On Automated Image Choice for Secure and Usable Graphical Passwords

Paul Dunphy
Culture Lab
School of Computing Science
Newcastle University
paul.dunphy@ncl.ac.uk

Patrick Olivier
Culture Lab
School of Computing Science
Newcastle University
p.l.olivier@ncl.ac.uk

ABSTRACT
The usability of graphical passwords based upon recognition of images is widely explored. However, it is likely that their observed high memorability is contingent on certain attributes of the image sets presented to users. Characterizing this relationship remains an open problem; for example, there is no systematic (and empirically verified) method to determine how similarity between the elements of an image set impacts the usability of the login challenge. Strategies to assemble suitable images are usually carried out by hand, which represents a significant barrier to uptake as the process has usability and security implications. In this paper, we explore the role of simple image processing techniques to provide automated assembly of usable login challenges in the context of recognition-based graphical passwords. We firstly carry out a user study to obtain a similarity ranked image set, and use the results to select an optimal per-pixel image similarity metric. Then we conduct a short-term image recall test using Amazon Mechanical Turk with 343 subjects where we manipulated the similarity present in image grids. In the most significant case, we found that our automated methods to choose decoy images could impact the login success rate by 40%, and the median login duration by 35 seconds.

Categories and Subject Descriptors
K.6.5 [Management of Computing and Information Systems]: Security and Protection: Authentication

General Terms
Experimentation, Security, Human Factors

Keywords
Usability, Security, User Authentication

1. INTRODUCTION
Users of alphanumeric passwords are widely known to choose credentials that are sufficiently predictable to undermine their principal security benefits [17]. This has led to the proposal of strategies of password selection designed to mitigate such predictability such as passphrases and mnemonics [19]. However, the success of these strategies, relies upon their adoption by conscientious users, and the strategies themselves may even contribute to user choice biases of their own. Graphical passwords [2][36] based upon the recognition of previously seen images have been proposed and evaluated in order to ascertain their suitability as a viable alternative to passwords and Personal Identification Numbers (PINs). Systems have been empirically evaluated with a password space ranging from 11 bits [13] to over 50 bits [40], and results highlight promising usability. The principal benefit of such systems is thought to be that users are more likely to remember assigned authentication credentials, thereby eliminating issues relating to predictable user choice. This genre of system provides a simple user interaction, as users must perform a visual search to identify previously assigned images (key images) amongst decoy images. The sequence of key images comprises the graphical password. Systems initialized using personal images [26,37,38] are particularly promising due to the ubiquity of large image collections, in addition due to observed usability benefits where users have had involvement to capture or create the images [28].

However, it appears likely that the usability benefits of this genre of graphical password are contingent upon subtle attributes of the image sets that are presented to users. For instance, it has been noted that visual and semantic similarity exhibited between images has the potential to disrupt the picture superiority effect [32] by causing errors in visual search [7] and rote learning [34]. For security and simplicity, it would be desirable that image sets are assembled randomly; however, the constraints that clearly exist when creating a usable login challenge imply that some degree of skill and effort is required to do this effectively, and in a manner that preserves security. Figure 1 illustrates different extremes of assembling decoy images for a particular key image. Zurko and Simon [43] remind us that user-centered security should be proportioned between end-users, developers and system administrators. Currently such a holistic consideration is lacking, as there is currently no systematic or empirically verified convention to reason over the similarity in an image set, and as a result this

Figure 1: Extremes of decoy image selection for the same key image: (left) decoys are semantically different; (centre) semantic and visual similarity to key image; (right) decoys are semantically similar yet different from the key image.
process must be performed by hand on the basis of common sense judgments of image semantics [1]. Users themselves could be asked to tag for similarity, but this can present security threats if users attempt to circumvent the process to obtain an overly simplistic login task (e.g. Figure 1, right image). This absence of a systematic means to evaluate image similarity, combined with the potential impact of inappropriate levels of similarity on authentication error rates, constitutes a significant barrier to the real world deployment of graphical passwords. Indeed, the ability to spontaneously generate usable image sets likely could present security benefits in terms of guessability, phishing and observation attack, and becomes more pressing when considering likely deployment level phenomena such as password resets, where new image sets would need to be generated in response to user demand.

An as-yet unexplored approach to solve this problem is to harness image processing research from the field of content-based image retrieval (CBIR) [12]. One fundamental challenge is to determine whether two images are similar. In this field the underlying assumption is that images with similar visual characteristics are more likely to be semantically similar [39]. Our contribution is to explore the efficacy of a systematic method to identify instances of visual similarity between images, and explore the impact of its careful manipulation upon the usability and security of images presented in a graphical password login. We carry out two user studies: firstly we gain a human consensus on the visual similarity within an image set, and use those results to identify a mechanism to detect digital image similarity that can best represent our collected human consensus. Then, in a second study we use the most promising method to generate graphical password logins for a user study conducted on Amazon Mechanical Turk. The study captured participant performance in a short-term recall task where the visual similarity present in the login was manipulated to differing degrees. Finally, we consider the security implications of systematic filtering, identify attacks that can result, and propose countermeasures.

