Passquerade: Improving Error Correction of Text Passwords on Mobile Devices by using Graphic Filters for Password Masking

Mohamed Khamis
[1] University of Glasgow
Glasgow, UK
[2] LMU Munich
Munich, Germany
Mohamed.Khamis@glasgow.ac.uk

Tobias Seitz
Leonhard Mertl
Alice Nguyen
Mario Schneller
Zhe Li
LMU Munich
Germany

Figure 1: We investigate how graphical filters impact the usability and security of text passwords on mobile devices compared to displaying them in plain text or asterisks. It is difficult to mentally reverse distortions, hence it is challenging for observers to know what the text passwords above are. At the same time, if a user knows that the leftmost word is Color-Halftone, they can easily map the word's letters to the distortions. This improves error correction, while maintaining observation resistance.

ABSTRACT

Entering text passwords on mobile devices is a significant challenge. Current systems either display passwords in plain text: making them visible to bystanders, or replace characters with asterisks shortly after they are typed: making editing them harder. This work presents a novel approach to mask text passwords by distorting them using graphical filters. Distorted passwords are difficult to observe by attackers because they cannot mentally reverse the distortions. Yet passwords remain readable by their owners because humans can recognize visually distorted versions of content they saw before. We present results of an online questionnaire and a user study where we compared Color-halftone, Crystallize, Blurring, and Mosaic filters to Plain text and Asterisks when 1) entering, 2) editing, and 3) shoulder surfing one-word passwords, random character passwords, and passphrases. Rigorous analysis shows that Color-halftone and Crystallize filters significantly improve editing speed, editing accuracy and observation resistance compared to current approaches.

CCS CONCEPTS

• **Security and privacy → Authentication**; • **Human-centered computing → Human computer interaction**.

KEYWORDS

Password Masking, Usable Security, Smartphones, Authentication

CHI 2019, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5970-2/19/05…$15.00
https://doi.org/10.1145/3290605.3300916

ACM Reference Format:
Mohamed Khamis, Tobias Seitz, Leonhard Mertl, Alice Nguyen, Mario Schneller, and Zhe Li. 2019. Passquerade: Improving Error Correction of Text Passwords on Mobile Devices by using Graphic Filters for Password Masking. In *CHI Conference on Human Factors in Computing Systems Proceedings (CHI 2019), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3290605.3300916

1 INTRODUCTION

Despite the significant advancements in authentication on smartphones and tablets, text passwords continue to be one of the most widely used schemes on handheld mobile devices. They are often used when accessing services such as emails, online banking, e-shopping, and more [21].

Entering text passwords on mobile devices is not only error prone and time consuming [13, 21, 23], but also subject to shoulder surfing [4, 27]. Melicher et al. reported that users take 20% longer to enter passwords on mobile devices, and made twice as many errors [21]. We are not aware of any statistics on how often users edit passwords. Likely because current methods do not allow noticing typos, which makes correcting typos harder than reentering the password. Schaub et al. [23] collected information on edits during password entry, but they aggregated it with incorrect entries to estimate the error rate without reporting how often edits occur As for security, a field study by Eiband et al. documented cases of shoulder surfing text passwords when logging into online shopping websites where biometric authentication is still widely unsupported [4]. This underlines the need for improving both the usability and security of password entry on handheld mobile devices.

Android and iOS either use asterisks to mask the characters of the password shortly after the user enters them, or keep them clearly visible in plain text during the entire interaction. While asterisks reduce the shoulder surfing risk, they make it more challenging for users to notice and correct typos, which are common on mobile devices [13, 21]. On the other hand, showing the passwords in plain

text simplifies entering and editing them. The downside is that this exposes passwords, making them easier to observe by bystanders.

In this work, we investigate the application of graphic filters onto text passwords to overcome the aforementioned usability and security challenges. We leverage properties of filters that take advantage of the user's memory and visual perception. Namely, humans can recognize visually distorted versions of content that they have seen before [2, 6, 11]. However, the distortions remain ambiguous to people who have not seen the original content, i.e., people struggle to mentally reverse distortions [3, 9, 10, 28]. For example, it is difficult to understand the text in Figure 1 before learning that they refer to Color-Halftone, Crystallize, Gaussian Blur, and Mosaic. While these properties of human perception has been used for graphical authentication [3, 9, 10] as well as privacy protection when browsing photos [28] our work is the first to investigate its impact on text passwords for mobile devices.

Inspired by promising results in prior works, and based on the results of an online questionnaire (N=163), we chose to investigate: 1) Color-halftone, 2) Crystallize, 3) Gaussian-Blur, and 4) Mosaic filters. In a follow up user study (N=25), we compared the filters to current implementations on Android and iOS as baselines, namely, displaying passwords in 5) Plain text and replacing characters by 6) Asterisks shortly after they are typed. We evaluated the impact on one-word passwords, passwords consisting of random characters, and passphrases consisting of multiple dictionary words. To gauge usability, we measured *entry* time and accuracy, and *editing* time and accuracy. As for security, we evaluated the susceptibility of the password to shoulder surfing. While the filters did not impact entry time and accuracy, we found that the Color-halftone and Crystallize filters significantly improve both *editing accuracy* compared to the other filters and *editing time* compared to Asterisks. At the same time, both Color-halftone and Crystallize filters are significantly more resilient to shoulder surfing than the baselines and the Mosaic filter. We conclude by discussing the implications of graphic filters on the usability and security of text passwords.

