Is a picture really worth a thousand words? Exploring the feasibility of graphical authentication systems

Antonella De Angeli[a1], Lynne Coventry[a], Graham Johnson[a] and Karen Renaud[b]
[a] Advanced Technology & Research, NCR Financial Solutions Group Ltd

3 Fulton Road, Dundee, DD2 4SW

[b] Department of Computing Science - University of Glasgow

17 Lilybank Gardens, Glasgow G12 8RZ, UK

Corresponding author: Antonella De Angeli

Present address:

Centre for HCI Design, School of Informatics, University of Manchester

PO BOX 88, Manchester, M60 1QD

Fax: +44 161 306 1281 Tel. +44 (0)161- 306 3383

E-mail address: Antonella.de-angeli@manchester.ac.uk

Abstract
The weakness of knowledge-based authentication systems, such as passwords and Personal Identification Numbers (PINs), is well known, and reflects an uneasy compromise between security and human memory constraints. Research has been undertaken for some years now into the feasibility of graphical authentication mechanisms in the hope that these will provide a more secure and memorable alternative. The graphical approach substitutes the exact recall of alphanumeric codes with the recognition of previously learnt pictures, a skill at which humans are remarkably proficient. So far, little attention has been devoted to usability and initial research has failed to conclusively establish significant memory improvement. This paper reports two user studies comparing several implementations of the graphical approach with PINs. Results demonstrate that pictures can be a solution to some problems relating to traditional knowledge-based authentication but that they are not a simple panacea, since a poor design can eliminate the picture superiority effect in memory. The paper concludes by discussing the potential of the graphical approach and providing guidelines for developers contemplating using these mechanisms.

Keywords: User authentication; visual memory; usability; security

1. Introduction

User authentication is a problem for every system providing secure access to valuables, confidential information, or personalised services. Most systems make use of knowledge-based authentication mechanisms, such as *Personal Identification Numbers* (PINs) and passwords, because they are simple to administer, well understood by users and system administrators alike, and require no extra hardware or software (Renaud and De Angeli, 2004). Despite this, passwords and PINs have a number of well-known deficiencies reflecting a difficult compromise between security and memorability (Adams and Sasse, 1999; Besnard and Arief, 2004). Secure codes must be composed of a long, random selection of alphanumeric keys but unfortunately humans struggle to remember meaningless strings. Thus they choose simple and predictable words or numbers related to everyday life, and engage in insecure practices, such as writing passwords down or sharing them. The problem is so serious that the user is often referred to as the 'weakest link' in the security chain (Sasse et al., 2001).

Biometric techniques – those that make use of physiological or behavioural characteristics of an individual to confirm identity during authentication – may alleviate memory load, but they too need to resolve the security-usability balance for general usage (Coventry et al., 2003a, 2003b). Biometrics cause additional consumer concerns about privacy. Until biometrics become more robust, easy to use and ubiquitous, knowledge-based authentication will prevail and research into this mechanism is needed.

A number of graphical authentication systems have emerged, especially in the area of handheld devices, for which typewritten input is less common than pointing at the screen (Jansen et. al., 2003). The basic idea behind graphical authentication is that exact password *recall* is replaced by *recognition* of pictures. It is claimed that these

mechanisms are more secure, easier to use and more appealing to the general public than PINs and passwords (Dhamija and Perrig, 2000; Jermyn, et al., 2000; Weinshall and Kirkpatrick, 2004). Unfortunately, most proposals emphasise security and tend to over-estimate visual-memory capabilities, with usability being given scant attention.

This paper addresses the usability of graphical mechanisms based on two user studies carried out within the Visual Identification Protocol (VIP) project to assess the potential of graphical authentication for Automatic Teller Machines (ATMs). This is a highly constrained environment with strong usability and security issues. Consumers of all types need to 'walk up and use' the same machine engaging in a very brief goal-oriented and secure interaction. Owners of terminals cannot allow the ATM to be an easy target for fraud but equally cannot afford customer dissatisfaction through false rejection. The goal of the project was to find the best compromise between usability and security, providing fast, easy-to-use, and secure authentication mechanisms.

The paper is structured as follows: Section 2 focuses on graphical authentication, presenting the rationale behind it, a taxonomy of different implementation types and associated problems. Sections 3 and 4 report the experiments. Section 5 provides design guidelines to maximise the usability of graphical mechanisms, and section 6 concludes.

2. Graphical authentication

The interest in graphical authentication is driven by the assumption that pictures are *easier to remember* and *more secure* than words. The increase in security is associated to the difficulty in communicating or recording pictures, which should inhibit insecure practices (Weinshall and Kirkpatrick, 2004). The increase in memorability is predicted by the *picture superiority effect* demonstrated in several cognitive psychology studies (Nickerson, 1965; Standing, 1973; Standing et al., 1970; Shepard, 1967). According to this effect, humans have a vast, almost limitless, visual memory, and pictures tend to be remembered far better and for longer than words (Madigan 1983, Paivio, 1971, 1983; Paivio et al., 1968).

A framework to explain the effect is the dual-code theory, which ascribes the superior retention of pictures to their greater likelihood of inducing both an imaginal code and a verbal code (Paivio, 1971, 1983). An alternative is the suggestion that pictures induce a richer, more detailed representation in memory than words, and this makes them more distinctive at the time of retrieval (Nelson et al., 1977). Accordingly, pictures are represented in a rich sensory-perceptual code and have direct access to semantic knowledge, while words are mediated by lexical access and lead to semantic processes only if required by the encoding or retrieval task (Dewurst and Conway, 1994).

A range of different mechanisms have been proposed to exploit the power of pictures as a means for user authentication. We propose to cluster them into three categories: *Cognometrics*, *Locimetrics*, and *Drawmetrics*. The term *cognometric* is used by Real User Corporation to denote authentication systems based on the measurement of innate cognitive abilities of the human brain, such as face recognition (Real User Corporation, 2004). In this paper we propose to restrict its meaning to those systems based on visual recognition of target images embedded amongst a set of distractor

images. *Locimetric* systems refer to mechanisms requiring the identification of a target point within an individual image. *Drawmetric* systems lie at the borderline between biometrics and graphical mechanisms, as they require the user to reproduce a pre-drawn outline drawing.

2.1 Cognometrics

Example cognometrics are: Passfaces by Real User Corporation, Déjà Vu (Dhamija and Perrig, 2000), and VIP (De Angeli et al., 2002, 2003) presented in this paper. All follow the same paradigm—requiring identification of target images amongst distractors— but each uses different procedures and different visual stimuli.

Passfaces relies on face recognition. At enrolment, the user is given a set of faces which constitutes the code. At authentication each face is displayed surrounded by 8 distractors, and the user must click on the target face. The procedure is repeated for each face in the code. Passfaces has been available on the Internet for several years, and the company claims a user population of over 15,000 users. However, published research has failed to demonstrate a clear advantage of the solution when compared to traditional passwords (Brostoff and Sasse, 2000). Furthermore, issues related to choice predictability (should users be given the opportunity to select their own password) have been raised, as people tend to select faces of their own race and gender (Davis et al., 2004).

Déjà Vu (Dhamija and Perrig, 2000) is based on abstract images, a type of stimuli considered particularly secure because of the increased difficulty in communication. The user selects a code of 5 images from a visual data-base and then have to recognise them in a larger challenge set. A user evaluation has investigated the memorability of this scheme against photographic pictures, passwords and PINs (Dhamija and Perrig, 2000). Results show that creating passwords and PINs is faster than selecting visual codes, with photographs requiring the longest time. Images were less error prone than passwords and PINs after a one week interval, and users expressed a preference for photographic images. It has to be noted, anyway, that the results of this study are not substantiated by a solid experimental design (De Angeli et al., 2002).

2.2 Locimetric systems

Locimetric systems are based on the method of loci, an old and well-known mnemonic (Higbee, 1988). The learner mentally associates objects to be recalled with different and familiar locations, such as rooms in a building, or sites along an oft-travelled road. She mentally revisits the specific locations along her journey to retrieve objects. The initial proposal of a graphical authentication system exploited this idea, requiring the user to touch, in order, pre-set areas of an image (Blonder, 1996). Commercial locimetrics are Visual Key by SFR and V-go Password by Passlogix, which requires the user to simulate familiar actions on a graphical interface (such as ‘mixing a cocktail’) and uses the sequence of mouse movements as the authentication code.