2. RELATED WORK

2.1 Graphical Passwords

Recognition-based graphical passwords [36][2] have received increasing research attention due to experimental results in cognitive psychology that suggest the existence of a picture superiority effect [32]: that images are retained in memory better than words or numbers. This has led to a number of instantiations of user authentication systems that aim to harness this effect. Passfaces [25] harnesses human ability to recognize faces. In a field study across three months, it was discovered that users of this system made one third of the errors made by those using traditional passwords [4]. Findings from the field suggest that asking users to remember a specific ordering of the images causes errors [1][18], that the rate at which users are introduced to new systems can impact retention [10], and that the memorability of images increases the more involvement the user has in their creation [28]. Tullis and Tedesco [38] performed a series of studies on a photograph-based system and found users to accurately recognize personal photographs amongst stock photographs, even when stock photographs were hand-picked to be semantically similar to their own images. The same authors also reported impressive user recall performance six years after the initial study[37].

Image processing is beginning to play a role in the usability and security of graphical passwords. Dirik et al. [6] proposed image processing techniques that could serve to predict user choice in the Passpoints [41] system. Focusing upon two particular images, they performed object segmentation, calculated probabilities for the objects users would likely focus their attention upon, and calculated the centroids of those objects to serve as candidate click points. By doing so they were able to guess 80% of user click points they collected in a user study. Salehi-Abari et al. [31] developed these ideas further and used corner detection to identify candidate click points, and optimized their guessing strategy according to techniques likely to be adopted by users in graphical password selection. They assembled a 34.7 bit attack dictionary and were able to guess (depending on the image) 48% to 54% of entire click point sequences. Dynahand [27] considered the problem of eliminating visual similarity from image sets where the image stimuli took the form of freehand doodles.

2.2 Similarity and Visual Search

There are a range of definitions of similarity. Medin [22] suggests that: “Similarity between two things increases as a function of the number of features or properties they share and decreases as a function of mismatching or distinctive features”. Smith et al. [34] used the Farnsworth-Munsell 100 Hue test in a paired associate learning context to explore the impact of visual similarity upon rote learning. Here, colors with different controlled levels of similarity were assigned textual tags and participants were challenged to recall the tag when presented with a color. They found that when stimuli were similar, the performance scores for recall were worse and suggest that rote learning decreases as a function of the number of stimuli relevant for a particular textual tag. In the field of visual search, what we refer to as key images are referred to as target images, and decoys are referred to as non-targets. Duncan and Humphreys [7] explored the efficiency of visual search when using alphabetic characters and hand manipulating the choice of distractors to varying degrees of similarity. They concluded that there exist two types of similarity: within-object conjunctions: where the spatial arrangement of strokes is similar (e.g. L vs. T); across-object conjunctions: where the target can be formed by recombining strokes from different non-targets (e.g. find R amongst a set of P and Q characters). In addition, they identify interesting properties of visual search: that it takes longer to decide that a target is absent than to say it is present, and that target images can be camouflaged if placed next to similar non-targets.

2.3 Digital Image Processing

The most common way to compare digital images for similarity is to create and store signatures of an image that represent particular features e.g. color, and compute distances between these signatures. Choosing the most appropriate image signature is a context specific task. Color is the most widely used attribute in image retrieval and object recognition [20]. The SIMPLIcity system [39] attempts to systematically categorize the image type before choosing the most appropriate feature representation. Histogram-based methods are widely studied and are considered to be effective. Rubner et al. [29] propose the use of the Earth Mover’s Distance (EMD) as a method for calculating a distance between two color histograms for purposes of determining their similarity. This reflects the minimal cost of transforming one distribution into another. Lv et al. [21] explore using a modified version of the EMD for similarity matching and propose the average precision metric. Pass and Zabih [24] suggest adding multiple dimensions to histograms to include characteristics other
than pure color, such as spatial information. Selection of the most appropriate color space can also be an important decision, as some are more perceptually linear than others and so lend themselves better to reasoning over image similarity [11].

3. IMAGE FILTERING AND GRAPHICAL PASSWORDS

We firstly define some terms: the image set comprises all the images available to the authentication system; the login challenge is a subset of the image set, which is comprised of both key images and decoy images and is presented to the user at login. There are a number of conventions regarding the presentation of images and decoy images and is presented to the user at login. There are a number of conventions regarding the presentation of the login challenge to users, however for simplicity we constrain our discussion to the mode where the login challenge is displayed across a sequence of grids, and where one key image is certain to appear in each grid. Image filtering is the process of reducing an image set into a login challenge through a process of choosing key images and their associated decoy images.

The absence of an accepted automated process to perform image filtering has likely, in part, motivated recent research pursuing the identification of an optimal image type for recognition-based graphical passwords [15] (see Table 1). This optimal image type is intuitively one that minimizes the burden placed upon a person to undergo the process of image filtering by hand, and one that allows users to perform favorably in recall tests with the login challenges assembled using that process. The drive to satisfy both constraints has led to a focus upon particularly contrived image types that by design permit a narrow range of possible interpretations as to their content (e.g. clipart). The lack of explicit attention given to image filtering even in these contexts appears to assume that it is a one-off procedure, and that the resulting login challenge can be reused for each user of the system. However, likely realities of deployment might make the use of a finite image set unrealistic. For instance, inevitable password resets would mean that previously seen images must be discarded from the image set for a particular user. In addition, if the image set or login challenge is static between users, then attackers can build up knowledge regarding user behavior with those images e.g. user choice, can permit phishing, and spontaneous distribution of credentials e.g. password sharing, observation attack (due to the images providing a common frame of reference shared between users). This approach also takes little account of results that have suggested users have better memory retention for images they have created [28], nor context-specific defenses that result from strategic selection of image content [9].

