Developing and Testing a Visual Hash Scheme

M. M. Olembo, T. Kilian, S. Stockhardt, A. Hülsing and M. Volkamer

Center for Advanced Security Research Darmstadt
Technische Universität Darmstadt
e-mail : {maina.olembo, timo.kilian, simon.stockhardt, andreas.huelsing,
melanie.volkamer}@cased.de

Abstract

Users find comparing long meaningless strings of alphanumeric characters difficult, yet they have to carry out this task when comparing cryptographic hash values for https certificates and PGP keys, or in the context of electronic voting. Visual hashes - where users compare images rather than strings - have been proposed as an alternative. With the visual hashes available in literature, however, people are unable to sufficiently distinguish more than 30 bits. Obviously, this does not provide adequate security against collision attacks. Our goal is to improve the situation: a visual hash scheme was developed, evaluated through pilot user studies and improved iteratively, leading to CLPS, which encodes 60 distinguishable bits using Colours, Patterns and Shapes. In the final user study, participants attained an average accuracy rate of 97% when comparing two visual hash images, with one placed above the other. CLPS was further tested in two follow-up studies, simulating https certificate validation and verifying in remote electronic voting. The results of this work and their implications for practical applications of visual hash schemes are discussed.

Keywords

Visual hash, Usability, Security, Hash function

1. Introduction

Cryptographic hash functions are widely used to guarantee integrity and provide authentication on the Internet. Different use-cases are available including verifying the authenticity of https certificates and PGP encryption keys as well as verifying the proper behaviour of an electronic voting system. In most use-cases, it is necessary to compare two hash values with each other; one is presented on the screen and the other one is available on paper. For example, in the case of https certificates, some certificate owners (like banks) distribute the hash values of their certificates (also called fingerprints) in print media to their clients. If the clients visit the corresponding webpage they can compare the printed fingerprint with the one displayed by the web browser. In many verifiable electronic voting systems, voters are asked to write down the hash value of their encrypted vote in order to verify the integrity of the voting software by later (in the vote casting process) comparing this hash value to a displayed one. In all the use-cases, hash values are represented by long strings (the length depends on the hash function and the encoding applied to the hash value). As a result, users are asked to compare long strings that hold little meaning to them. Consequently, they are not very likely to perform this task which decreases the security of the applications dramatically. In addition, users are known to be poor at this task (Perrig and Song, 1999).

Visual hashes offer an alternative, with studies as early back as Shepard (1967) showing that people perform better at interacting with images compared to text. With existing schemes proposed in literature, people were unable to sufficiently distinguish more than 30 bits. However visual hash schemes need to encode more bits to provide adequate security against collision attacks. Our objective is to improve the situation by developing a visual hash scheme where more bits can be distinguished by people, i.e., that provides a higher level of entropy in practical use.

The contribution of this work - CLPS - is a visual hash scheme encoding 60 bits using Colours, Patterns and Shapes. When tested in a user study where images were placed above and below each other for comparison, the average accuracy rate on images with obvious differences (easy pairs) was 98.8% and 94.6% on images with no differences or hard-to-detect differences (hard pairs), i.e. users could sufficiently distinguish two hash values. The combined average accuracy rate for both easy and hard pairs was 97%. CLPS was further simulated in realistic scenarios and tested in two follow-up studies: in verifiability in remote electronic voting where participants achieved an accuracy rate of 73.4% on hard pairs, and in https certificate validation, where they achieved an average accuracy rate of 78.6% for hard pairs. We discuss the implications of these results for practical applications of visual hashes.

2. Related work

Visual hashes were first explored by Perrig and Song (1999) using images generated from a computer program Random Art available at (Gallery of Random Art, 2013). Random Art was initially developed to automatically generate artistic images. It takes a binary string as input from which an image is generated randomly. Since then some more visual hash schemes have been proposed and studied in literature: Flag (Ellison and Dohrmann, 2003) and T-Flag (Lin *et al.* 2009).