The difference between the traditional method of loci and locimetrics is in the anchoring context: mental images versus 2-D digital pictures. The efficacy of such a transferral has proved difficult (Renaud and De Angeli, 2004), as people tend to forget

the exact position they chose previously, point at it inaccurately with the mouse, and are worryingly predictable in their selection.

2.3 Drawmetric systems

Drawmetrics require the user to draw a preset outline figure, either on top of an image or on a grid. A well-known system in this category is Draw-a-Secret, a PDA application, claimed to be much more secure than traditional passwords as it boasts a larger password space (Jermyn et al., 2000). A user evaluation revealed that people remembered drawings with stroke order as a match determinant less accurately than alphanumerical passwords (Goldberg et al., 2002). Users could recall all visual elements of the drawing as well as they could recall alphanumeric passwords, but they had difficulties redrawing their doodles accurately. Furthermore, another study found that users tended to draw symmetrical figures, which significantly reduces the 'dictionary' size and impacts the security of these mechanisms (Thorpe and van Oorshot, 2004).

2.4 Challenges in graphical authentication

Despite the growing interest in graphical authentication methods, their actual usability still needs to be proven, and initial results are discouraging (Brostoff and Sasse, 2000; De Angeli et al., 2002, 2003; Renaud and De Angeli, 2004; Goldberg et al., 2002). To the best of our knowledge, no studies so far have conclusively demonstrated a strong advantage of the graphical approach over passwords or PINs. On the contrary, a number of issues have been raised, including that of predictability (Davis et al. 2004; Renaud and De Angeli, 2004) showing that self-selected graphical codes have lower entropy than passwords.

There are many unanswered questions about the use of visual passwords, and more research is required to understand the role of visual memory in authentication mechanisms. This paper contributes to this understanding by reporting two studies of people's behaviour with, and attitudes towards, different implementations of VIP, a cognometric mechanism designed to simplify user authentication at the ATM interface.

3. Study 1

Three graphical mechanisms were designed and compared to the traditional PIN to investigate limits and potentials of the graphical approach. All of the graphical prototypes displayed detailed, colourful and meaningful photos of objects and scenes (Figure 1), as these are the typical stimuli used in classic research reporting the picture superiority effect (e.g., Nickerson, 1965; Shepard, 1967; Standing, 1973) and they are believed to be easier to remember (Paivio, 1971). In order to avoid security problems related to choice predictability, users were assigned a series of images to represent their visual code. At authentication they had to recognise these images from a wider challenge set of pictures.

_____

Insert Figure 1 about here

_____

3.1 Authentication mechanisms

Basic characteristics of each authentication mechanism are summarised in Table 1. *PIN* required the user to learn a sequence of 4 digits and to enter it on a touch-screen ATM-like keypad. *VIP1* was the pictorial equivalent of PIN (Figure 1). The user had to select a sequence of 4 pictures, which were displayed in the same position at each authentication attempt. The visual keypad resembled an ATM keypad but a new set of distractors was extracted from the visual database at every authentication attempt. Images in the database were clustered into 9 semantic categories according to the subject matter (flowers, animals, rocks, landscapes, humans, vegetables, buildings, skies, boats). Each image in the authentication code belonged to a different category and the distractors were selected from the remaining categories. In case of authentication failure, three attempts were given as in a normal ATM transaction.

VIP2 differed from VIP1 in that the 4 pictures forming the authentication code were displayed in random positions around the visual keypad at each authentication attempt. In case of authentication failure, the same visual configuration was displayed in order not to disclose any clue about the authentication code.

_____

Insert Table 1 about here

_____

VIP3 explores the limits of the graphical approach, by assigning a portfolio of 8 pictures to the user. At every authentication attempt, 4 of these pictures were randomly displayed in the challenge set together with 12 distractors (Figure 2). Thus,

there are $\frac{8!}{4!x4!} = 70$ possible code variations to display. Distractors were randomly selected from the database, avoiding duplication of the categories of the code items displayed in the current challenge set. To authenticate, users selected their images in any order. In case of authentication failure, the same visual configuration was displayed.

Insert Figure 2 about here

3.2 Security

In this section, we propose a criterion to evaluate the security of authentication systems and use it to compare the security achieved by each prototype under investigation. The proposal considers three basic security dimensions:

1. *guessability*: ability of a fraudster to guess the code;
2. *observability*: ability of a fraudster to observe the code as the user enters it; and
3. *recordability*: ability of the user to record the code, thereby making it easier for a fraudster to steal it.

For the sake of simplicity, we decided that each of these dimensions was equally important. Hence, we ranked the systems according to each metric and assigned a maximum value of 1 to the highest security solution and a minimum value of 0 to the lowest security solution.

PIN, VIP1 and VIP2's guessability is 1 in 10000 ($\frac{1}{10}x\frac{1}{10}x\frac{1}{10}x\frac{1}{10}$). The guessability of VIP3 is 1 in 1820 ($\frac{4}{16}x\frac{3}{15}x\frac{2}{14}x\frac{1}{13}$), as no duplicates can be used in this approach and order is unimportant. Note that since the user is only permitted 3 tries, this difference is less important than it initially appears. To provide a global guessability value, we assigned the highest value (1) to those systems which make it most difficult for a fraudster to guess the key (PIN, VIP1 and VIP2) and 0.2 to VIP3, as it reduces the guessability to almost 20% as compared to the best solution (Table 2).

Observation of the code involves two equally important features: (a) being able to actually see the key on the screen; and (b) being able to judge the position of the key based on where the person is pointing. Each of these features contributes half of the total weight of observability, thus their maximum score is 0.5. With respect to the the key being revealed on the screen, we assigned a value of 0 to PIN, VIP1 and VIP2 as the entire key can be observed and 0.5 to VIP3 as only part of the key can be observed. As regards the ability to use location to guess at the key, a value of 0 was assigned to PIN and VIP1 and a 0.5 was assigned to VIP2 and VIP3, because the targets are displayed in different positions at each authentication attempt.

As regards recordability, the systems were assigned the following values:

a. 0 if the code was easily recorded (PIN).
b. 0.5 if it was harder to record or describe, such as a simple representational image (see experiment 2).
c. 1 if it was difficult to record or describe, such as, for example, a complex image with many features (VIP1, VIP2 in the first experiment).

The total security score was obtained by adding the values for guessability (*G*), observability (*O*) and recordability (*R*). For the authentication mechanisms compared in this experiment, the security count is reported in Table 2.

_____

Insert Table 2 about here

_____

3.3 Hypotheses

The experiment was designed to collect a broad overview of usability and security issues of the visual approach in comparison with PIN and to test two hypotheses.

1) *Picture superiority hypothesis*: pictures are easier to remember than numbers, thus VIP1 should be better than PIN.
2) *Spatial coding hypothesis*: codes displayed in fixed locations are easier to remember as motor memory and memory for locations facilitate the retrieval, thus VIP1 should be better than VIP2.

3.4 Method

3.4.1 Participants

Participants were recruited by journal advertisement in the Dundee area, and paid £10 to participate in 2 experimental sessions over a week. Sixty-one persons were selected based on a brief phone interview, to guarantee that all were ATM users and did not have pathological memory deficits.

3.4.2 Procedure

On arriving for the experimental session, participants were given a questionnaire on their behaviour with, and attitude towards, ATMs and PINs, and randomly assigned to an experimental condition. After receiving verbal instructions, they underwent a 3-step automatic enrolment. In step 1 and step 3, all the pictures/numbers composing the authentication code were displayed on a line for 5 sec. In step 2, pictures were displayed individually for 5 sec. each. After enrolment, participants performed 10 authentications (training). They were instructed to swipe their card and enter the code as fast and accurately as possible. The first memory test (test1) took place after a 40 minute delay, during which participants performed a distraction task. The second test

(test2) took place a week later, followed by a questionnaire. Both tests adopted the same procedure as the training phase: participants were required to swipe their cards and enter their code as fast and accurately as possible for 10 times. In case of erroneous entry, participants were automatically given up to 3 attempts.