This work contributes 1) the introduction of graphical filters to improve the usability and security of text passwords, 2) an understanding of the impact of filters on editing time and accuracy, and on resilience to shoulder surfing, and 3) recommendations for improving the UX of authentication UIs.

2 RELATED WORK

Our work builds on: 1) text passwords on mobile devices, and 2) security and privacy protection using obfuscation.

2.1 Text Passwords on Mobile Devices

While the research community introduced a plethora of authentication schemes for mobile devices (e.g., [16, 18, 20, 25, 29]) in this work we focus on text passwords. Although text passwords were long predicted that they will cease to exist [22], they are still widely used for on mobile apps as well as online websites and services accessed from mobile devices [26].

Several works studied text password entry and creation on mobile devices. It is generally agreed upon that entering text passwords on mobile devices is error prone [21], time consuming [26], and frustrating [13, 21]. For example, a study by von Zezschwitz et al. [26] showed that authentication is slower on mobile devices compared to desktop computers. Greene et al. [7] found that password entry on mobile devices is more error prone and takes more time. Like Yang et al. [30], they attributed this to the lack of tactile feedback and the size of virtual keyboards on mobile devices. Schaub et al. [23] highlighted that switching layouts on virtual keyboards (e.g., to enter symbols) is one reason behind increased authentication time on mobile devices. Haque et al. [8] found that users create weaker passwords on mobile devices, and proposed layouts to encourage the use of special characters and digits on mobile devices. Melicher et al. [21] found that users need 20% longer time to authenticate on mobile devices, and make twice as much errors. Jakobsson and Akavipat [13] proposed Fastwords, which employs pass phrases with flexibility to synonyms and order of words, auto-correction, and auto-completion to make entries faster and more accurate.

While the aforementioned works highlighted usability issues related to text passwords, a survey by Eiband et al. asked participants for their experiences with shoulder surfing, and concluded that text passwords are among the content that is being successfully shoulder surfed in daily situations [4]. For example, in one of the reported stories, a person observed a user sitting next to them and learned their Amazon account's username and text password.

From previous work, we learn that there is a need to improve the usability and security of text passwords on mobile devices. In particular, there is a need to improve entry time [23, 26], error correction [13, 21], and observation resistance [4] of text passwords. We hence explore alternatives to mask text passwords, and compare them to baselines from commonly used operating systems.

2.2 Security and Privacy Protection through Obfuscation

Humans excel at recognizing patterns in images and relating them to patterns that they know [19]. Interestingly, humans can recognize patterns even in distorted or low quality images. For example, humans can recognize faces they know in low quality surveillance videos [2, 11]. Humans can also recognize the content of distorted versions of images that if they had seen the original undistorted versions before [6, 9]. These abilities stem from multiple human properties that were studied extensively in previous research on visual perception and cognitive psychology [2, 6, 19].

These properties were exploited in multiple works to improve graphical passwords [3, 9, 10]. The idea was to show the user an undistorted image when creating a password, and allow users to authenticate by selecting the image from a set of distorted images. Users were easily able to distinguish the images they chose as passwords because they saw the original undistorted versions before, but observers were unable to understand the content of the images. Von Zezschwitz et al. also leveraged this property to protect privacy when browsing photo albums in public [28]. Eiband et al. distorted text messages using the user's handwriting, making it difficult for observers to read but not for legitmate users [5]. In EyeSpot [17], the phone's screen is distorted when texting or writing emails except for the area the user is gazing at. Crystallized masks were favored over blackout masks in EyeSpot because they maintained some contextual information (e.g., the user could still see who the last person to send a text message was), while still providing a relative

protection from shoulder surfing. Compared to the aforementioned works, our work is the first to explore the impact of graphic filters on text passwords on mobile devices.

3 THREAT MODEL

In our threat model, the user is entering a text password on their mobile device in a context in which they are subject to shoulder surfing. For example, text passwords are shoulder surfed in public transport, public spaces, and at work [4, 12]. The observer has a perfect view of the interface during and after authentication.

4 DESIGNING THE FILTERS

Prior work experimented with a variety of filters to obfuscate content. The most promising ones were crystallize [17, 28], mosaic (aka pixelate) [28], and Oil Paint [10, 28]. The latter was shown in multiple works to be limited in terms of observation resistance [10, 28]. On the other hand, filters such as Color-halftone and Gaussian-Blur were never investigated before. Thus, we chose to experiment with Color-halftone, Crystallize, Gaussian-Blur and Mosaic.

Filters can be applied in a wide range of ways and degrees (e.g., radius of the distortions). To determine the optimal parameters, we ran multiple pilot tests and an online questionnaire (N=163) where we experimented with different strengths of each filter.