Hsiao *et al.* (2009) carried out an online user study of textual and all three visual hash schemes along with their own proposal called Flag Extension. The textual schemes that were tested are Base32 (Josefsson, 2006), English words (Ford *et al.* 2006), and Chinese, Japanese, and Korean characters. To enable comparison between these schemes, the entropy was set to a value between 22 and 28 bits. Easy and hard image pairs were constructed for each scheme, where the authors defined an easy pair as containing two images that were equal, or obviously different, while hard pairs contained two images with hard-to-detect differences. Participants performed the best on accuracy rates and response times for Base32, Random Art, T-Flag and Flag Extension. Results from the work by Hsiao *et al.* (2009) are shown in Table 1.

Hsiao *et al.* (2009) argue that Random Art, Flag, and T-Flag can only guarantee limited entropy as the only way to increase the number of encoded bits is to use more colours, which makes the resulting images harder to distinguish. Thus, the number of encoded bits would increase but the level of entropy would not increase in practical use. For this reason, it seems necessary to come up with a new proposal to achieve a higher level of entropy for practical use, i.e., people are able to distinguish any two images that encode two different hash values.

		Easy Pairs		Hard Pairs	
Category	**Encoded bits**	**Accuracy (%)**	**Time (s)**	**Accuracy (%)**	**Time (s)**
Base32	25	97	3.39	86	3.51
Random Art	24*	98	4.77	94	3.21
T-Flag	24	98	6.31	85	5.30
Flag Extension	24	98	3.93	88	4.02

Table 1: Average accuracy rate and response time results from Hsiao *et al.* (2009).

*Note: the authors estimated the perceptual entropy of Random Art. Here, we provide the maximum number of bits that could be encoded.

3. Scheme development and pilot studies

In this section we discuss how the visual hash scheme was developed iteratively, describe how participants were recruited and their tasks, and summarize findings from the pilot user studies.

3.1. Original visual hash idea

People are known to be good at identifying geometrical shapes, patterns and colours (Reynolds, 1972). As a result, we decided to base our proposal for a new visual hash scheme on colours, patterns and shapes. An object is therefore defined by its shape and the pattern and colour it is filled with.

A wide range of possible values for the parameters were selected through several iterative discussions between the co-authors as well as with other colleagues leading to the following selection: four patterns (2 bits), 32 shapes (5 bits), two positions (up or down – 1 bit) and four objects in one image. Additionally, we used a colour contrast analyser and selected eight colours (3 bits) that can easily be distinguished by humans, taking into account colour-blindness. This resulted in 11 bits per object and 44 bits for an image. Four characters from a Base32 alphabet (5 bits per character, leading to 20 bits in total) were added to the image to further increase the number of encoded bits. Base32 had obtained good results in the study from Hsiao *et al*. (2009). In total, we can encode 64 bits with this approach.

3.2. Evaluation of the visual hash

We evaluated and improved this approach based on lab user studies, which allowed the participants to be observed. Timing was important as response time data was the usability measure applied to evaluate the effectiveness of the visual hash scheme. As such, the lab studies were useful in ensuring that the comparison task was carried out in a reasonable amount of time. The methodology used was the same in all user studies during the development as well as for the final evaluation (see Section 4). This methodology is described and justified in this subsection.

Hard and easy pairs: For the studies, we designed easy and hard pairs of images, where an easy pair consisted of two images that were obviously different, i.e., in which many parameters changed, while a hard pair consisted of two images that were either equal or had slight differences between them, i.e., parameters were changed to values that were visually close (for example, changing 'Z' to '2'). Note that this definition differs from that given in Hsiao *et al.* (2009). In our work, equal pairs are considered as hard pairs since all, or close to all, parameters of the pairs would have to be compared by the user to determine whether or not they were equal.

Users' tasks and methodology to collect and analyse the data: A PowerPoint presentation was developed to display the image pairs on each slide, to allow participants to answer whether the images were equal or not, to store the answers given per slide, and to automatically deduce the participants' accuracy rate and response times and store this result into an Excel data sheet for analysis. Visual Basic for Applications (VBA) scripts were written for this purpose.

The first page of the presentation gave an explanation of the study and the task that the participants were to carry out. It also showed an example of the images for comparison and explained what could differ. On the second slide, participants entered demographic data, specifying their age, gender, and level of comfort with computers. They could then begin the interactive part of the study. The slides were displayed randomly; each slide contained two images, one displayed above, or next to, the other, depending on the study. A participant then had to decide if the images were identical or different and pushed a green 'tick' button to indicate that the images matched, or a red 'cross' button to indicate that the images did not match. When participants had gone through all the slides, they commented on their perception of the study on one slide of the presentation, after which their results were displayed, showing them the number of images they had correctly identified to be equal or different. Examples of the PowerPoint slides used in the pilot studies are shown in Figure 1(a) and (b).