3.4.3 Design

The comparative evaluation was based on a between-subjects design. Participants were randomly assigned to one of four systems (PIN, VIP1, VIP2, VIP3). Data was collected at three stages: training, test1 and test2.

3.4.4 Dependent measures

The evaluation metric was defined along three major dimensions of usability: *effectiveness*, *efficiency* and *user satisfaction*.

- Effectiveness is associated with code memorability and defined in terms of number of people who forgot their code and numbers of wrong entries.
- Efficiency refers to speed of data entry.
- User satisfaction refers to the perception of the system relative to the perception of traditional keypad-based PIN devices.

Behavioural data were collected by automatic logging. Subjective data were collected by questionnaires and semi-structured interviews to capture user satisfaction and the strategies applied by participants in encoding and retrieving the authentication code.

3.5 Results

Participants were 29 males and 32 females, from 16 to 66 years (mean = 30). They reported using an average of 4 different PINs or passwords for a variety of devices. One person out of two also reported having difficulty remembering their bank PIN, and some 38% having had their card retained by an ATM as they were unable to remember the correct PIN. The main reasons were a mismatch between PIN and card, inexperience, or occasional use. Some 36% of the sample admitted having communicated their PIN to another person and 35% storing a written copy.

Before the experiment, participants were asked to evaluate their memory self-efficacy, by estimating their abilities to learn a set of 4 numbers/pictures and retrieve it a week later on a 4-point Likert scale (1 = not confident at all, 4 = very confident). On average, participants declared that numbers were easier to remember (mean = 2.57) than pictures (mean = 2.20). The difference is significant, $t_{(59)} = 3.75$ $p < .001$.

A total of 2196 authentication attempts were collected. It may be noted that this sample is greater than expected (61 people * 3 session * 10 authentications = 1830). This is due to the fact that when committing an error (N=118) participants were invited to repeat the authentication. Furthermore, several participants spontaneously entered their codes more times than required (N=248). No differences between the experimental conditions emerged with respect to this behaviour, thus we decided to retain and analyse all the collected authentications.

3.5.1 Effectiveness

No participant forgot their code, but some 5% of the authentications (118/2196) resulted in error (Figure 3). A crosstabulation analysis indicated that errors were concentrated in VIP3 which accounted for more wrong entries than all the other three conditions together (PIN, VIP1 and VIP2), $\chi^2_{(3)}= 57.08$, $p <.001$. A slight advantage of VIP1 over VIP2 was also observed, $\chi^2_{(1)}= 3.42$, $p = .07$, providing partial support to the spatial coding hypothesis. Contrary to our expectations, no significant difference between numbers and pictures (PIN vs. VIP1) emerged.

Insert Figure 3 about here

Overall, wrong actions were concentrated in the training phase and after the week interval (test2), while participants were consistently more precise at test1 (Figure 3). PIN and VIP1 presented an interesting difference. In the PIN condition, all but 1 error occurred at test2, while in the VIP1 condition, most of the errors occurred during training. Thereafter, they tended to disappear, with only one occurrence at test2. This trend supports the picture superiority hypothesis.

To understand the factors triggering errors, every wrong entry was tabulated according to its type. The following categories emerged from the analysis.

- *Erroneous selection*: one or more of the selected items did not belong to the authentication code.
- *Sequence*: the correct items were retrieved, but entered in the wrong order.
- *Duplicate selection*: the same item was selected twice in non-consecutive positions.
- *Double click*: the same item was unintentionally selected two consecutive times (the prototypes did not allow corrections).
- *Composite error:* it is composed of more than one error type (e.g., sequence error plus erroneous selection).

The graph in Figure 4 shows that different systems triggered specific errors. The poor performance of VIP3 was mainly due to erroneous selections (note that sequence errors could not occur in this condition). In some 92% of these errors, people tended to falsely identify distractors belonging to the same category of code items which were not displayed in the current challenge set (intra-category error). Flower images appeared to be particularly prone to this type of error, accounting for 63% of intra-category errors.

Insert Figure 4 about here

Inter-category errors (mismatch between different pictures with different subject matter) occurred mainly in condition VIP2 when targets were neither meaningful nor

easy to name (e.g., minerals). In this case, they were confused with distractors which had similar visual configurations even if they belonged to other semantic categories (e.g., an unusual yellow flower mistaken for a yellow mineral).

3.5.2 Efficiency

Efficiency was measured by the entry time (i.e., lag between code appearance and last selection) of correct authentications. A mixed design Anova, with experimental stage (3) as the within-subjects factor and system (4) as the between-subjects factor, was run using individual authentications as the unit of analysis. Post-hoc tests based on the LSD model (Least Significance Difference) were run, to test the picture superiority hypothesis and the spatial coding hypothesis. Mean values are illustrated in Figure 5.

Insert Figure 5 about here

Both main effects were highly significant (experimental stage: $F_{(2,1302)}= 58.93$, $p < .001$; system: $F_{(3,651)} = 84.35$, $p <.001$) and the interaction showed a slight tendency $F_{(6,1302)}= 2.04$, $p = .07$. The effect of stage is due to learning, as participants were significantly faster at test1 and test2 than during training. The effect of system indicates that design affected speed of data entry. VIP3 achieved the slowest performance. This is due to the need for visual scanning of a larger challenge set in order to locate the code, without knowing what items to look for. Post-hoc comparisons provided support for the spatial coding hypothesis, as participants using VIP1 were significantly faster than participants using VIP2 ($p <.001$). The picture superiority hypothesis was rejected, as users who entered numbers were faster than users who entered pictures ($p <.05$).

Analysing the global framework of performance data related to effectiveness and efficiency, we can notice that performance was not affected by any speed-accuracy trade-off. As the time increased there were also more errors (Figure 3 and Figure 5).

3.5.3 User satisfaction

For each user, two basic subjective measures were considered. A pre-test collected their satisfaction with the number-pad implementation of PIN devices, they had experienced before the experiment (baseline). Satisfaction with the device tested in the experiment was assessed after test2 (evaluation). In both cases, 7 bi-polar semantic differential items were used, addressing a range of usability and security dimensions (Cronbach $\alpha > .80$). The overall satisfaction indexes (average of individual items) were superior to the scale mid-point, reflecting a positive evaluation in all conditions.

A relative satisfaction index was obtained by subtracting baseline to evaluation and analysed by an ANOVA with condition (4) as the between-subjects factor. Results indicated a marginally significant effect of condition, $F_{(3,59)}= 2.58$, $p = .06$, due to a stronger improvement in satisfaction for participants who used VIP1 (mean values

PIN= - .29; VIP1= .85; VIP2= .17; VIP3= -.11). Post hoc analysis showed that VIP1 was perceived as better than PIN ($p < .01$), and VIP3 ($p <. 05$).

3.5.4 Mnemonic

Almost half of the sample (56%) reported having used a particular strategy to support code retrieval. Mnemonics for numbers were different from mnemonics for pictures. Repetition, chunking and association with date or math were used with the same frequency for numbers. Association was the prevailing strategy for pictures (98%), sometimes supported by repetition. Most participants associated pictures with words, often creating a story to support sequence retrieval. Participants occasionally reported associating objects with bigger pictures, emotions, or other memories.

3.6 Discussion

The study revealed important information about the memorability of pictures and an insight into cognitive constraints of visual memory in the context of user-authentication. Major findings are summarized below.

- The picture superiority hypothesis was rejected as no advantage in graphical mechanisms emerged.
- The spatial coding hypothesis was supported as a clear improvement in visual code retrieval occurred when targets were displayed in fixed locations.
- Errors in visual memory tended to follow a regular pattern: interference occurred between items sharing the same subject matter or similar visual configuration.