There is a general lack of knowledge regarding the impact of strategies of image filtering upon usability and security. The assumption so far has been that images should all be semantically and visually different for purposes of usability. A different assumption so far has been that images should all be semantically distinct so that Passfaces cannot be described by gender or obvious characteristics”[25] pg. 5. This illustrates sensitivity to risks of large semantic differences in the login challenge and the ability for users to share the graphical passwords. Usability concerns must result, however one study of human memory involving 2500 images presented in pairs showed that participants could be accurate at remembering precise image details. Even where images were visually and semantically identical and exhibited only small differences in detail, e.g. orientation, user recognition rates were only marginally worse than when images exhibited semantic differences [3]. The assembly of a login challenge based upon distinct semantics is perceived to improve usability, but those assembled to incorporate similar semantics could be harnessed to improve security.

In either case, while the curation of a usable and secure login challenge remains a skill residing with those with the greatest experience of doing it, the propagation of such systems more generally is limited. The spontaneous use of everyday uncurated image collections (e.g. photographs) in this context is perceived to be particularly challenging, however, this image type is in some ways attractive, as sets of uncurated images are ubiquitous in personal collections and online. It is possible that if methods of automated image filtering based upon judicious analysis of image content are identified, this could reduce the imperative to identify an optimal image type.

3.1 Image Similarity in the Login Challenge

There is currently little convention to follow regarding where to apply systematic analysis of image similarity. Figure 2 outlines points in a typical recognition-based graphical password login challenge that could comprise the image filtering procedure. Analysis can occur on a per-grid and a per-login basis. On a per-grid basis, intra grid key-decoy similarity refers to the similarity between a key image and collocated decoy images. The most usable visual search is one where decoy images appear distinct from the key image [7]. High similarity in this dimension suggests that users might confuse the key image with a collocated decoy image, whereas low similarity suggests the key image would appear to be easier to identify amongst the decoy images. Intra grid decoy similarity refers to the difference between collocated decoy images. In isolation such consideration provides few

---

Table 1: Illustrative results from graphical password studies focused upon different image types.

<table>
<thead>
<tr>
<th>System</th>
<th>Stimulus</th>
<th>Entropy</th>
<th>Success %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Komanduri [18]</td>
<td>Icons</td>
<td>35 bits</td>
<td>100%</td>
</tr>
<tr>
<td>Passfaces [4]</td>
<td>Faces</td>
<td>12.7 bits</td>
<td>85%</td>
</tr>
<tr>
<td>Tullis &amp; Tedesco [38]</td>
<td>Photos</td>
<td>20 bits</td>
<td>100%</td>
</tr>
<tr>
<td>Dynahand [27]</td>
<td>Scribbles</td>
<td>9.5 bits</td>
<td>99.4%</td>
</tr>
<tr>
<td>Weinshall [40]</td>
<td>Clipart</td>
<td>47 bits</td>
<td>&gt;90%</td>
</tr>
<tr>
<td>Déjà vu [5]</td>
<td>Fractals</td>
<td>16 bits</td>
<td>90%</td>
</tr>
</tbody>
</table>

---

Figure 2: Points at which to consider image similarity across an example login challenge. Red indicates a per grid consideration, and blue indicates a per login consideration.
usability issues, however high intra grid decoy-decoy similarity and high intra grid key-decoy similarity indicates that a grid may overall appear visually similar, which could complicate usability, observation attack and description [9].

The per-grid consideration should be complemented with per-login analysis. Inter grid key similarity refers to the similarity between key images. If there is high similarity in this dimension there is a threat that an attacker might infer that pattern (e.g. all key images are of specific objects or contain particular colors). If this difference is too great, it is likely that images will be more difficult to remember for the user who must remember visually and semantically disconnected images. Inter grid key-decoy similarity refers to the similarity between a decoy image and non-collocated key images. This is important from a usability perspective, as a decoy image may appear to be similar to a non-collocated key image and entice users to select it erroneously. Inter grid decoy similarity considers the similarity of decoy images across a whole login challenge. High similarity in this regard, along with high intra grid decoy similarity, indicates that decoys across the whole login could appear visually similar, whereas high inter grid decoy similarity and low intra grid decoy similarity indicates that there exists similarity within the decoys in each grid, however each grid appears visually different.

4. USER STUDY 1 – HUMAN CONSENSUS OF IMAGE SIMILARITY

Perceptions of image similarity are subjective. However, in order to measure the success of a proposed image processing intervention, it is necessary to first obtain some ground truth notion of pairwise similarity that exists within a particular image set. To do this we carried out a user study to capture a human consensus of similarity within an image set to provide the basis for further study.

4.1 Procedure

We assembled a set of 101 digital photographs and recruited 20 participants (14 male, 6 female with ages μ=27, σ=6) who were staff and students in the research lab. Each participant was asked to organize the printed set of 101 images into piles on a tabletop according to a similarity ranking method proposed elsewhere [35]. This involved the participant being asked to organize the set of images into piles, with the only criteria being that those perceived to be similar should be placed in the same pile. No further advice is offered. The raw data per-participant were the image numbers present within each pile. Across all participants this was aggregated into a score \( n \) for each image pair \((x,y)\), where \((x,y)=n\) means that image \( x \) and \( y \) appeared on the same pile \( n \) times, where \( n \leq 20 \), and high values of \( n \) indicate high agreement of similarity. The set of 101 images was intended to be representative of a typical photograph collection. The size of the image set was chosen to provide a manageable sorting task for participants. The images were printed onto high quality paper (100mm x 80mm) and the reverse of each was numbered. For descriptive purposes only we labeled the images according to the following informal categories: People (9): focus is a person or group of individuals; Scene (30): the focus is purely a landscape scene; Object (14): the focus is purely an object; People/Scene (47): the focus is upon people and scenery; People/Object (1): the focus is upon both people and an object. The image collection contained images taken to a wide range of photographic quality, and was sourced by aggregating a number of personal collections.