While we collected three usability measures - effectiveness (accuracy rate), efficiency (response time), and satisfaction (users' subjective responses) – only a few participants gave subjective responses on their perception of the visual hash scheme. As a result, we only report accuracy rate and response time results. While high accuracy rates are important, they are especially important for hard pairs as they indicate the extent to which participants can successfully distinguish differences in visual hashes, showing the scheme to be useful for practical use, e.g. when collision attacks are attempted.

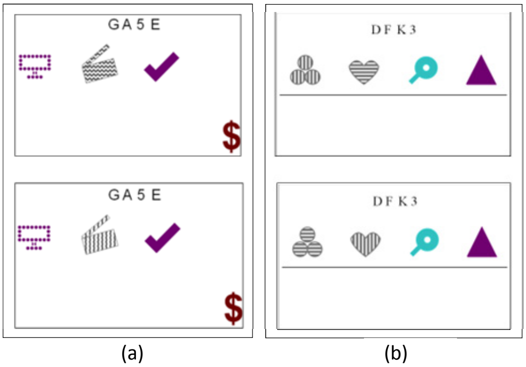


Figure 1: Image pairs in the pilot user studies: (a) original idea with images above and below each other, (b) original idea extended with horizontal partitioning line

Study Participants: They were administrative employees from a company, and students and employees from a research institute and a university. German was used for the studies, and participants were either German native speakers, or proficient in German (i.e., used it for work and study). They were verbally invited to participate in the user study. No compensation was offered. Participants were informed that their task was to compare images, and they were to indicate if what they viewed was the same or different. They were not trained on the image comparison task.

3.3. Findings of pilot user studies and scheme improvement

The original idea was tested with 16 participants (see Figure 1(a)). Eleven participants did not notice that the objects in the hard pairs moved from the ‘up’ to ‘down’ position and nine participants did not detect a change from the ‘down’ to ‘up’ position. Additionally, seven participants were unable to distinguish between dotted and horizontal-wavy patterns (not shown in the figure). In order to improve the situation we inserted a horizontal partitioning line as proposed by participants of the first pilot study. Furthermore, we replaced wavy line patterns with straight ones instead.

This new version (see Figure 1(b)) was tested with 16 new participants. Here seven participants did not distinguish the first object changing from the ‘down’ to ‘up’ position. Additionally, six participants were unable to distinguish the two centre objects switching between the ‘up’ and ‘down’ position and vice versa. As the partitioning line did not sufficiently improve the errors regarding the position parameter, we discarded the position parameter (4 bits as each object in the visual hash used one bit for position), and retained only the colours, shapes and pattern parameters for the final visual hash scheme - CLPS - encoding 60 bits (see Figure 2).

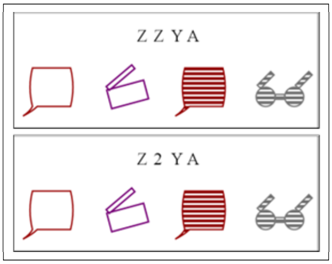


Figure 2: CLPS image pair

4. CLPS and Base32

A study was carried out to evaluate participants' accuracy rate and response time results with CLPS and Base 32, both encoding 60 bits. Participants performed well with Base32 in Hsiao *et al.* (2009), motivating inclusion of Base32 in this study. Two groups of participants were recruited. One group interacted with CLPS, and the second group with Base32. The study design was the same as for the pilot studies. The study design and results of the user study are reported in this section.

Study Design: The easy pair images in this study were obviously different image pairs, while the hard pair images were either equal or contained one or two differences in them. Ten slides had equal images, five slides had obviously different images and ten slides had slight differences in the images as follows: two slides with shapes changing, two slides with letters changing ('Z' to '2'; '5' to 'S'), two slides with colours changing (black to grey; turquoise to blue), and four slides with patterns changing. This selection was based on parameters where participants performed poorly in the pilot user studies. Image pairs were displayed randomly for every participant. There were 30 participants and the average age was 34.7 years. The youngest and oldest participants were 24 and 43 years old, respectively.