4.1 Online Questionnaire

In the online questionnaire we evaluated the four different filters with two strengths each. We distributed the questionnaire via university mailing lists. We asked participants about visual impairments: 1 had Keratoconus, 4 were long sighted, 51 were short sighted (7 of which had astigmatism), and 8 reported poor eyesight without further details. The majority reported using sight correction Participants answered 20 questions out of a pool of 48 questions about 1) guessability with prior knowledge of the clear text (e.g., which of those distorted texts correspond to the word *university*?), and 2) guessability without prior knowledge of the clear text (e.g., which of the following texts is the clear version of this distorted image?). Questions of type (1) are to simulate cases where a user knows the clear version of their password, and matches it to a distorted version. While questions of type (2) simulate cases where an observer guesses distorted versions of passwords but they do not know the undistored version. Out of the 48 questions, 24 were type (1) and 24 were type (2). Each participant was randomly allocated 10 of each type. Each of the 48 questions was answered at least 50 times All participants answered questions from both types.

We used Adobe Photoshop to apply the filters on pre-prepared text snippets. All snippets used 11pts font-size, normal font-weight, character-spacing of 0%, and were in black color (#000000). The text snippets were then rasterized with a resolution of 300 Pixels/inch in sRGB IEC61966-2.1 Mode with a bit depth of 8 bit. Afterwards, the distortion filters were applied on the text-elements, each in two variants that were found promising through pilot tests:

- Color-halftone: 5 pixels and 6 pixels radii.
- Crystallize: 7 pixels and 8 pixels cell size.
- Gaussian-Blur: 7.5 pixels and 8 pixels radii.
- Mosaic: 10 pixels and 11 pixels cell sizes.

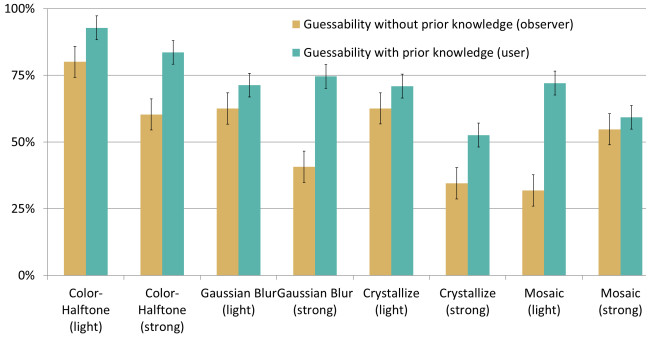

Figure 2: Results from an online questionnaire (N=163) indicate that prior knowledge of distorted text (e.g., a user examining a distorted password that they have just entered) improves guessability. On the other hand, not having prior knowledge of the distorted word (e.g., an observer shoulder surfs a distorted password that a user has entered) makes it more difficult to guess the distorted password.

The light and strong variants were based on pilot tests in which we heuristically tested all possible variants among the authors and 4 participants to exclude those very clear to observers (lower limit) and those unclear to users (upper limit).

A total of 163 participants completed the questionnaire. In open questions where participants had to provide their guesses in text, we measured the Levenshtein distance between the response and the correct answer. While multiple choice questions were evaluated as either correct (1/1) or incorrect (0/1). The results for each filter type and strength are aggregated and summarized in Figure 2. The figure shows that for all filters of all strengths, participants are better at interpreting distorted words when they know the original clear version, which is inline with prior work [2, 3, 6, 9, 10, 19]. We raffled 3 online shop vouchers to compensate participants.

We chose to prioritize usability over security in our selection of the filters. This was done to ensure that we reach an implementation that users would actually accept to be willing to use. Otherwise, optimizing for security at the expense of usability might result in lower acceptance, which eventually reduces security and reduces the overall value. Hence, we picked the filter strengths that provide higher guessability to the user who has prior knowledge of the text, i.e., the user who has just entered a password rather than the observer who does not know the password. Therefore, we chose to focus on these filters in our follow up study: the light variation of Color-halftone (5 pixels radius), the light variation of Crystallize (7 pixels cell size), the strong variation of Gaussian-Blur (8 pixels radius), and the light variation of Mosaic (10 pixels).

4.2 Implementation

Based on the results of the online questionnaire, we implemented the filters in Javascript to allow experimenting with the different filters in real time while authenticating on mobile devices. We used the JSManipulate library[1] to implement the Gaussian-Blur and Mosaic filters using the `blur_default` and `pixelate_default` functions respectively. We implemented Color-halftone and Crystallize filters

[1] https://github.com/JoelBesada/JSManipulate

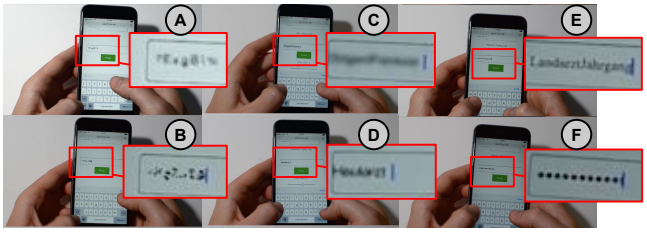

Figure 3: Participants entered passwords using A) Color-halftone filter, B) Crystallize filter, C) Gaussian-Blur filter, D) Mosaic filter, E) Plain text and F) Asterisks.

by creating custom web fonts. To create the web fonts, we first applied the respective filters on all alphanumeric characters using Adobe Photoshop the same way we did for the online questionnaire. Distorted characters were then exported as TFF images, which were then used to create the web font.