4.2 Results

Figure 3 gives an overview of the raw output for this study, which highlights the subjective nature of image similarity judgments even across a relatively small image set. For each image, the graph illustrates the number of other images considered to be a strong match for similarity. For a pairing to be considered a strong match, we applied a threshold to \( n \) that represented the minimum number of times images should have been placed on the same pile. As we increase the threshold, fewer images are classed as a strong match. The median number of piles participants sorted the images into was 21.5 (IQR = 12.25) with a minimum of 6 piles and a maximum of 32 piles. Images in the people and scene categories generally had the highest number of image matches (Median=45, IQR=13) and those in the Object category had the least (Median=33; IQR=23). No systematic investigation of the strategies used to group images was conducted, but in general it was apparent that these ranged from matching particular objects in the image, to matching the overall context, contrast level or principal colors. Although we only use 101 images in the results, the graph shows 102 (the original number) since one image was misplaced in the course of the study (#84).

4.2.1 Choosing an Automated Similarity Measure

A final phase of this study was to test a number of image processing methods to identify one that was most appropriate to detect the most severe instances of similarity as identified in the sorting task. The threshold of \( n \geq 14 \) (that is: in our first study, fourteen or more participants judged two images as similar) provided a basis for us to identify the most severe cases of similarity that any automated mechanism should detect. The field of Content-based Image Retrieval is fast moving; our approach was to test a number of candidate image signatures that would not require extensive expertise in image processing to understand and
implement. We reused the images from the first study and performed analysis with those images in the CIE(L*a*b*) color space [11] which is more perceptually linear than RGB or HSV. We implemented each of the following in OpenCV:

- **Statistical Moments**: treat each channel of a digital image as a probability distribution and calculate the first three statistical moments. To compare two image signatures we calculated the Euclidean distance between the statistical moments of each color channel and threshold the result.

- **Color Histogram**: the histogram bins contain the frequencies of particular pixel values. Firstly we initialize a histogram with 16x16x16x16 bins, which divides each 8 bit color channel into 16 bins. In the normalization phase, each bin is set to a value between zero and one representing its relative frequency with regard to the other bins. Then we remove any bins with less than 1% of the volume as this can be attributed to noise. To compare histograms we calculate the (EMD) [30] which treats the histograms as *piles* and provides the minimum cost of turning one pile into the other. The threshold for similarity was 0.9.

- **PerceptualDiff** [42]: this is not an image signature but is a suite of algorithms that contains a model of the human visual system. Its canonical task is to optimize the computer graphics task of global illumination, by determining whether two scenes are perceptually similar. We were interested to see if a more sophisticated approach held promise.

For each method we made a single pass of the digital images from the sorting study where each was resized to 384x286. We took each image in turn, calculated the corresponding image signatures and compared to every other image in the set, noting the images that were judged to be similar in each case. To calculate the success of these routines, we employed widely used metrics for information retrieval: recall and precision:

\[
\text{Recall} = \frac{|\text{relevant images} \cap |\text{retrieved images}|}{|\text{relevant images}|}
\]
\[
\text{Precision} = \frac{|\text{relevant images} \cap |\text{retrieved images}|}{|\text{retrieved images}|}
\]

We had knowledge of relevant images from the first user study in the form of the strong matches identified by participants for each image. Retrieved images are the set of images that the particular method judged to be similar. The metric of recall provides a measure of the fraction of relevant images that a particular method returned. The precision provides the fraction of the returned images that were relevant and is sensitive to false positives. Where thresholds had to be chosen to make a decision of similarity for a particular image processing intervention, they were selected to balance precision and recall. To incorporate spatial information into the calculations we also augmented the statistical moment and color histogram methods with a vertical or horizontal region of interest (ROI). This involved partitioning images with a vertical or horizontal line, calculating the image signature for both halves and using the mean of the two as the result for that image. Table 2 summarizes the filtering results obtained for each method. In addition to precision and recall we calculated the F₁ score, which is used to aggregate both precision and recall and represents a weighted average of the two.

Overall, the color histogram image signature applied to whole images provided the best recall at .58. The addition of spatial information to the image signature through the vertical ROI gave higher recall than the horizontal ROI but also introduced more false positives. The use of statistical moments was less effective than the color histogram in all configurations, as recall was .34, and this in fact dropped with the introduction of ROI, although ROI eliminated false positives. PerceptualDiff produced a lower recall than both color histogram and statistical moments. The use of PerceptualDiff was most effective at returning very strict matches where visually the objects and colors in the scene appeared similar. As might have been anticipated, the recall using the ROI was consistently lower than signatures based upon whole images, but ROI augmentation also yielded fewer false positives. This was reflected in a high score for precision.

Since the color histogram approach provided the best recall and the highest F₁ score, we chose to use this in our second study. The efficacy of the color histogram is likely because that representation captured the diversity of color without being restrictive spatially. This method did not provide a perfect recall score; however, we believed this score was difficult to better given the set of images in use.