In the Base32 study, the text for comparison was 12 characters long (5 bits encoded into one Base32 character, thus, 60 bits in total). There were ten slides with equal alpha-numeric characters, five slides with obviously different ones, and ten slides with slight differences (e.g., one character changing, such as, 4NNKV4XTLPB7S and 4NNKV4XTRPB7S). Image pairs were displayed randomly for every participant. This study had 35 participants, who had an average age of 27.7 years. The youngest and oldest participants were 19 and 63 years old, respectively.

Results: Participants took slightly less time to compare CLPS images than they did to compare Base32 characters for both easy pairs (6.3s for CLPS and 6.9s for Base32), and hard pairs (4.6s for CLPS and 6.3s for Base32), showing participants to be more efficient at CLPS comparison than Base32. Additionally, the average accuracy rate for easy pairs was comparable to that of Base32 (98.8% - CLPS and 99.6% - Base32), and better than Base32 for the hard pairs (94.6% - CLPS and 89.2% - Base32). A high accuracy rate in hard pairs is important as it shows the extent to which participants can successfully distinguish differences, e.g., where collision attacks are attempted.

The results from this study show CLPS to be a viable visual hash scheme, with acceptable accuracy rate and response time results. Therefore, CLPS was tested in follow-up studies simulating realistic use: verifiability in Helios (Adida, 2008), a verifiable Internet voting system, and https certificate validation. These two studies and the accompanying results are reported in Sections 5 and 6.

5. Study simulating verifiability in Helios

This study was designed to evaluate the use of visual hashes in the Helios Internet voting system (Karayumak *et al.* 2011) to perform the so-called cast as intended verification.

Study design: The easy pair images in this study had obvious differences in the parameters, while the hard pair images were either equal or designed to have one or two differences in one parameter, i.e., colour, pattern or shape. This selection was motivated by the results of participants' performance with CLPS reported in Section 4. Thus we selected parameters that participants made errors in. Image pairs were displayed randomly for every participant.

Participants first saw a CLPS image displayed on a PowerPoint slide. They were asked a brief, distracting question on a second slide, for example, 'What is your favourite ice-cream?', and provided with multiple-choice responses. The third slide displayed to participants contained another visual hash image and participants indicated if it was similar to or different from the first one they had seen. This process simulated the Helios interface, where voters would see a hash value (first visual hash), select an option from several available options to carry out the verification process (distracting question), and then view the results of the verification, determining whether a second hash value displayed in a new window (second visual hash) matched the first one they had seen previously. Each participant repeated this process five times. Forty-five participants took part in the study. They had an average age of 26.8 years. The youngest and oldest participants were 19 and 57 years old, respectively.

Results: Participants had an average accuracy rate of 96.7% for easy pairs and took an average of 18.9s, while the accuracy rate was 73.4% for hard pairs where they took an average of 20.9s. CLPS is thus seen as a promising alternative for practical use in this use-case given the results obtained and considering the results from related work shown in Table 1.

Colour and pattern parameters proved problematic for participants, with 18 and 11 errors being made, respectively. Improvements will be investigated in future work, along with changes to the images to aid participants' recollection. We anticipate this will improve performance for practical use even further.

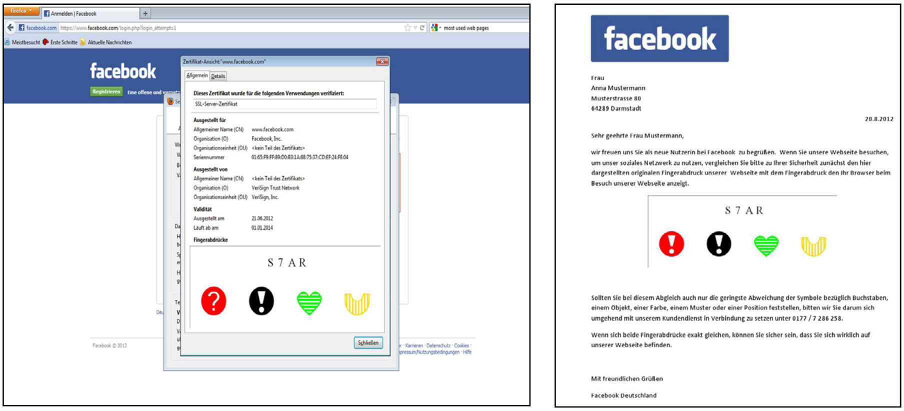
6. Study simulating https certificate validation

A study was carried out with participants comparing hash values for https certificates represented using CLPS.