5 USER STUDY

To evaluate the impact of the filters on the text password's usability and security, we conducted a study that is divided into two experiments: a usability evaluation experiment and a security one. Both experiments complied with university's ethics regulations. All participants started with the usability experiment, followed by the security experiment. We used a repeated-measures design for both experiments with two independent variables:

- IV1 Filter: we experimented with two baselines: Plain text, Asterisks, and four filters: Color-halftone, Crystallize, Gaussian-Blur, and Mosaic filters.
- IV2 Password type: we experimented with three types: one-word passwords, random characters passwords (including digits, uppercase, lowercase, and special characters), and passphrases [15].

We invited 25 participants (11 females) aged between 18 and 32 years (M = 26.64, SD = 3.74) through the word of mouth and university mailing lists. When asked, participants indicated that they use text passwords regularly: 9 enter text passwords more than once per day, 8 enter them more than once per week, 4 enter them more than once per month, and 4 enter them once or less per month. Participants were compensated with an e-voucher.

5.1 Usability Experiment

We evaluated impact of using the different filters on the usability of text passwords on mobile devices. Participants performed 36 blocks (6 filters × 3 password types × 2 runs). In each block, they had to *input* a unique password and then *edit* the password they had just entered. The order of conditions was counter balanced using a Latin square. We measured the input time, input accuracy, editing time, and editing accuracy.

In the *input task*, participants had to enter a text password on their mobile device. They were shown a password on a computer screen, and were then asked to enter it in an input field on a customized website that ran on a local server (see Figure 3). The website was accessed through the participant's own mobile device. The respective filters were applied on the entered password according to the Latin square. We measured the input time and accuracy.

Figure 4: To evaluate observation resistance, participants watched high quality videos of users entering passwords masked with the different filters. Their task was to provide up to 3 guesses of the observed password.

In the *editing task*, participants had to edit the password they had just entered. They were shown the same password on the computer screen, but this time with a red arrow pointing between two characters. Participants were asked to insert the character 'X' at the shown position. Although actions like replacing and deleting characters are also typical editing tasks, we opted for this task because 1) it requires recognizing certain characters before knowing where to add the 'X', and 2) it is a prerequisite step for replacing, and deleting characters. We measured the editing time and accuracy.

Participants used their own smartphones in both tasks to avoid the impact of using an unfamiliar device. Some mobile devices provide feedback about the tapped key by enlarging the tapped key briefly. We disabled this feature on the participants' smartphones to ensure a fair comparison.

5.2 Security Experiment

Since the filters affect the way the passwords are shown, we evaluated the impact of using filters on their observation resistance.

To this end, we showed participants videos of users entering text passwords on a mobile device from an optimal angle (see Figure 4). Those videos were previously recorded from the same angle as one of the authors authenticated. The participant's task was to act as a shoulder surfer and try to infer the entered password. Participants watched two high quality video recordings of a user entering a text password in 18 conditions (6 filters × 3 password types). Thus, participants performed a total of 36 attacks. Participants were provided with pen and paper to take notes, and were allowed to provide up to 3 guesses. We measured the successful guessing rate, and the accuracy of the guesses. To motivate participants to put an effort in their attacks, we arranged a raffle where every correct guess increases the chance to win an additional voucher.

6 RESULTS

This section presents the results of the usability and security evaluations.

Editing Time

	Color-halftone 7.1 s	Crystallize 7.97 s	Gaussian-Blur 10.54 s	Mosaic 9.67 s	Plain text 6.18 s	Asterisks 9.94 s
Color-halftone 7.1 s	-	p > 0.05	p < 0.05	p > 0.05	p > 0.05	p < 0.005
Crystallize 7.97 s	p > 0.05	-	p > 0.05	p > 0.05	p > 0.05	p < 0.05
Gaussian-Blur 10.54 s	p < 0.05	p > 0.05	-	p > 0.05	p < 0.005	p < 0.001
Mosaic 9.67 s	p > 0.05	p > 0.05	p > 0.05	-	p > 0.05	p > 0.05
Plain text 6.18 s	p > 0.05	p > 0.05	p < 0.005	p > 0.05	-	p < 0.001
Asterisks 9.94 s	p < 0.005	p < 0.05	p < 0.001	p > 0.05	p < 0.001	-

Editing Accuracy

	Color-halftone 0.207	Crystallize 0.227	Gaussian-Blur 1.067	Mosaic 1.067	Plain text 0.04	Asterisks 9.94 s
Color-halftone 0.207	-	p > 0.05	p < 0.001	p < 0.005	p > 0.05	p > 0.05
Crystallize 0.227	p > 0.05	-	p < 0.001	p < 0.005	p > 0.05	p > 0.05
Gaussian-Blur 1.067	p < 0.001	p < 0.001	-	p > 0.05	p < 0.001	p < 0.001
Mosaic 1.067	p < 0.005	p < 0.005	p > 0.05	-	p < 0.001	p < 0.001
Plain text 0.04	p > 0.05	p > 0.05	p < 0.001	p < 0.001	-	p > 0.05
Asterisks 0.193	p > 0.05	p > 0.05	p < 0.001	p < 0.001	p > 0.05	-

Table 1: The tables highlight the pairs with significantly different editing time (left) and accuracy (right). Plain text, Color-halftone, and Crystallize filters result in both the fastest and most accurate editing. Editing when using Asterisks is accurate but slow, because participants count the Asterisks until they find the desired position rather than recognizing the characters.

