pictures and then trying to memorize the order in which they picked them. Not surprisingly, this contributed very significantly to incorrect password entries due to misordering their selections. For example, of the 236 incorrect password entries in Story, over 75% of them consisted of the correct images selected in an incorrect order. This is also supported anecdotally by several of the exit surveys:

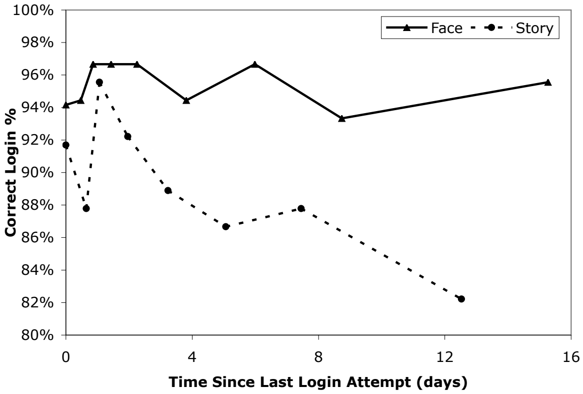

Figure 8: Memorability versus time since last login attempt. Each data point represents the average of 90 login attempts.

> "I had no problem remembering the four pictures, but I could not remember the original order."

> "No story, though having one may have helped to remember the order of the pictures better."

> "... but the third try I found a sequence that I could remember. fish-woman-girl-corn, I would screw up the fish and corn order 50% of the time, but I knew they were the pictures."

As such, it seems advisable in constructing graphical password schemes to avoid having users remember an ordering of images. For example, we expect that a selection of k images, each from a distinct set of n images (as in the Face scheme, though with image categories not necessarily of only persons), will generally be more memorable than an ordered selection of k images from one set. If a scheme does

Pop.	Animals	Cars	Women	Food	Children	Men	Objects	Nature	Sports
Female	20.8%	14.6%	6.3%	14.6%	8.3%	4.2%	12.5%	14.6%	4.2%
Male	10.4%	17.9%	13.6%	11.0%	6.8%	4.6%	11.0%	17.2%	7.5%

Table 6: Category selection by gender in Story

Pop.	Animals	Cars	Women	Food	Children	Men	Nature	Objects	Sports
Asian	10.7%	18.6%	11.4%	11.4%	8.6%	4.3%	17.1%	11.4%	6.4%
Hispanic	12.5%	12.5%	25.0%	12.5%	0.0%	12.5%	12.5%	12.5%	0.0%
White	12.5%	16.8%	13.0%	11.5%	6.3%	4.3%	16.8%	11.1%	7.7%

Table 7: Category selection by race in Story

rely on users remembering an ordering, then the importance of the story should be reiterated to users, since if the sequence of images has some semantic meaning then it is more likely that the password is memorable (assuming that the sequences are not too long [21]).

6 Conclusion

The graphical password schemes we considered in this study have the property that the space of passwords can be exhaustively searched in short order if an offline search is possible. So, any use of these schemes requires that guesses be mediated and confirmed by a trusted online system. In such scenarios, we believe that our study is the first to quantify factors relevant to the security of user-chosen graphical passwords. In particular, our study advises against the use of a Passfaces$^{\text{TM}}$-like system that permits user choice of the password, without some means to mitigate the dramatic effects of attraction and race that our study quantifies. As already demonstrated, for certain populations of users, no imposed limit on the number of incorrect password guesses would suffice to render the system adequately secure since, e.g., 10% of the passwords of males could have been guessed by merely two guesses.

Alternatives for mitigating this threat are to prohibit or limit user choice of passwords, to educate users on better approaches to select passwords, or to select images less prone to these types of biases. The first two are approaches initially attempted in the context of text passwords, and that have appeared in some graphical password schemes, as well. The Story scheme is one example of the third strategy (as is [4]), and our study indicates that password selection in this scheme is sufficiently free from bias to suggest that reasonable limits could be imposed on password guesses to render the scheme secure. For example, the worst 10% of passwords in the Story scheme for the most predictable population (Asian males) still required twenty guesses to break, suggesting a limit of five incorrect password guesses might be reasonable, provided that some user education is also performed.

The relative strength of the Story scheme must be balanced against what appears to be some difficulty of memorability for users who eschew the advice of using a story to guide their image selection. An alternative (besides better user education) is to permit unordered selection of images from a larger set (c.f., [4, 7]). However, we believe that further, more sizeable studies must be performed in order to confirm the usability and security of these approaches.

7 Acknowledgments

The authors would like to thank Joanne Houlahan for her support and for encouraging her students to use the graphical login server. We also extend our gratitude to all the students at Carnegie Mellon University and Johns Hopkins University who participated in this study.

Notes

[1] For a random variable X taking on values in $\mathcal{X}$, if $G(X)$ denotes its guessing entropy and $H(X)$ denotes its Shannon entropy, then it is known that $G(X) \geq 2^{H(X)-2} + 1$ [18] and that $H(X) \geq \frac{2 \log |\mathcal{X}|}{|\mathcal{X}|-1}(G(X) - 1)$ [19].

References

[1] T. Alley and M. Cunningham. Averaged faces are attractive, but very attractive faces are not average. In *Psychological Science*, 2, pages 123-125, 1991.