The failure to replicate the *picture superiority effect* is probably the most striking outcome of the study. No difference between VIP1 and PIN emerged in error occurrence. Rather, participants were significantly faster when entering numbers than pictures. We noticed, however, a trend in error occurrence over time, suggesting that sequences of numbers may be more susceptible to long-term decay than sequences of pictures. Pictures, on the other hand, were more subject to errors during the training phase, probably because of the lack of familiarity with the task. Differences in user satisfaction were in the expected direction: users preferred VIP1 and perceived it as more secure and easy to remember than PIN. Although these reactions need to be weighted due to the novelty factor, they suggest acceptance of the VIP paradigm after usage. However, participants were initially sceptic about the graphical solution, as they estimated their ability to remember pictures as weaker than their ability to remember numbers.

On the one hand, results of our study may be attributed to a ceiling effect, as participants' performance was extremely good in both numeric and visual conditions. Therefore, it is reasonable to believe that the experimental setting, requiring the user to learn one code and retain it for a short interval (a week), did not put enough stress on the user memory to evince differences due to the nature of processed stimuli. On the other hand, these results can be due to an overgeneralization of the picture superiority effect, typical of current interest in graphical authentication systems. Thus, it can hide a basic theoretical misunderstanding, which has led researchers to believe

that picture recognition capacity is essentially unlimited and that this could be directly exploited in user authentication.

As discussed in section 2, the studies of memory which motivated graphical authentication systems mostly date back to the mid 1960's. Since then the picture superiority effect has been often challenged and there is a consistent body of experiments where the effect was reduced or eliminated by experimental manipulation. For example, Nelson and colleagues (1977) inhibited the effect by presenting participants with conceptually dissimilar pictures that were drawn to be schematically similar. Dewhurst and Conway (1994) reversed the effect by instructing participants to imagine pictorial representations of verbal stimuli. There is a consensus that the picture superiority effect is relative to the set of mechanisms generating the observer's response. This raises a number of issues about the viability of cognometric systems, as the retrieval conditions necessary for secure and fast authentication are inherently different from the procedures used in memory testing, where items to be recognized are displayed individually or in pairs (see for example Standing et al., 1970; Shepard, 1967; Madigan, 1983).

Another important challenge to the picture superiority effect derives from a large body of research demonstrating a striking blindness to change in visual configurations under a variety of conditions (see, Rensink 2002 for a review). Change blindness is generally explained by the assumption that visual object representations decay rapidly after attention is withdrawn from an object. This implies that *Long Term Memory* (LTM) representation of a visual object is limited to meaning, layout, and perhaps the abstract identities of objects. A recent attempt to accommodate the picture superiority effect and change blindness is the *model of scene perception and memory* proposed by Hollingworth and Henderson (2002). The model assumes that when attention is oriented to an object in a scene, both low-level sensory processing and higher level processes occur. Higher level processes lead to a visual representation of the object specific to the observer's viewpoint, containing quite detailed information about its visual form, and to conceptual representations of its identity and meaning. These representations are indexed to a position in a map, coding the spatial layout of the scene and forming an object file, which is consolidated in LTM. Retrieval of LTM codes for previously attended objects is influenced by the allocation of visual attention, so that local object information can be retrieved by attending to the position in the scene at which information about that object was originally encoded.

The assumption that recognition is influenced by focussed attention is consistent with the spatial coding effect demonstrated in this experiment. Participants made fewer errors and were significantly faster when using VIP1 than VIP2, as object information was retrieved in the same position where it was encoded. In the experiment, the effect was intensified by kinaesthetic memory, coming from the hand movement. The memory model also provides an explanation to the errors evinced during the experiment. Fine discrimination was particularly difficult with familiar items sharing common subject matter and having similar visual configuration (intra-category error). For example, if a person had a flower in their portfolio, they were induced to identify similar flowers in the challenge set as their flower. It is plausible to assume that due to their familiarity these images were processed without the amount of visual attention needed to detect change. Inter-category interference, on the other hand, tended to

occur with those objects which did not generate a clear conceptual representation of their identity and meaning in LTM during encoding.

4. Study 2

In the second design iteration we concentrated on VIP1 and VIP3 mechanisms, as they were the two extremes of the usability continuum. Two web-based prototypes were designed, with a new and larger set of images, which were deemed to be easier to remember (Figure 6). These images represented simple, familiar and concrete everyday objects on a white background, and were easier to name, more distinctive and less complex than those used in Experiment 1. We also used a larger set of semantic categories (N=16), so that it was easier to control the display of the challenge set, avoiding using any images of the same categories as the challenge set amongst the distractors, which appeared very detrimental in VIP3.

_____

Insert Figure 6 about here

_____

4.1 Authentication mechanisms

The study compared the PIN mechanism with a revised version of VIP1 and VIP3. PIN and VIP1 were based on the same mechanism presented in study 1, but this time participants were asked to memorise a sequence of 5 rather than 4 items, and the input device was a mouse rather than the touch screen.

VIP3 was simplified by reducing the length of the portfolio to 6 pictures and avoiding any duplicates in the challenge set. At authentication, 5 pictures were randomly displayed in the challenge set together with 11 distractors (Figure 6) which means only 6 different variations are possible. Ideally the decrease in security of a smaller portfolio would be offset by an increase in usability.

4.2 Security

The evaluation of security followed the criterion and methodology presented in Study 1. The guessability value is increased for PIN and VIP1 to 1 in 100000 ($\frac{1}{10} x \frac{1}{10} x \frac{1}{10} x \frac{1}{10} x \frac{1}{10}$) and for VIP3 to 1 in 4368 ($\frac{5}{16} x \frac{4}{15} x \frac{3}{14} x \frac{2}{13} x \frac{1}{12}$). The value of recordability is also reduced in this experiment, since the new set of images were easier to record. Security indexes are reported in Table 3.

_____

Insert Table 3 about here

_____

4.3 Hypotheses

The new prototypes were tested in a longitudinal web-based experiment over a period of four months. We believed that the increase in ecological validity associated to this paradigm would compensate for the decrease in experimental control typical of on-line experiments. The study was designed to test the picture superiority hypothesis over a longer time period (PIN vs. VIP1), with prototypes believed to be more usable, and to evaluate the effect of different authentication mechanisms on usability (VIP1 vs. VIP3).

- *Picture superiority hypothesis*: pictures are easier to remember than numbers, thus VIP1 should be better than PIN.

4.4 Method

4.4.1 Participants

Sixty-three third year computing students at the University of Glasgow volunteered to participate in the study. All participants reported normal or corrected-to-normal vision.

4.4.2 Procedure

The evaluation consisted of a web-based experiment running from January to May 2003. The authentication mechanisms controlled access to a personalised web page containing information on, and resources for, a University module, thus creating the need for participants to be authenticated. When they first accessed the module page, participants were randomly assigned to an experimental condition and given an authentication code via an automatic enrolment procedure, similar to that described in experiment 1. This code had to be entered, every time they wanted to access the web-site. In case of 3 erroneous entries participants were automatically given a new code and invited to re-enrol. All accesses to the web site were logged.

4.4.3 Design

System (3) was manipulated between-subjects. Participants were randomly assigned to one of three experimental conditions corresponding to different systems (PIN, VIP1 and VIP3).

4.4.4 Dependent measures

The evaluation concentrated on *effectiveness* (errors in memory) and *efficiency* (speed of data entry). User satisfaction could not be analysed due to the high dropout rate in the post test questionnaire.

4.5 Results

We discarded all participants (N=4) who used the authentication system less than once a month. The statistics reported are thus based on a sample of 59 users, (21 in the PIN condition, 19 in VIP1 and VIP3) and 943 authentication attempts. On average,

each participant performed 16 authentication attempts. No difference in frequency of use emerged ($F<1$).

4.5.1 Effectiveness

During the four months of the study, 24 people (41% of the sample) had to be given a new authentication code at least once, as they were incapable of remembering their code after 3 attempts. Six of them had to be given a new code twice. The likelihood of forgetting the code was affected by the system used. Only 6 people using VIP3 forgot their code, versus 10 people in the PIN condition and 9 people in the VIP1 condition.

A total of 198 authentication attempts (21%) resulted in errors. Comparing the proportion of errors out of the number of authentications for each participant and experimental condition, a significant effect of system emerged ($\chi^2_{(2)}= 6.33$, $p <.05$). Some 27% of PIN authentications resulted in errors; versus 24% for VIP1 and 11% for VIP3. A Mann-Whitney U test indicated that the difference between PIN and VIP3 is significant ($U=117.5$, $N=41$, $p <.05$), while there is no difference between PIN and VIP1. Errors were clustered in 5 categories according to their cause.