6.1 Usability Evaluation Results

We measured 1) input time, 2) input accuracy, 3) editing time, and 4) editing accuracy.

6.1.1 Input time. was measured in milliseconds from the moment the user started entering the first character until the moment the last character was entered. The input times are: 12.1 s for Color-halftone (SD = 0.55 s), 12.64 s for Crystallize (SD = 0.69 s), 12.73 s for Gaussian-Blur (SD = 0.67 s), 12.33 s for Mosaic (SD = 0.66 s), 12.32 s for Plain text (SD = 0.66 s), and 12.45 s for Asterisks (SD = 0.63 s). We found no significant effect of filter type in input time ($p = 0.68$). This means we found no evidence that the filters impact input time.

On the other hand, a repeated measures ANOVA revealed a significant main effect of password types on input time ($F_{2,48} = 236.723$, $p < 0.001$). Post hoc pair-wise comparisons with Bonferroni correction showed significant differences between all pairs (all $p < 0.001$): one-word passwords are significantly the fastest to enter ($M = 7.66\,s$, $SD = 0.42\,s$) followed by passphrases ($M = 12.53\,s$, $SD = 0.73\,s$), and then random character passwords ($M = 17.32\,s$, $SD = 0.615\,s$). This confirms prior work on passphrases [14, 15]. We found no interaction between filter and password type ($p = 0.856$).

6.1.2 Input accuracy. was measured by calculating the Levenshtein distance between the entry and the actual password. The input accuracy for each filter type is: 0.07 for Color-halftone (SD = 0.26), 0.11 for Crystallize (SD = 0.36), 0.09 for Gaussian-Blur (SD = 0.38), 0.05 for Mosaic (SD = 0.23), 0.05 for Plain text (SD = 0.24), and 0.05 for Asterisks (SD = 0.23). While the input accuracy for each password type is: 0.05 for one-word passwords (SD = 0.22), 0.04 for randomized-characters passwords (SD = 0.22), and 0.12 for passphrases (0.41). We found no significant effect of filter type or password type on input accuracy (both $p > 0.05$). No interaction was found between filter type and password type as well ($p > 0.05$). This means that there is no evidence that the filters or password types influence input accuracy.

6.1.3 Editing time. was measured from the moment the user tapped a button that shows the last entered password with the filter applied on it, until the user has entered the character 'X'. A repeated measures ANOVA with Greenhouse-Geisser correction (due to violation of the sphericity assumption) revealed a significant main effect of filter type on editing time ($F_{2.47,59.17} = 6.45$, $p < 0.005$). Table 1 shows the results of the pair-wise comparisons (all Bonferroni corrected). In summary, users edit passwords in Asterisks in significantly more time compared to Color-halftone, Crystallize, Gaussian-Blur and Plain text. Color-halftone results in significantly shorter editing time compared to Gaussian-Blur. Editing in Plain text is significantly faster than in Gaussian-Blur, but no significant differences were found between Plain text and either of Color-halftone or Crystallize. This means that that Plain text, Color-halftone, and Crystallize are fastest in terms of editing time.

Editing time is 8.15 s for one-word passwords (SD = 0.58), 8.62 s for randomized-characters passwords (SD = 0.57), and 8.94 s for passphrases (SD = 0.61). No Significant main effect of password type was found on editing time ($p > 0.05$).

6.1.4 Editing accuracy. was measured by counting the number of characters between the correct position and the expected position. A repeated measures ANOVA with Greenhouse-Geisser correction (due to violation of the sphericity assumption) revealed a significant main effect of filter type on editing time ($F_{2.47,59.17} = 6.45$, $p < 0.005$). Table 1 shows the results of the post hoc pair-wise comparison with Bonferroni correction. Editing is most accurate when using Plain text. Interestingly, participants' edits were accurate when using Asterisks because, as they explained, they counted the characters until they found the correct position. This however takes time as reflected by the long editing times associated with Asterisks (see Table 1). On the other hand, Color-halftone and Crystallize result in significantly higher accuracy than Gaussian-Blur and Mosaic. This means that that Plain text, Asterisks, Color-halftone, and Crystallize result in the most accurate editing.

Mean editing accuracy is 0.44 for one-word passwords (SD = .79), 0.43 for randomized-characters passwords (SD = .82), and 0.56 for passphrases (SD = .92). We found no significant effect of password type on editing accuracy $p > 0.05$.