[2] G. E. Blonder. Graphical password. US Patent 5559961, Lucent Technologies, Inc., Murray Hill, NJ, August 30, 1995.

[3] S. Brostoff and M.. A. Sasse. Are Passfaces$^{\mathrm{TM}}$ more usable than passwords? A field trial investigation. In *Proceedings of Human Computer Interaction*, pages 405–424, 2000.

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

[5] D. Feldmeier and P. Karn. UNIX password security—Ten years later. In *Advances in Cryptology—CRYPTO '89* (Lecture Notes in Computer Science 435), 1990.

[6] A. Feingold. Good-looking people are not what we think. In *Psychological Bulletin*, 111, pages 304-341, 1992.

[7] W. Jansen, S. Gavrila, V. Korolev, R. Ayers, and R. Swanstrom. Picture password: A visual login technique for mobile devices. NISTIR 7030, Computer Security Division, Information Technology Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, July 2003. Available at `http://csrc.nist.gov/publications/nistir/nistir-7030.pdf`.

[8] I. Jermyn, A. Mayer, F. Monrose, M. Reiter and A. Rubin. The design and analysis of graphical passwords. In *Proceedings of the 8th USENIX Security Symposium*, August 1999.

[9] D. Klein. Foiling the cracker: A survey of, and improvements to, password security. In *Proceedings of the* 2*nd* *USENIX Security Workshop*, pages 5–14, August 1990.

[10] J. Langlois, L. Kalakanis, A. Rubenstein, A. Larson, M. Hallam, and M. Smoot. Maxims and myths of beauty: A meta-analytic and theoretical review. In *Psychological Bulletin* 126:390–423, 2000.

[11] D. Levin. Race as a visual feature: using visual search and perceptual discrimination tasks to understand face categories and the cross race recognition deficit. *Quarterly Journal of Experimental Psychology:General*, 129 (4), 559-574.

[12] D. Lindsay, P. Jack, and M. Chrisitan. Other-race face perception. *Journal of Applied Psychology* 76:587–589, 1991.

[13] T. Luce. Blacks, whites and yellows: They all look alike to me. *Psychology Today* 8:105–108, 1974.

[14] S. Madigan. Picture memory. In *Imagery, Memory, and Cognition*, pages 65–86, Lawrence Erlbaum Associates, 1983.

[15] R. S. Malpass. They all look alike to me. In *The Undaunted Psychologist*, pages 74-88, McGraw-Hill, 1992.

[16] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*, Chapter 6, MIT Press, May 1999.

[17] A. M. Martinez and R. Benavente. The AR Face Database. Technical Report number 24, June, 1998.

[18] J. L. Massey. Guessing and entropy. In *Proceedings of the 1994 IEEE International Symposium on Information Theory*, 1994.

[19] R. J. McEliece and Z. Yu. An inequality on entropy. In *Proceedings of the 1995 IEEE International Symposium on Information Theory*, 1995.

[20] C. Meissner, J. Brigham. Thirty years of investigation the own-race advantage in memory for faces: A meta-analytic review. *Psychology, Public Policy & Law*, 7, pages 3-35, 2001.

[21] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review* 63:81–97, 1956.

[22] R. Morris and K. Thompson. Password security: A case history. *Communications of the ACM* 22(11):594–597, November 1979.

[23] D. L. Nelson, U. S. Reed, and J. R. Walling. Picture superiority effect. *Journal of Experimental Psychology: Human Learning and Memory*, 3:485–497, 1977.

[24] *The Science Behind Passfaces.* Revision 2, Real User Corporation, September 2001. Available at `http://www.realuser.com/published/ScienceBehindPassfaces.pdf`.

[25] *Strategies for using Passfaces$^{\mathrm{TM}}$ for Windows.* Real User Corporation, 2002. Available at `http://www.realuser.com/published/PassfacesforWindowsStrategies.pdf`.

[26] *Asian Face Image Database PF01.* Pohang University of Science and Technology, Korea, 2001.

[27] T. Valentine. An evaluation of the Passfaces$^{\mathrm{TM}}$ personal authentication system. Technical Report, Goldsmiths College University of London, 1998.

[28] T. Valentine. Memory for Passfaces$^{\mathrm{TM}}$ after a long delay. Technical Report, Goldsmiths College University of London, 1999.

[29] T. Valentine and M. Endo. Towards an exemplar model of face processing: The effects of race and distinctiveness. *Quarterly Journal of Experimental Psychology* 44, 671-703.

[30] Visual Key – Technology. Available at `http://www.viskey.com/tech.html`.

[31] P. Walker and W. Tanaka. An encoding advantage for own-race versus other-race faces. In *Perception*, 23, pages 1117-1125, 2003.

[32] T. Wu. A real-world analysis of Kerberos password security. In *Proceedings of the 1999 ISOC Symposium on Network and Distributed System Security*, February 1999.

[33] M. Zviran and W. J. Haga. Cognitive passwords: The key to easy access and control. *Computers and Security* 9(8):723–736, 1990.

[34] M. Zviran and W. J. Haga. A comparison of password techniques for multilevel authentication mechanisms. *The Computer Journal* 36(3):227–237, 1993.