- *Erroneous selection*: one selected item did not belong to the authentication code; but the code was entered in the correct sequence, if required.
- *Sequence*: all the correct items were retrieved but entered in a wrong order.
- *Erroneous code*: three or more of the selected items did not belong to the authentication code suggesting that participants entered a different code.
- *Composite error*: one or two selected items did not belong to the authentication code and, if required, the sequence was wrong.
- *Blank*: no items or a partial number of items were entered.

Insert Figure 7 about here

Different systems afforded specific errors (Figure 7). The most typical failure in condition VIP3 was the wrong selection of one individual item. Almost half of the errors in condition VIP1 were failures in retrieving the exact sequence of the code, while the most common error in the PIN condition was due to the retrieval of a completely different code.

Figure 8 shows the percentage of correct authentication at four intervals after the last successful authentication. Note that the last category reports a broad interval ranging from 1 week to 101 days, data were nevertheless collapsed to ensure a representative sample ($N=173$). The graph reports only data relative to 1st authentication attempts and thus do not includes cases of multiple errors (the user had 3 possibilities before the system was locked) as they tend to concentrate in the same day and would have biased the analysis. The trend in correct 1st authentication attempts suggests that when users logged in the same day of a successful authentication, the PIN system achieved the best performance. Thereafter, the picture superiority effect clearly appeared as pictures were less affected by time decline.

Insert Figure 8 about here

4.5.2 Efficiency

Effectiveness was measured considering entry time (lag between code appearance and last entry) in correct authentications (N=746). On the average participants took almost 7 sec. to select their code. An ANOVA revealed a significant effect of system $F_{(2,745)}= 5.57$, $p < .01$. LSD post-hoc analysis indicated that the longest time required by VIP3 is entirely responsible for this effect, and that there is no difference between PIN and VIP1. Mean values and standard errors are reported in Table 5.

Insert Table 4 about here

4.6 Discussion

The second design iteration of the VIP project succeeded in creating a graphical mechanism which generated fewer errors in memory than traditional PIN. Surprisingly, this mechanism was not the visual equivalent of the PIN procedure (VIP1) that in Experiment 1 had demonstrated greater potential, but a revised version of the portfolio-based mechanism (VIP3) that in Experiment 1 had achieved a very poor performance. The difference may be due to a number of factors: including shorter portfolio (6 instead of 8 pictures), a new set of pictures (simpler, more concrete and clearer objects), and a more controlled challenge set configuration (avoiding intra-category distractors). The new version of the portfolio-based solution thus succeeded in providing a mechanism which decreased the probability of forgetting the code of almost 17%, and of entering a wrong code of almost 16%, as compared to numeric codes. However, such system was less efficient than the other solutions, as code entry required almost 2.5 seconds longer than traditional PIN.

The longer time required for code entry can be partially responsible for the higher accuracy in memory performance. However, in visual memory literature, longer reaction time has traditionally been associated to increased difficulty in retrieval. In our study the longer entry time is also due to a different retrieval context which could not benefit from 'spatial coding' (Hollingworth & Henderson, 2002), but which required scanning a larger visual configuration without knowing in advance what items to look for and where to look for them. The beneficial effect of spatial coding as a cue to code retrieval was demonstrated in Experiment 1.

The advantage of VIP3 over VIP1 may be due to the fact that VIP3 is a pure recognition-based system, with no requirement for the user to remember a sequence, which generated most of the errors in the VIP1 condition. The difficulty in remembering a visual sequence already appeared in Study 1, but this time it was aggravated by a more realistic situation with longer intervals between retrieval, and the lack of important kinaesthetic cues provided by the hand movement in touch-screen entry, as the interaction was mediated by means of a mouse.

Another important result of this study regards the difference in error occurrences over time between graphical and numerical codes. The picture superiority effect emerged when participants did not authenticate for a few days and tended to increase as the time passed. It has to be noted that a typical error in PIN retrieval was caused by interference with different numerical codes participants may have used for other authentication systems. Due the novelty of the solution, this type of error could not occur in the graphical conditions.

The good performance of VIP3 is an important result for public technology. Indeed, a portfolio-based graphical mechanism has the potential for improving some aspects of security. In particular, it protects against shoulder surfing since there are a number of permutations which may lead to the display of different codes. Furthermore, it complicates the communication of the code to others, as it introduces a level of uncertainty about the displayed items. This implies that when another person has to enter the code, she has to undergo higher level cognitive processes comparing her 'verbal knowledge' with the visual layout, which may be difficult.

5. Guidelines for successful authentication

The experiments reported in this paper provide further evidence of the need for better authentication systems, evidenced by personal preferences, and obvious memory difficulties experienced by participants in both experiments. They also show that whereas the graphical approach demonstrates some potential for simplification of user authentication, it is not a simple panacea due to the many usability and security problems of the mechanism. Not only are cognometric systems affected by the same tricky trade-offs *between* security and usability as PINs and passwords, but they also face additional conflicts *within* security and usability dimensions. For instance, a portfolio-based solution decreases the risk related to observability and recordability of the code but strongly increases the probability that the code may be guessed, as compared to a 5 key order-significant code (1/100000). To get approximately the same odds for order-insignificant portfolio-based solutions, we would have to display 23 distractors for a 5 image key, which would have a detrimental effect on the time taken to authenticate and thereby impact usability. As regards usability, we evinced a similar conflict between effectiveness and efficiency, with portfolio-based solutions inducing fewer errors but requiring longer authentication time as compared to order-significant solutions.

From the previous discussion, it is clear that every authentication mechanism has its own set of security and usability problems, and that only a detailed analysis of task, system, and user requirements will lead to the best choice (G1). This is a trivial, but frequently disregarded, tenet because the same level of complexity (recalling a PIN) is currently required to withdraw money at an ATM, or to place a call on a mobile,

whereas the task and system requirements are very different. The results reported in this paper have the potential for informing the design of cognometric systems for PC-based applications where the attainable level of security of graphical mechanisms is deemed appropriate for the application. As observed before, the security level achieved by graphical mechanisms has inferior guessability compared to most current authentication mechanisms based on 6-8 alphanumeric characters. It should be noted, however, that the superior guessability can only be attained if the user chooses a completely random selection of characters and not a simple word, which is often the case (Adams and Sasse 1999). More research is needed to investigate graphical mechanisms for hand-held devices, where screen size and resolution are likely to inhibit the picture superiority effect (system level). At the user level, it is important to consider that graphical authentication may be difficult, if not impossible, for users with visual disabilities, thus an alternative mechanism is required to ensure universal accessibility.

To design successful authentication mechanisms, system, task and user requirements need to be evaluated against several dimensions of security and usability. Hence the rest of this section presents some guidelines developed within this framework.

G1. Authentication mechanisms must be selected according to system, task and user requirements.
   - Maximise usability whenever security is not paramount; otherwise make sure to educate the user to comply with the security policy.
   - Security is a multi-facet concept, whose basic aspects are guessability, observability and recordability (see 3.1. in this paper). The relative importance of each of these dimensions varies according to task requirements and need to be established in advance as a fundamental system requirement.
   - Usability is a multi-facet concept, whose basic aspects are effectiveness, efficiency and user satisfaction (ISO, 1997). The relative importance of each of these dimensions varies according to task and user requirements and need to be established in advance as a fundamental system requirement.

G2. Concrete, nameable, and distinctive colour images are easier to remember; thus, they tend to improve all aspects of usability but they decrease security (recordability).
   - Test your image set with real users whenever possible, or use a visual database similar to those tested and reported in the literature.

G3. Control the visual configuration of the challenge set by
   - displaying distractors from different semantic categories from those in the challenge set (to increase usability);
   - displaying visually dissimilar distractors from those in the challenge set (to increase usability); and
   - using as many categories as possible so that distractors can be drawn from a wide set of possibilities (to increase security)

G4. System-allocated codes have a positive effect on security (reducing predictability) but they may affect usability (being more difficult to remember).

G5. Keys displayed in fixed locations at each authentication attempt increase usability but decrease security (observability).