### Table 2: Results from filtering procedure on an image set with 800 photographs, resized to 384x286.

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall</th>
<th>Precision</th>
<th>F₁</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color Histogram</td>
<td>.58</td>
<td>.95</td>
<td>.4</td>
</tr>
<tr>
<td>Color Histogram &amp; Vertical ROI</td>
<td>.48</td>
<td>.8</td>
<td>.3</td>
</tr>
<tr>
<td>Color Histogram &amp; Horizontal ROI</td>
<td>.41</td>
<td>1</td>
<td>.3</td>
</tr>
<tr>
<td>Statistical Moments</td>
<td>.34</td>
<td>1</td>
<td>.3</td>
</tr>
<tr>
<td>Statistical Moments &amp; Vertical ROI</td>
<td>.20</td>
<td>1</td>
<td>.2</td>
</tr>
<tr>
<td>Statistical Moments&amp; Horizontal ROI</td>
<td>.07</td>
<td>1</td>
<td>.1</td>
</tr>
<tr>
<td>PerceptualDiff [42]</td>
<td>.24</td>
<td>.9</td>
<td>.2</td>
</tr>
</tbody>
</table>

5. **USER STUDY 2 – RECALL TEST USING AUTOMATICALLY SELECTED IMAGES**

The first user study suggested that the optimal image signature we tested was the color histogram, as it provided the closest predictor of the human similarity judgments we collected. Our remaining research question concerned whether systematic manipulation of thresholds chosen for this image signature could have a predictable impact upon the short-term recall of users in a typical graphical password login.

### 5.1 Procedure

We chose a between-subject study design where the independent variable was the similarity between a key image and its decoy images, and the dependent variables were user performance in terms of recall and login time. We developed a web-based system that would challenge the user to identify four key images across four grids of nine images in a 3x3 layout, with one key image certain to appear in each grid, providing theoretical entropy of 12.7 bits. We chose three experimental conditions where similarity between the key image and its decoy images was controlled by a threshold upon the EMD distance $d$ between the color histograms of the images:

- **Similar**: where $1<d<=0$
- **Middle**: where $4<d<=3$
- **Dissimilar**: where $6<d<=5$

Studies in psychology [7] have observed how the difficulty of the visual search should decrease with decreasing similarity between
target and non-targets. We were hoping to recreate a similar trend. To generalize our results more effectively we firstly discarded the image set used in the first study and obtained a set of 1000 images used in other image processing research [39], and removed any portrait oriented images for display consistency (reducing to 800). The database is highly categorical, which provides a worst case scenario for this study. In advance, we also chose 8 key images that represented exemplars of particular categories in the image set (see Figure 4). These key images were persistent across conditions. For each key image and experimental condition we automatically chose eight decoy images according to the condition-specific similarity criteria. We also enforced distances between other images in the login to respect the image filtering concerns discussed in Section 3.1. Within a particular condition, once an image was selected as a decoy to be associated with a particular key image, it could not be selected to appear as a decoy image for a different key image within that condition. Also, a key image could not reappear as a decoy image. Figure 5 provides an example of decoy image selection for one particular key image across all three conditions.

We recruited participants from the crowdsourcing platform Amazon Mechanical Turk. Kittur et al. [16] provide hindsight from conducting users studies on this platform, in particular that the most suitable tasks are those that have a verifiable answer. Clearly those carrying out studies on crowdsourcing platforms must design robust experiments that do not rely on literacy, and due to its remote nature take measures to detect behavior that may undermine the integrity of the study. The user sample on Mechanical Turk was suitable as they are likely to be technology savvy adults. There were a number of study phases: registration: participants were requested to give information such as worker ID and demographic information; enrolment: where the participant would be given 30 seconds to view four key images randomly selected from our set of eight; wait: A JavaScript enforced stoppage of 30 minutes where participants could not progress to the next phase, but were free to carry out other tasks on Mechanical Turk. If participants attempted to progress beyond the wait period too quickly this could be detected via use of server-side timestamps. The last phase was recall: the participant attempts to recognize the images assigned to them and has a single attempt to do so.

Due to the remote nature of the study we designed the following study defenses in order to have increased confidence in our results: anti-image caching: the images presented at login were drawn from a different location on the server than those at enrolment. This removed the threat that the key images would load faster due to caching; anti-print screen: participation was restricted to Internet Explorer and via a JavaScript we cleared the clipboard of the participant every 100 milliseconds. If consent to do this was not granted the experiment would not continue; Dynamic key images: key image sequences were not static across participants and there were 3C4 different possibilities for the sequence of key images that could be presented. This meant that if images were recorded they may not be immediately reusable by another participant. HTTP GET parameter protection ensured that we could detect where parameters were maliciously altered in the browser or the back button was pressed.

5.2 Results

5.2.1 Participation

We received 364 completed logins across a 6 day period. We treated as outliers those who completed the login procedure identifying no key images in less than 5 seconds. This reduced the numbers down to 343 with 117 in the similar condition, 112 in the dissimilar condition and 114 in the middle condition. In terms of demographics, 72% of participants were from India, with the United States the next prominent location at 7%. Most of the participants were male (73%). In terms of age, 269 were in the age group 18-30, 67 in the group 31-40, 15 in the age group 41-50, and 13 were 51+ years of age.