In a pre-test, with 30 participants, we identified which image pairs to use for the study. As no errors were made by participants in the easy image pairs (containing obvious differences), we decided to only evaluate equal image pairs and images containing slight differences. Both of these are defined as hard pairs in Section 3.2. We therefore refer to them as equal pairs and slight-difference pairs in this section, where the study design and results of the user study are reported.

Study design: Three out of eight image pairs were equal, while the remaining five pairs were slight-difference pairs (for example, swapping one character). Participants were given eight different letters from well-known online environments, specifically online stores, social networking sites, and banks. The letters contained instructions for participants to verify the hash values of the https certificates. The hash values represented with CLPS for each website were displayed on a PowerPoint presentation, and the image pairs were displayed randomly for every participant.

As participants clicked through the presentation, they would pick up the letter fitting to the certificate on the screen and carry out the comparison. An example of a comparison simulation for Facebook is shown in Figure 3. Thirty participants took part in the study. Their average age was 38.8 years. The youngest and oldest participants were 20 and 58 years old, respectively.



Zertifikat-Ansicht:"www.facebook.com"

Dieses Zertifikat wurde für die folgenden Verwendungen verifiziert:

SSL-Server-Zertifikat

Ausgestellt für

Allgemeiner Name (CN)	www.facebook.com
Organisation (O)	Facebook, Inc.
Organisationseinheit (OU)	<kein Teil des Zertifikats>
Seriennummer	01:65:F9:FF:89:D0:B3:1A:88:75:37:CD:EF:24:FE:04

Ausgestellt von

Allgemeiner Name (CN)	<kein Teil des Zertifikats>
Organisation (O)	VeriSign Trust Network
Organisationseinheit (OU)	VeriSign, Inc.

Validität

Ausgestellt am	21.06.2012
Läuft ab am	01.01.2014

Fingerabdrücke

Frau
Anna Mustermann
Musterstrasse 80
64289 Darmstadt

20.8.2012

Sehr geehrte Frau Mustermann,

wir freuen uns Sie als neue Nutzerin bei Facebook zu begrüßen. Wenn Sie unsere Webseite besuchen, um unser soziales Netzwerk zu nutzen, vergleichen Sie bitte zu Ihrer Sicherheit zunächst den hier dargestellten originalen Fingerabdruck unserer Webseite mit dem Fingerabdruck den Ihr Browser beim Besuch unserer Webseite anzeigt.

Sollten Sie bei diesem Abgleich auch nur die geringste Abweichung der Symbole bezüglich Buchstaben, einem Objekt, einer Farbe, einem Muster oder einer Position feststellen, bitten wir Sie darum sich umgehend mit unserem Kundendienst in Verbindung zu setzen unter 0177 / 7 286 258.

Wenn sich beide Fingerabdrücke exakt gleichen, können Sie sicher sein, dass Sie sich wirklich auf unserer Webseite befinden.

Mit freundlichen Grüßen

Facebook Deutschland

Figure 3: Comparing CLPS image for Facebook

Results: Participants had an average accuracy rate of 100% and an average response time of 16.6s for equal pairs, while with slight-difference pairs, the average accuracy rate was 78.6% and the average response time was 13.7s.

Fifteen participants made errors when the line pattern in the fourth object in the image changed from vertical to horizontal. Colour was also problematic for participants, with 12 participants making errors with this parameter. The causes of these errors and possible solutions will be investigated in future work.

From these results, as well as those reported in related work (shown in Table 1), CLPS is seen as promising for further investigation in future work for this use-case. Since participants received no training in the comparison task, and as https certificates change infrequently, we anticipate that improvements in accuracy rates in practical use can be achieved, especially for hard pairs.

7. Discussion and Future Work

The practical use of CLPS and its scalability and security are discussed in this section.