G6. Portfolio-based solutions increase usability (effectiveness) and affect security, increasing guessability but decreasing observability, as a new challenge set is presented at each authentication attempt.

6. Conclusion

This paper reported a user-centred approach to the design of cognometric mechanisms, based on the exploitation of visual memory for user authentication in self-service technology. Our experience has demonstrated that the design of successful authentication mechanism is a complex task, as it requires considering and weighting several important factors to reach maximum security and usability. There is no 'magical solution' and the conflict between these two objectives sometimes appears insurmountable. The contribution of this paper lies in defining some of the factors which may affect usability and security of graphical authentication mechanisms. The most severe limitation of the work is that no attention has been given to accessibility issues and on how the proposal should be modified to meet the needs of different categories of visually impaired users. Another interesting area to explore is the potential of the graphical approach for the elderly, as some evidence suggests that visual memory is less affected by the general cognitive decline associated with ageing as other types of memory (Park et al., 1986).

More user research is needed to fully understand the potential and limitations of the visual paradigm in authentication systems, with particular attention paid to the effects of multiple visual codes, and biases which occur in user selection of codes (Davis et al. 2004; Renaud and De Angeli, 2004). It is important to understand that if graphical authentication mechanisms are here to stay people are likely to get better at using them due to practice, but that their efficiency may be affected by competing codes, as currently occurs with PINs and passwords. Possible solutions to this problem, subject to empirical validation, are '*application specific'* visual codes, where something in the code is associated with the application (a theme or a background colour). This paper demonstrates how this research could be conducted and offers a solid empirical base to start the foundation of a specific model of visual memory in the authentication context.

7. References

Adams, A., Sasse, M.A. 1999. Users are not the enemy. Communications of the ACM 42 (12), 40-46.

Besnard, D., Arief, B., 2004. Computer security impaired by legitimate users. Computers & Security 23, 253-264.

Blonder, G. E. 1996, Graphical password. United States Patent 5559961.

Brostoff, S., Sasse, A., 2000. Are passfaces more usable than passwords? A field trial investigation. Proceedings of HCI 2000, 405-424.

Coventry, L., De Angeli, A., Johnson, G.I., 2003a. Usability and biometric verification at the ATM interface. CHI 2003 Proceedings, 153-160.

Coventry, L., De Angeli, A., Johnson, G.I., 2003b. Biometric verification at a self-service interface, in: McCabe, P.T. (Ed.), Contemporary Ergonomics 2003. Taylor & Francis, London, pp. 247-252.

Davis, D., Monrose, M.,and Reiter, M., 2004. On user choice in graphical password schemes. Proceedings of the 13th USENIX Security Symposium, 151-164.

De Angeli, A., Coutts, M., Coventry, L., Johnson, G.I, Cameron, D., Fischer, M., 2002. VIP: a visual approach to user authentication. Proceedings of the Working Conference on Advanced Visual Interfaces AVI 2002, 316-323.

De Angeli, A., Coventry, L, Johnson, G.I., Coutts, M. 2003. Usability and user authentication: Pictorial passwords vs. pin, in McCabe, P.T. (Ed.), Contemporary Ergonomics 2003. Taylor & Francis, London, pp. 253-258.

Dewhurst, S. A., Conway, M. A., 1994. Pictures, images, and recollective experience. Journal of Experimental Psychology: Learning, Memory, and Cognition 20, 1088-1098.

Dhamija, R., Perrig, A., 2000. Déjà vu: A user study using images for authentication. Proceedings of 9th USENIX Security Symposium, 45-58.

Goldberg, J., Hangman, J. and Sazawal, V. 2002. Doodling our way to better authentication. CHI 2002 Extended Abstracts, 868-869.

Higbee, K.L., 1988. Your memory: how it works and how to improve it. Prentice-Hall Press, New York, 2nd Edition.

Hollingworth, A., Henderson, J. M., 2002. Accurate visual memory for previously attended objects in natural scenes. Journal of Experimental Psychology: Human Perception and Performance 28, 113–136.

ISO, 1997. Ergonomics requirements for office work with visual display terminal (vdt) - parts 1-17. International Standard Organization 9241, geneva, switzerland

Jansen, W., Gavrila, S., Korolev, V., Ayers, R., Swanstrom., R., 2003. Picture password: A visual login technique for mobile devices. NISTIR 7030. Available online: http://csrc.nist.gov/publications/nistir/nistir-7030.pdf.

Jermyn, I., Mayer, A., Monrose, F., Reiter, M.K., Rubin, A.D., 1999. The design and Analysis of Graphical Passwords. Proceedings of the 8th USENIX Security Symposium, 1-14.

Madigan, S., 1983/ Picture memory, in: Yuille J.C. (Ed.), Imagery, memory, and cognition: Essays in honor of Allan Paivio. Erlbaum, Hillsdale, NJ, pp. 66-89.

Nelson, D. L., Reed, V.S., McEvoy, C.L. (1977). Learning to order pictures and words: A model of sensory and semantic encoding. Journal of Experimental Psychology: Human Learning and Memory 3, 485-497.

Nickerson, R. S., 1965. Short-term memory for complex meaningful visual configurations: A demonstration of capacity. Canadian Journal of Psychology 19, 155–160.

Paivio, A, Rogers, T.B., Smythe, P.C., 1968. Why are pictures easier to recall than words? Psychonomic Science 11 (4), 137-138.

Paivio, A., 1971. Imagery and verbal processes. Holt, Rinheart & Winston, New York.

Paivio, A., 1983. The empirical case for dual coding. In: Yuille J.C. (Ed.), Imagery, memory, and cognition: Essays in honor of Allan Paivio. Erlbaum, Hillsdale, NJ, pp. 9-33.

Park, D.C., Puglisi, J.T., Smith, A.D., 1986. Memory for pictures: does an age related decline exist? Journal of Psychology and Aging 1, 11-17.

Real User Corporation. 2004. The Science Behind Passfaces. Available online: http://www.realuser.com/published/ScienceBehindPassfaces.pdf

Renaud, K., De Angeli, A., 2004. My password is here! An investigation into visuo-spatial authentication mechanisms. Interacting with Computers 16 (6), 1017-1041.

Rensink R.A., 2002. Change Detection. Annual Review of Psychology, 53, 245-277.

Sasse, M.A., Brostoff, S., Weirich, D., 2001. Transforming the 'weakest link': a human-computer interaction approach to usable and effective security. BT Technology Journal 19 (3), 122-131.

Shepard, R.N., 1967. Recognition memory for words, sentences and pictures. Journal of Verbal Learning and Verbal Behavior 6, 156-163.

Standing, L., 1973. Learning 10,000 pictures. Quarterly Journal of Experimental Psychology 25, 207-222.

Standing, L., Conezio, J., Haber, R. N., 1970. Perception and memory for pictures: Single-trial learning of 2500 visual stimuli. Psychonomic Science 19, 73–74.

Thorpe, J and van Oorschot, P, 2004. Graphical dictionaries and the memorable space of graphical passwords. Proceeding of the 13th USENIX Security Symposium, 135-140.

Weinshall, D., Kirkpatrick, S., 2004. Passwords you'll never forget, but can't recall. CHI 2004 Extended Abstract, 1399-1402.