Table 3: The number of key images correctly identified (out of four) in study two. Success is 4/4.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Score Distribution</th>
<th>Success</th>
</tr>
</thead>
<tbody>
<tr>
<td>Similar (n=117)</td>
<td>0 (0%) 14 (12%) 26 (22%) 37 (32%) 34 (29%)</td>
<td></td>
</tr>
<tr>
<td>Middle (n=114)</td>
<td>5 (4%) 8 (7%) 14 (12%) 20 (18%) 67 (59%)</td>
<td></td>
</tr>
<tr>
<td>Dissimilar (n=112)</td>
<td>0 (0%) 6 (5%) 6 (5%) 8 (7%) 74 (70%)</td>
<td></td>
</tr>
</tbody>
</table>

5.2.2 Accuracy

We firstly calculated a login success rate on a per-participant basis i.e. to compare the participants who correctly identified all 4 images. This was calculated by (successful logins/total logins). The raw data comprised success/fail value to represent a login. There was a significant difference between the performance of participants in the dissimilar group (70%) and the similar group (29%) \( \chi^2(1,N=229)=37.716, p<0.01 \). In addition there was a significant difference between the success rate in the middle (59%) and similar condition \( \chi^2(1,N=231)=20.716, p<0.01 \). The difference between the dissimilar and the middle condition was not statistically significant. Table 3 presents a more detailed illustration of participant performance. We also calculated a per-click success rate that represented (correct clicks/total clicks) for each condition. The benefit of this calculation is that it can give insight into accuracy in a manner less sensitive to a single mistake.
Figure 6: The number of login errors made per key image and per experimental condition.

by a participant. For example, a single problematic image grid could reduce login success rates considerably, whereas in reality this would reduce the per click success rate less severely. There was a significant difference between the success rates in the dissimilar condition (90%) and the similar condition (67%) \( X^2(1, N=916)=57.679, p<0.01 \). There was also a significant difference between the success in the middle (80%) and the similar condition, \( X^2(1, N=924)=19.758, p<0.01 \). The difference between the dissimilar and the middle condition using this metric was not significant.

Figure 6 illustrates the number of errors made per condition per image. Overall the errors do follow an intuitive pattern, they increase as the decoy images become more similar to the key image. However there are two exceptions, image 3 and image 7. Looking closely at the grids for these images, it is likely these errors can be explained by inter grid key-decoy similarity (see Section 3.1): a decoy image was visually similar to a non-collocated key image. This likely created confusion as to which image the user should select. This could indicate that the threshold we imposed on this instance of similarity was not sufficiently high. The graph also illustrates the interesting case of image 6 in the similar condition: there was a large number of user errors recorded when they were asked to identify this image. The decoy images for this image appeared visually and semantically similar. Analysis of errors made on a per-image basis across conditions highlighted a number of significant results too (see Table 4).

5.2.3 Login Duration

We also recorded time required for users to login in each condition. This was recorded from the first grid appearing on-screen, until the final click. We treated the data as non-parametric due to the existence of a number of particularly long login durations distorting the mean. The median login duration was 57 seconds in the similar group, in the middle group 40 seconds, and in the dissimilar group 36 seconds. The difference between the similar and dissimilar conditions was significant in a Mann-Whitney U test \( Z=-4.730, p < 0.01 \). Using a Wilcoxon 1-sample sign test we estimated the 95% confidence interval for the medians. This estimates that participants in the dissimilar condition would take between 33-40 seconds, in the middle condition 38-51 seconds, and in the similar condition 48-68 seconds. This suggests that the choice of decoy selection method could also have a significant impact upon the login durations. Figure 7 shows the distribution of login durations recorded for successful logins for each condition.

Table 4: Significant differences noted in user performance across experiment conditions on a per-image basis.

<table>
<thead>
<tr>
<th>Image</th>
<th>Success %</th>
<th>Success %</th>
<th>( X^2 )</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Similar</td>
<td>Dissimilar</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>71%</td>
<td>91%</td>
<td>( X^2(1,120)=5.729, p&lt;0.05 )</td>
</tr>
<tr>
<td>2</td>
<td>66%</td>
<td>95%</td>
<td>( X^2(1,119)=14.540, p&lt;0.01 )</td>
</tr>
<tr>
<td>5</td>
<td>55%</td>
<td>87%</td>
<td>( X^2(1,105)=14.207, p&lt;0.01 )</td>
</tr>
<tr>
<td>6</td>
<td>32%</td>
<td>93%</td>
<td>( X^2(1,117)=46.230, p&lt;0.01 )</td>
</tr>
<tr>
<td>7</td>
<td>77%</td>
<td>95%</td>
<td>( X^2(1,81)=31.777, p&lt;0.01 )</td>
</tr>
<tr>
<td></td>
<td>Similar</td>
<td>Middle</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>55%</td>
<td>83%</td>
<td>( X^2(1,116)=10.449, p&lt;0.01 )</td>
</tr>
<tr>
<td>6</td>
<td>32%</td>
<td>70%</td>
<td>( X^2(1,116)=16.726, p&lt;0.01 )</td>
</tr>
<tr>
<td></td>
<td>Middle</td>
<td>Dissimilar</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>74%</td>
<td>95%</td>
<td>( X^2(1,101)=8.614, p&lt;0.01 )</td>
</tr>
<tr>
<td>6</td>
<td>70%</td>
<td>92%</td>
<td>( X^2(1,115)=10.125, p&lt;0.01 )</td>
</tr>
<tr>
<td>7</td>
<td>73%</td>
<td>95%</td>
<td>( X^2(1,123)=10.908, p&lt;0.01 )</td>
</tr>
</tbody>
</table>

Figure 7: Length of successful logins in each condition: top left) similar; top right) middle; bottom left) dissimilar; bottom right) overall.