6.2 Security Evaluation Results

We measured the successful attack rate, and the accuracy of the guesses. Out of the 3 guesses made in each attack, only the best guess (i.e., the one with the shortest Levenshtein distance to the actual password) was considered in the analysis. This was done to evaluate the filters in worst case scenarios.

6.2.1 Successful Attacks. Figure 5 summarizes the rate of successful attacks against each filter and each password type. Participants were least successful when attacking passwords on which Color-halftone and Crystallize filters were applied. Gaussian-Blur and Mosaic were

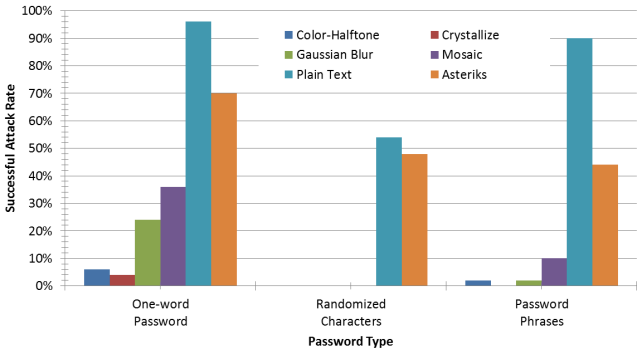

Figure 5: Successful attack rates are lowest against passwords masked with Color-halftone and Crystallize filters. Gaussian-Blur and Mosaic filters perform well with random characters and passphrases. Guesses against Plain text are the most successful, followed by Asterisks because of revealing the entered character briefly before masking it.

moderately secure against shoulder surfing, but not when using one-word passwords.

6.2.2 Accuracy of Guesses. To measure the accuracy of the guesses, we calculated the Levenshtein distance between the attacker's guess and the actual password. Since the passwords were of different lengths, the Levenshtein distances were first normalized to a 0 to 1 range. A repeated measures ANOVA revealed a significant main effect of the filter type ($F_{5,120} = 186.9, p < 0.001$) on the accuracy of guesses. Post hoc pair-wise comparisons with Bonferroni correction revealed significant differences between multiple pairs as illustrated in Table 2. In summary, Table 2 shows that guesses against passwords on which any of the filters is applied are significantly farther away from the actual password (i.e., less correct) compared to those against passwords displayed in Asterisks and Plain text. It shows that Color-halftone, Crystallize and Gaussian-Blur outperform Mosaic, that Crystallize outperforms both Gaussian-Blur and Mosaic, and that Asterisks outperforms Plain text. The low observation resistance of Asterisks is due to current implementations on widely used operating systems, where typed characters are revealed for some milliseconds before being masked.

A repeated measures ANOVA with Greenhouse-Geisser correction (due to violation of the sphericity assumption) revealed a significant main effect of the password type ($F_{1.62,38.78} = 5.7, p < 0.05$) on the Levenshtein distance between the guess and the password. Post hoc pair-wise comparisons with Bonferroni correction revealed that guesses against passphrases ($M = 51.7\%$, $SD = 1.6\%$) are significantly less successful than guesses against one-word passwords ($M = 58.3\%$, $SD = 2.5\%$), $p < 0.005$ and those against random character passwords ($M = 56.3\%$, $SD = 1.5\%$) $p < 0.05$. This is inline with previous work [14].

A statistically significant interaction effect was found between the filter type and password type ($F_{5.31,127.46} = 4.57, p < 0.001$). Therefore we ran subsequent one-way repeated measures ANOVA tests to compare the filters for each password type. In all tests, attacks against passwords with either Color-halftone or Crystallize filters were found to be significantly less accurate compared to the baselines.

Guessing Accuracy						
	Color-halftone 31%	Crystallize 23.4%	Gaussian-Blur 38.9%	Mosaic 55.8%	Plain text 96.5%	Asterisks 86.8%
Color-halftone 31%	-	p > 0.05	p > 0.05	p < 0.001	p < 0.001	p < 0.001
Crystallize 23.4%	p > 0.05	-	p < 0.005	p < 0.001	p < 0.001	p < 0.001
Gaussian-Blur 38.9%	p > 0.05	p < 0.005	-	p < 0.005	p < 0.001	p < 0.001
mosaic 55.8%	p < 0.001	p < 0.001	p < 0.005	-	p < 0.001	p < 0.001
Plain text 96.5%	p < 0.001	p < 0.001	p < 0.001	p < 0.001	-	p < 0.001
Asterisks 86.8%	p < 0.001	p < 0.001	p < 0.001	p < 0.001	p < 0.001	-

Table 2: Pair-wise comparisons show that Crystallize and Color-halftone filters are significantly more shoulder surfing resilient compared to Mosaic, Plain text and Asterisks. Crystallize is also significantly more secure compared to Gaussian-Blur. All filters provide higher protection than Plain text and current implementations of Asterisks.

6.3 Qualitative Feedback

After both the usability and security experiments, we condluded with a semi-structured interview. Feedback was analyzed using thematic analysis and clustered into 3 themes.