Table list

Figure list

Table 1. Systems tested in experiment 1

System	Type of code	Key Location	Key Order
PIN	Sequence of 4 numbers from 10	Constant	Fixed
VIP1	Sequence of 4 pictures from 10	Constant	Fixed
VIP2	Sequence of 4 pictures from 10	Random	Fixed
VIP3	Portfolio based (4 pictures from 12)	Random	Random

Table 2. Security count for experiment 1

		PIN	VIP1	VIP2	VIP3
Guessability		1	1	1	0.2
Observability	key	0	0	0	0.5
	position	0	0	0.5	0.5
Recordability		0	1	1	1
Security value		**1**	**2**	**2.5**	**2.2**

Table 3. Security count for experiment 21

		PIN	VIP1	VIP3
Guessability		1	1	0.04
Observability	key	0	0	0.5
	position	0	0	0.5
Recordability		0	0.5	0.5
Security value		**1**	**1.5**	**1.54**

Table 4. Entry time as a function of experimental conditions

	PIN	**VIP1**	**VIP3**
Mean (ms)	6529.92	5027.13	9184.98
Standard errors	756.72	493.81	1203.84

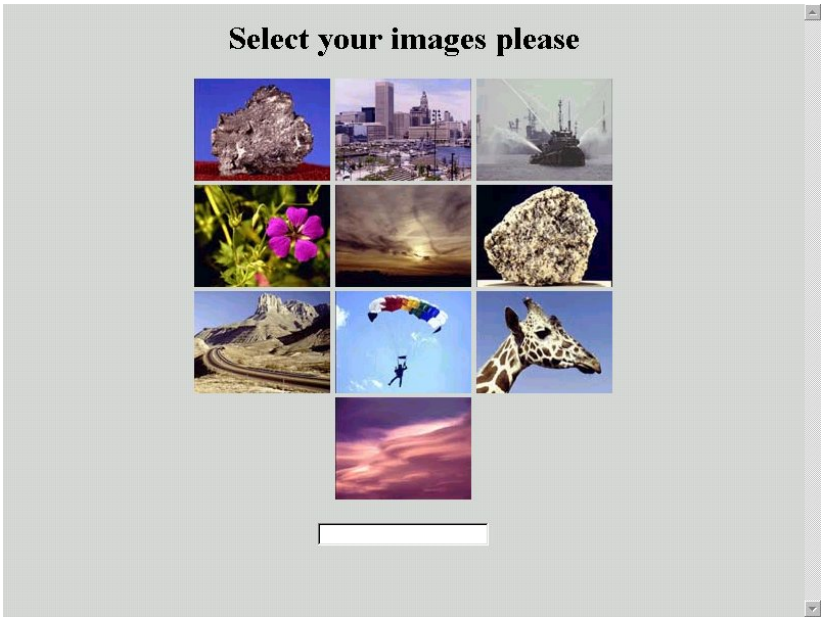

Figure 1. VIP1/VIP 2 interface

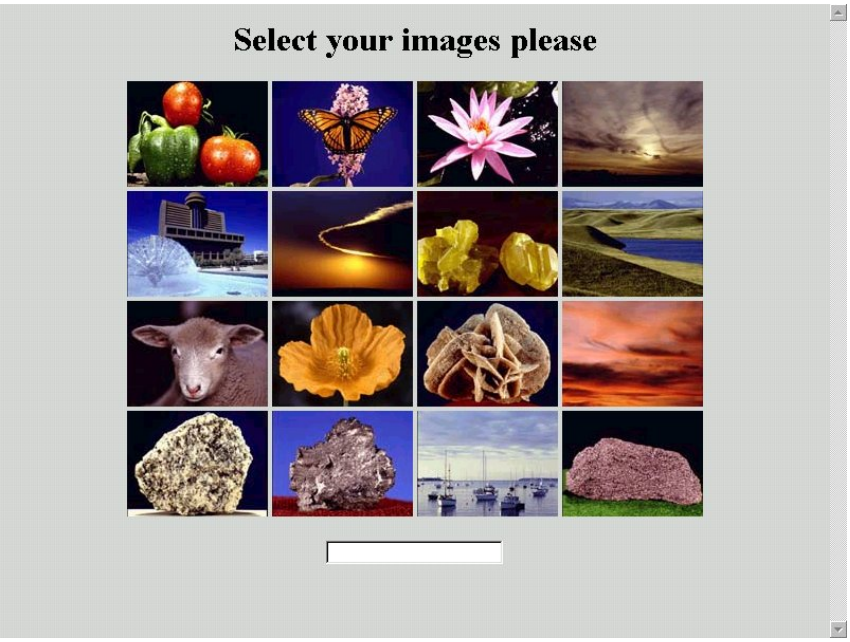

Figure 2. VIP3 interface.