7.1. Practical use of CLPS

We have shown CLPS to be a viable visual hash scheme, achieving comparable average accuracy results to those of Hsia *et al.* (2009) yet also attaining higher entropy in user studies. In applying CLPS to practical use, the results obtained in both the user study simulating https certificate validation and that simulating verifiability in Helios suggest that visual hashes are of particular relevance for both use-cases.

Since https certificates change infrequently in many cases on the Internet, users can easily notice any introduced changes to a visual hash they interact and become familiar with over a period of time. However, the visual hash value should be displayed e.g., in the first few seconds when a user visits a web page. In future work training participants on the comparison task will be investigated as we anticipate that this will improve accuracy on slight-difference pairs and speed up the comparison time. A realistic scenario will be implemented, where participants visit a web page several times and the visual hash is changed completely or slightly. As users are likely to get accustomed to this over time and perhaps no longer carry out the comparison, means of avoiding this also need to be evaluated.

In the study simulating verifiability in Helios, users had to recall images after viewing them for a short period of time, affecting their accuracy rate on hard pairs. For future work, we will explore whether using a story in the visual hash images generated in this work, will help participants better remember the images, and speed up the response time.

7.2. Scalability of CLPS

CLPS can be used to encode 60 bits and achieve the same amount of entropy. This is a first step towards a higher security level, but it is still not enough to guarantee collision resistance in practice. CLPS is however easily scalable: for this initial proposal, we used only a limited number of varying colours, patterns and shapes. We identified that colours and certain patterns are problematic for users to distinguish. The current results suggest that we can increase the number of parameters that were not problematic for users to some extent, e.g., shapes and objects, while the entropy is increased by the same amount, as long as the distinguishability by people does not shrink faster, i.e., the entropy still grows.

8. References

Adida, B. (2008), "Helios: Web-based open-audit voting", *Proceedings of the 17th Symposium on Security*, San Jose, CA, 2008, pp. 335 - 348.

Gallery of Random Art, (2013), http://www.random-art.org, (Accessed 28 March 2013)

Ellison, C., and Dohrmann, S. (2003), Public-key support for group collaboration, *ACM Transactions on Information and System Security (TISSEC)*, Vol. 6, No. 4, pp. 547 – 565.

Ford, B., Strauss, J., Lesniewski-Laas, C., Rhea, S., Kaashoeck, F., and Morris, R. (2006), "Persistent personal names for globally connected mobile devices", *Proceedings of the 7th USENIX Symposium on Operating Systems Design and Implementation (OSDI).*

Hsiao, H.-C., Lin, Y.-H., Studer, A., Studer, C., Wang, K.-H., Kikuchi, H., Perrig, A., Sun, H.-M., and Yang, B.-Y. "A study of user-friendly hash comparison schemes", *Proceedings of the 2009 Annual Computer Security Applications Conference, ACSAC '09,* Washington DC, USA, 2009, pp. 105 - 114.

Josefsson, S. (2006), *The Base16, Base32, and Base64 Data Encodings, RFC4648,* https://tools.ietf.org/html/rfc4648 (Accessed 28 March 2013)

Karayumak, F., Kauer, M., Olembo, M. M., Volk, T., & Volkamer, M. (2011), "User study of the improved Helios voting system interfaces", *2011 1st Workshop on Socio-Technical Aspects in Security and Trust (STAST),* pp. 37 - 44).

Lin, Y.-H., Studer, A., Hsiao, H.-C., McCune, J. M., Wang, K.-H., Krohn, M., Lin, P.-L., Perrig, A., Sun, H.-M., and Yang, B.-Y. (2009), "Spate: Small-group PKI-less Authenticated Trust Establishment", *Proceedings of the 7th International Conference on Mobile Systems, Applications and Services,* pp. 1 – 14.

Perrig, A., and Song, D. (1999), "Hash visualization: A new technique to improve real-world security", *Proceedings of the 1999 International Workshop on Cryptographic Techniques and E-Commerce*, pp. 131 - 138.

Reynolds, R. E., White, R. M., and Hilgendorf, R. L. (1972), "Detection and recognition of colored signal lights", *Human Factors* Vol. 14, No. 3, pp. 227 - 236.

Shepard, R. N. (1967), "Recognition memory for words, sentences and pictures", *Journal of Verbal Learnings and Verbal Behaviour*, Vol. 6, Issue 1, pp. 156 - 163.