6.3.1 Theme 1: Easier editing using the proposed filters. Several participants (N=7) indicated that editing text passwords that are masked with the proposed filters is easier than when they are masked using Asterisks. P2 highlighted that unlike the Asterisks, the filters allow noticing typing errors and thereby easier error correction. P7, P18 and P25 similarly stated that they found recognizing typed characters to be easier when using the filters. P18 added that he needs to count characters to correct errors when using Asterisks. P21 found it hard to distinguish some letters (e.g., a and e) but still commended that "it is nice to see some context rather than none at all". P23 added that she prefers the filters over Asterisks, because with the latter she often needs to reenter the whole password if she thinks she made a mistake.

6.3.2 Theme 2: Difficulty compared to Plain text. P24 and P14 highlighted that they found typing using Plain text to be the easiest, since it is much easier to notice typing mistakes.

6.3.3 Theme 3: Better observation resistance. P13 and P14 admired that they are able to perceive the distorted words that they wrote themselves, but not those that others wrote. P14 added that he "noticed himself [that] the input is better protected from prying eyes". P22 expressed concerns that although the filters resist observations better, attackers might eventually become more proficient in being able to read distorted text.

6.3.4 Preferences. We asked participants which was easiest to use and hardest to observe as a proxy for the subjective usability/observability trade off. When asked to pick their favorite filter, 9 participants picked Crystallize, 8 picked Color Halftone, 6 picked Gaussian Blur and 3 picked Mosaic.

7 DISCUSSION AND FUTURE WORK

Overall, the main strength of the filters is in improving editing performance. Namely, we found that editing speed and accuracy are significantly higher when using Color-halftone and Crystallize

filters compared to Gaussian-Blur and Mosaic. However, Color-halftone and Crystallize are outperformed by Plain text in terms of editing accuracy and speed. This is expected, since no masking is taking place at all. Participants were significantly slower when editing Asterisks due to the need to count the characters. At the same time, the security experiment also shows that the same two filters, Color-halftone and Crystallize, are significantly more resilient to shoulder surfing compared to Plain text and Asterisks. The results show that the idea of showing characters briefly after they are typed (i.e., Asterisks) indeed improves resistance to shoulder surfing compared to Plain text, but observers can still successfully shoulder surf the vast majority of inputs. We expect that a desktop version of Asterisks masking would be more secure against shoulder surfing, but would make editing time and accuracy even worse than when using the mobile version of Asterisks.

Asterisks are good at making it difficult to find the entire password. But a closer look at the accuracy of guesses reveals that attackers are still able to recover big part of the password even when masked using Asterisks. This suggests that attackers are more likely to recover the password after few observations. On the other hand, the filters reduce both the accuracy of guesses as well as the successful attack rates.

7.1 Improved Guessing Attacks

Attackers who have a hint what the distorted password could be, might have a stronger chance at guessing the password. For example, if the attacker knows that the distorted character is a "vehicle", it will be easier to notice the password if it is "car". Another example can be seen in Figure 1. Once the reader knows that one of the words (e.g., Crystallize) refers to a graphic filter, it will become easier to guess some of the other distorted words. This is because any hints about the domain of the word helps the observer recall relevant words, hence changing the challenge from a recall challenge (i.e., "what is this word") to a recognition challenge (i.e., "which filter does this word refer to"). Another interesting question is whether attackers can improve their skills in attacking distorted passwords overtime. This can be explored in follow-up longitudinal study.

7.2 Impact on Password Creation and Memorability

An interesting aspect to be explored in future work is how the filters influence text password creation. Users could create text passwords that are more secure against observations if they intentionally choose characters whose distortions are difficult to interpret. However it is not clear if users are willing to do that. Another direction for future work is to explore how filters impact password memorability. Graphical elements (e.g., emojis) in text passwords were shown to improve memorability [24]. This suggests that graphic filters might positively impact memorability if users associate the shape of the distortions with their text password.

7.3 Impact of Password Policies on the Usability and Security of Filters

In our experiment, we considered passwords that contain digits, uppercase, lowercase, and special characters. An open direction for future work is to study the individual impact of each of those password features on the usability and security of passwords that are masked with graphical filters.

7.4 Filters on Other platforms

We focused on mobile devices because they would benefit the most from improved usability of password entry [13, 21]. A direction for future work is to experiment with filters on desktop settings. Here we expect that there would be improvement in error correction, but we do not expect the result to be as significant as in case of mobile devices. Furthermore, we do know that shoulder surfing is an issue on mobile devices [4], but we do not have equally strong evidence for desktop devices. Therefore, overall there is less evidence that the filters will be significantly useful for desktop settings.

7.5 Usability and Security Trade off

The trade off between usability and security was widely discussed in previous work. We argue that the ideal security and privacy protection mechanisms are those that do not reduce usability; for example, using filters that greatly distort the text password will make it more secure against observations, but harder to edit. This drop in usability might discourage users from adopting the filters, leading to even less security This is why we prioritized usability in the analysis of the survey results. Still, the trade off is present in our results: the Crystallize and Color-Halftone filters are more secure but not as usable as Plain text. However, they significantly improve both usability and security compared to asterisks.