6. SECURITY IMPLICATIONS

The introduction of deliberate and measurable differences in visual similarity between images creates the potential for traces of this process to be left behind and exploited by attackers, who could infer key images and gain unauthorized access to systems. The particular threat is guessability, that patterns in the composition of the login challenge could be reverse engineered to allow an attacker to make better than random guesses. The goal
for an attacker is to obtain a successful login without any interaction with the user through activities such as coercion or observation attack. Intersection attack would allow a similar attack vector if unsecured [5][8]. We should assume that an attacker knows the method used for image filtering and any thresholds employed, and is able to capture the login challenge specific to a particular user. If the system provides direct authentication [33] we assume the attacker has compromised the username of the legitimate user and can capture the login images for offline analysis. If the infrastructure is local authentication then an attacker may be able to take a high quality photograph of the images, although this could be considered less likely. Other threats to be considered when introducing visual differences in the image filtering procedure include observation attack and description [9].

6.1 Key Relative Filtering

Key relative filtering [27] has been proposed in previous work as a method that is suitable for small image sets, as it imposes few constraints upon the login challenge composition. The procedure is to firstly identify the key image, reject all images within a similarity distance \( d \) of the key image and select decoy images randomly from the remaining images. However, an attacker could identify candidate key images even without knowledge of the thresholds being used, by recalculating pairwise similarities to search for patterns. One way to do this is to create a similarity matrix (see Figure 8) which is a simple visualization that captures pairwise EMD distances between images in a single \( nxn \) grid. In Figure 8 each large square represents the location of a single image, and the smaller squares within contain the EMD distance between that image and every other image in the \( 3x3 \) grid. The figure represents a particularly vulnerable case where the key relative similarity threshold is \( d<3 \). In this case the attacker could conclude that the centermost image has a good chance of being the key image, since it is the only image that exhibits such a careful pattern (\( d>3 \)) in pairwise similarity values. Even without knowledge of the threshold the attacker could make a good guess at its value based upon the minimal \( d \) observed for each image in the matrix.

6.2 Exhaustive filtering

The analysis of key relative filtering has shown that for purposes of security a more holistic approach to image filtering should be taken, in order to hide traces of the filtering procedure. One approach is based upon ensuring a minimum distance exists between all images in the grid. One limitation of this approach is that while it enforces a minimal distance between all images, there is no upper bound, which could leave the login vulnerable to observation attack, as images exhibiting large visual differences could remain in the login challenge. An alternative approach that could eliminate this threat is based upon similarity intervals, where additionally an upper bound of similarity is also enforced. However, this approach would likely be difficult to implement in small image collections, as a greater number of images are likely to be rejected due to the increased number of similarity constraints upon a permissible image. There is a trade-off between the volume of images rejected in the filtering procedure and the number of constraints that are enforced. As a compromise, a minimum distance approach is likely to be suitable in smaller image sets and where observation attack or description is less likely to be a threat. An example of the visual differences that may result is illustrated in Figure 9. In order to minimize the number of images that must be rejected, a useful strategy in general involves:

1. Choosing a strategy for decoy selection i.e. similarity or dissimilarity.
2. Choosing a candidate key image, and calculating the distribution of pairwise similarity between it and the rest of the image set.
3. Sorting the images in ascending order of EMD, then, if choosing for dissimilarity, choosing from the back of the list, and if choosing for similarity, choosing from the front of the list.
4. Repeating 2-4 for each key image.

<table>
<thead>
<tr>
<th>-</th>
<th>1.8</th>
<th>3.1</th>
<th>1.8</th>
<th>2.3</th>
<th>3.1</th>
<th>2.3</th>
<th>-</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.0</td>
<td>3.2</td>
<td>0.8</td>
<td>4.0</td>
<td>3.1</td>
<td>5.0</td>
<td>2.0</td>
<td>3.3</td>
</tr>
<tr>
<td>1.3</td>
<td>3.0</td>
<td>6.0</td>
<td>2.9</td>
<td>4.9</td>
<td>3.8</td>
<td>4.0</td>
<td>5.0</td>
</tr>
<tr>
<td>4.0</td>
<td>4.0</td>
<td>2.0</td>
<td>3.2</td>
<td>3.1</td>
<td>3.3</td>
<td>0.8</td>
<td>5.0</td>
</tr>
<tr>
<td>-</td>
<td>3.5</td>
<td>3.0</td>
<td>3.5</td>
<td>-</td>
<td>3.0</td>
<td>3.0</td>
<td>-</td>
</tr>
<tr>
<td>0.6</td>
<td>4.1</td>
<td>1.4</td>
<td>3.2</td>
<td>3.4</td>
<td>4.0</td>
<td>3.0</td>
<td>1.9</td>
</tr>
<tr>
<td>1.3</td>
<td>2.9</td>
<td>4.0</td>
<td>3.0</td>
<td>4.9</td>
<td>5.0</td>
<td>6.0</td>
<td>3.8</td>
</tr>
<tr>
<td>0.6</td>
<td>3.2</td>
<td>3.0</td>
<td>4.1</td>
<td>3.4</td>
<td>1.9</td>
<td>1.4</td>
<td>4.0</td>
</tr>
<tr>
<td>-</td>
<td>4.0</td>
<td>5.0</td>
<td>4.0</td>
<td>-</td>
<td>4.0</td>
<td>5.0</td>
<td>4.0</td>
</tr>
</tbody>
</table>

Figure 8: Similarity matrix that illustrates the pairwise EMD distance \( d \) between images in a single \( 3x3 \) image grid. This grid has been assembled with key relative filtering [27].