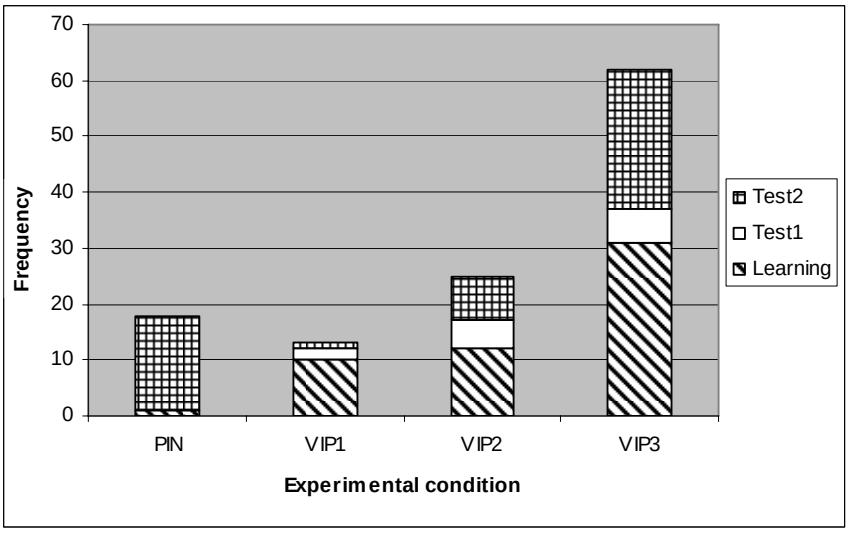

Figure 3. Distribution of errors in the experimental conditions

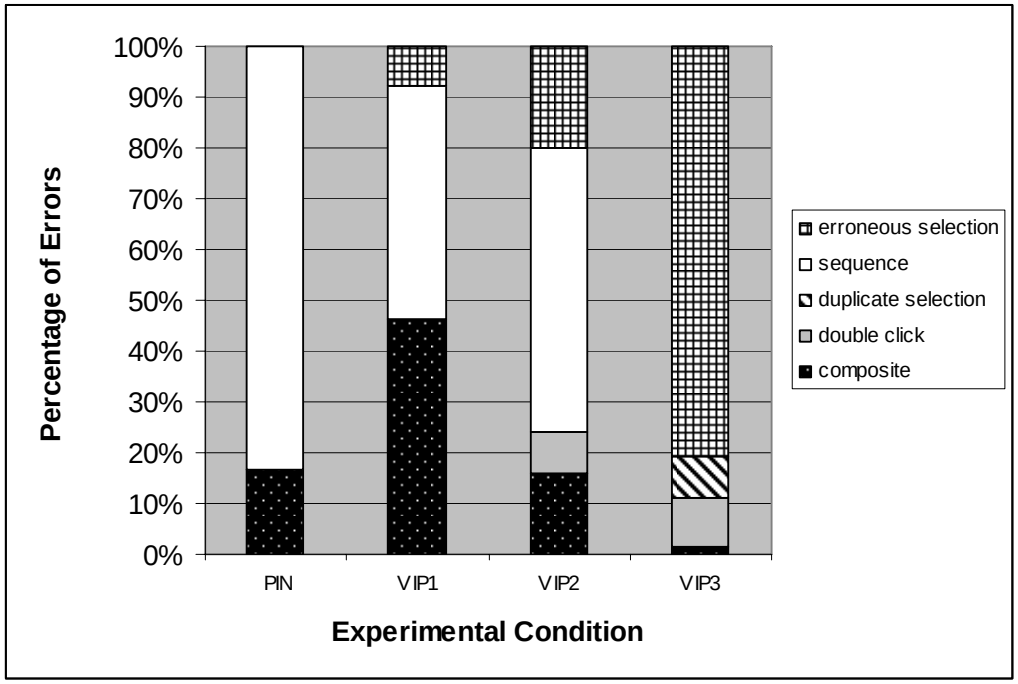

Figure 4. Error types as a function of system

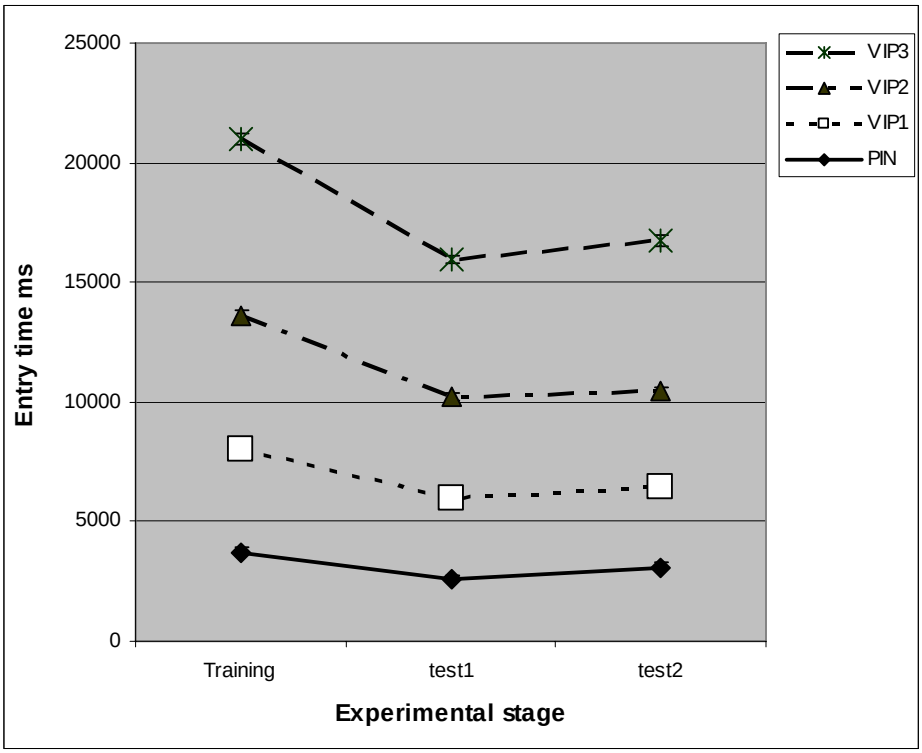

Figure 5. Entry time as a function of experimental conditions

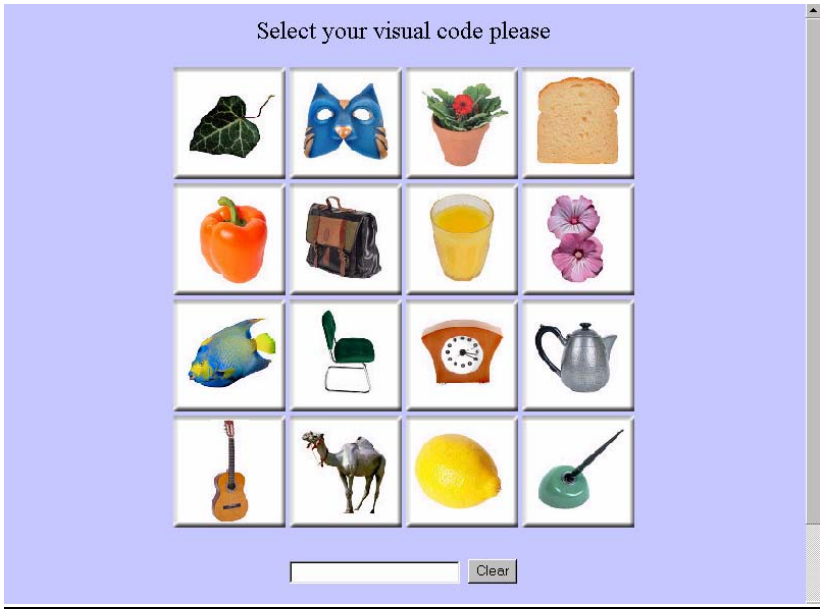

Figure 6. New interface of VIP3

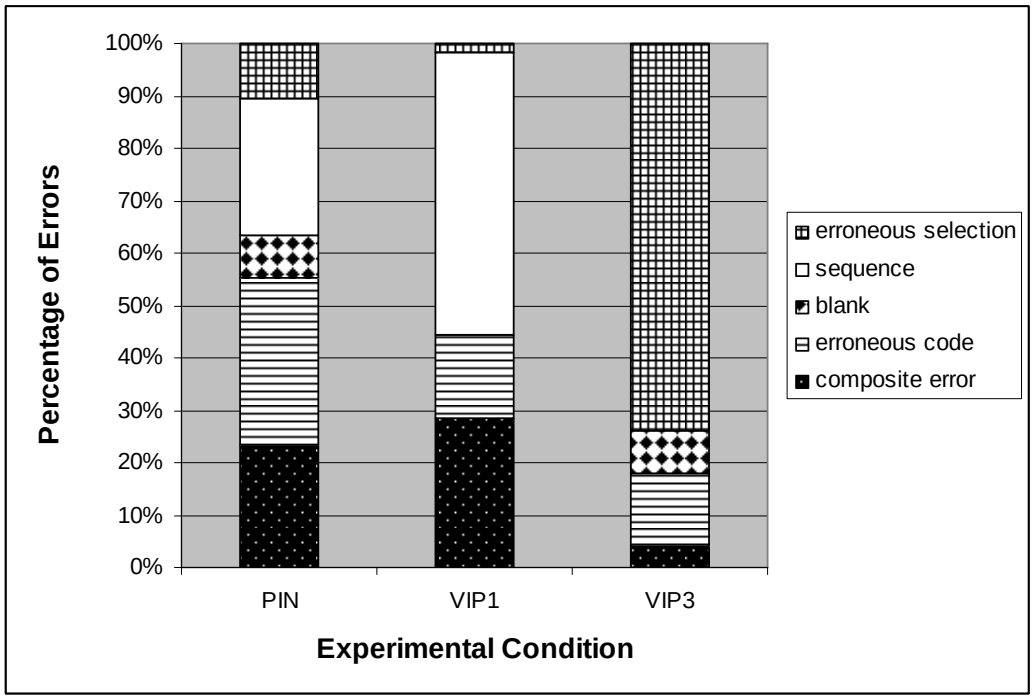

Figure 7. Percentage of error types in the experimental conditions

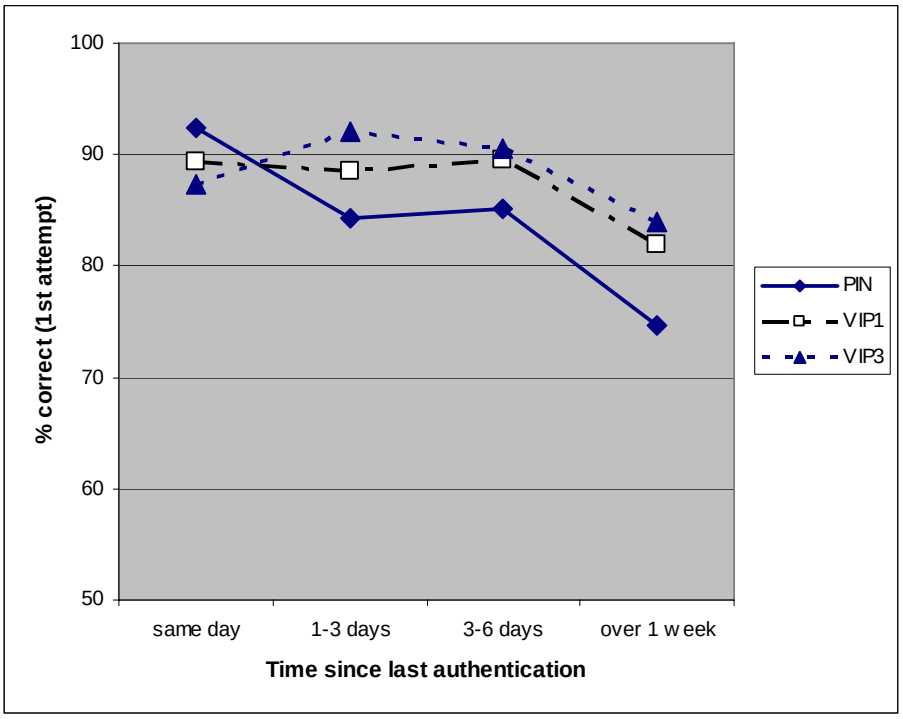

Figure 8. Percentage of correct authentications as a function of System and Time