7. DISCUSSION

The vision of this research is that a system can take an arbitrary set of images and perform a filtering operation to generate a usable and secure login challenge, or else conclude that an image set does not contain suitable images for this purpose. Such a spontaneous approach to the generation of a login challenge becomes more useful when considering deployment level phenomena such as password resets, where ineffective recycling of images could cause confusion between new key images and old. A perfect automated semantic separation of images appears to be a difficult goal; however, we have shown that taking a coarse grained approach to the problem can affect usability. As a result it seems possible that ensuring a specific visual difference between images using pixel-level image signatures could assist in automated image selection strategies for recognition-based graphical passwords.

The recall test results suggest that comparison of pixel-level image signatures can affect the usability of recognition-based graphical passwords in terms of both user accuracy and the time required to login. We observed significant accuracy results between the similar and dissimilar condition, and the similar and middle condition. We did not observe significant differences between the middle and dissimilar condition. The login durations were also significantly impacted between the similar and dissimilar group with a significant difference of 21 seconds in the medians. The most damaging type of similarity we noted was inter grid key-decoy similarity, which is the similarity between a decoy image and a non-collocated key image. In this case the user erroneously selects a decoy image that appears similar to a non
Figure 9: Left) grids assembled using minimum distance approach where d=2; right) similarity intervals where 4>d>0.

collocated key image. Particularly conservative thresholds should be employed when considering this type of similarity. The remote scenario places success rates in a realistic zone, as participants were not within the sphere of influence of experimenters. Such results have important implications for deployment of recognition-based graphical passwords, as they serve to highlight the impact that seemingly subtle image choices can have upon the usability of the system.

We also identified the security risks inherent in automated image filtering where this is not performed holistically. The approach to exhaustive filtering based upon similarity intervals appears to be the most secure approach, resistant to reverse engineering of the similarity procedure and likely to be useful against observation attacks. However, the benefits must be traded off against the size of the required image set. Questions of an optimal image choice appear to be a pressing question for this genre of graphical passwords. However, in order to determine any benefits had by one image type or the other, it is necessary to carry out longitudinal studies that compare image types with differing levels of entropy and similarity involved. Recent work has compared objects and faces and questioned the superiority of face images [15].

The results have implications for graphical password systems of all genres. This work has focused upon recognition-based graphical passwords where there are multiple grids and one key image on each screen. However, our taxonomy of image filtering is relevant to systems where images are static for all users, or for configurations where key images are randomly distributed across the grids [5,13]. The results also have implications for other systems such as Passpoints [41], where future research could focus upon some notion of similarity between images to predict whether similar user choices could be expected between a number of images. Recent work correctly asserts that alternatives to alphanumeric passwords have so far failed to make an impact [14], juxtaposing novel mechanisms with alphanumeric passwords or PINs in terms of flexibility and convenience usually results in failure. The challenge for alternative authentication solutions is to demonstrate evidence of new types of value that impact the trade-off between security, usability, deployability, and convenience. In recent work [23] we made initial investigations how image similarity could impact observation attack and description in the context of a novel recognition-based graphical password system. Future work exploring automated image choice could help graphical passwords to provide such added value.

8. STUDY LIMITATIONS
In this study we did not consider the longitudinal memory impact of the recall task; we chose to model a short-term memory task as password enrolment is a particularly traumatic period for committing new credentials to memory. However, we believe our scenario introduced sufficient stress into the enrolment procedure to enable us to have confidence in the results. The success rates are constrained by the fact that users were not working with their own images and only had one attempt to identify the images. In addition, the participants were not logging into a real system and so not authenticating to access anything of value. Finally, the image processing intervention we chose only operated at the pixel-level. Study of more sophisticated techniques that incorporate object segmentation may prove to be a fruitful future research direction.

9. CONCLUSION
In this paper we explored the extent to which the process of assembling a usable login challenge for recognition-based graphical passwords based upon photographs could be automated using image processing techniques. In our tests we found that using a color histogram as an image signature and computing the distance between those histograms using the Earth Mover’s Distance [30] provided a useful approach. In a short-term recall test with more than 300 people using Amazon Mechanical Turk, we found that automated choice of decoys to differing levels of visual similarity could impact the number of errors that users made at login, along with the login durations. We found significantly fewer errors made by users viewing grids with progressively dissimilar decoys compared to those viewing the most similar decoys. In the most significant case, we found that our automated decoy selection method could affect login success rates by 40%. This study illustrates that the performance benefits of recognition-based graphical passwords can be closely related to the chosen image sets.

10. ACKNOWLEDGMENTS
This work was funded by the UK Digital Economy Research Hub SiDE: Social Inclusion through the Digital Economy (EP/G066019/1).

11. REFERENCES
34. Smith, T.A., Jones, L.V., and Thomas, S. Effects upon verbal learning of stimulus similarity, number of stimuli per response, and concept formation. *Journal of Verbal Learning and Verbal Behavior* 1, 6 (1963), 470-476.