7.6 Implications for Designers

We drew the following implications based on our results:

(1) Since they improve editing time and accuracy, and observation resistance, we recommend defaulting to Color halftone (5 pixels radius) or Crystallize (7 pixels cell size) filters instead of Asterisks, Mosaic and Gaussian Blur.
(2) Filters should be required in contexts in which shoulder surfing occurs [4]. This can be determined using the device's sensors (e.g. user is in public transport), or by detecting shoulder surfers through the front facing camera [1].
(3) Plain text is ineffective against observations, but editing it is fast and accurate. To improve usability, allow switching to Plain text if the context does not suggest the presence of shoulder surfers, or the user confirms being alone.

8 CONCLUSION

In this work, we investigated the use of filters to mask text passwords on mobile devices. We picked the filters to investigate based on prior work, pilot tests and an online questionnaire. Based on this, we experimented with the Color-halftone, Crystallize, Gaussian-Blur, and Mosaic filters. We then conducted follow up usability and security within-subjects experiments (N=25) where we compared the filters to two baselines: Plain text and current implementations of Asterisks. In the usability experiment, we measured input time, input accuracy, editing time, and editing accuracy. While we found no significant differences in terms of input time, and input accuracy, we found that Plain text, Color-halftone and Crystallize result in the fastest and most accurate editing. In the security experiment, we found that Color-halftone and Crystallize are significantly better at

observation resistance compared to the other conditions. Based on the results, we recommend using Color-halftone and Crystallize to improve editing performance and observation resistance.

In addition to the directions for future work that were suggested in the previous section, we plan to conduct a field study to study how well users notice their typing errors when using the filters in the wild.

REFERENCES

[1] Mohammed Eunus Ali, Anika Anwar, Ishrat Ahmed, Tanzima Hashem, Lars Kulik, and Egemen Tanin. 2014. Protecting Mobile Users from Visual Privacy Attacks. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1–4. https://doi.org/10.1145/2638728.2638788

[2] A. Mike Burton, Stephen Wilson, Michelle Cowan, and Vicki Bruce. 1999. Face Recognition in Poor-Quality Video: Evidence from Security Surveillance. *Psychological Science* 10, 3 (1999), 243–248. http://www.jstor.org/stable/40063419

[3] Tamara Denning, Kevin Bowers, Marten van Dijk, and Ari Juels. 2011. Exploring Implicit Memory for Painless Password Recovery. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 2615–2618. https://doi.org/10.1145/1978942.1979323

[4] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 2017. Understanding Shoulder Surfing in the Wild: Stories from Users and Observers. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. ACM, New York, NY, USA, 4254–4265. https://doi.org/10.1145/3025453.3025636

[5] Malin Eiband, Emanuel von Zezschwitz, Daniel Buschek, and Heinrich Hussmann. 2016. My Scrawl Hides It All: Protecting Text Messages Against Shoulder Surfing With Handwritten Fonts. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 2041–2048. https://doi.org/10.1145/2851581.2892511

[6] Gail S Goodman. 1980. Picture memory: How the action schema affects retention. *Cognitive Psychology* 12, 4 (1980), 473 – 495. https://doi.org/10.1016/0010-0285(80)90017-1

[7] Kristen K. Greene, Melissa A. Gallagher, Brian C. Stanton, and Paul Y. Lee. 2014. I Can't Type That! P@$$w0rd Entry on Mobile Devices. In *Human Aspects of Information Security, Privacy, and Trust*, Theo Tryfonas and Ioannis Askoxylakis (Eds.). Springer International Publishing, Cham, 160–171.

[8] S M Taiabul Haque, Matthew Wright, and Shannon Scielzo. 2013. Passwords and Interfaces: Towards Creating Stronger Passwords by Using Mobile Phone Handsets. In *Proceedings of the Third ACM Workshop on Security and Privacy in Smartphones & Mobile Devices (SPSM '13)*. ACM, New York, NY, USA, 105–110. https://doi.org/10.1145/2516760.2516767

[9] Atsushi Harada, Takeo Isarida, Tadanori Mizuno, and Masakatsu Nishigaki. 2006. *A User Authentication System Using Schema of Visual Memory*. Springer Berlin Heidelberg, Berlin, Heidelberg, 338–345. https://doi.org/10.1007/11613022_28

[10] Eiji Hayashi, Rachna Dhamija, Nicolas Christin, and Adrian Perrig. 2008. Use Your Illusion: Secure Authentication Usable Anywhere. In *Proceedings of the 4th Symposium on Usable Privacy and Security (SOUPS '08)*. ACM, New York, NY, USA, 35–45. https://doi.org/10.1145/1408664.1408670

[11] Zoë Henderson, Vicki Bruce, and A. Mike Burton. 2001. Matching the faces of robbers captured on video. *Applied Cognitive Psychology* 15, 4 (2001), 445–464. https://doi.org/10.1002/acp.718 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/acp.718

[12] Ponemon Institute. 2016. Global Visual Hacking Experimental Study: Analysis. multimedia.3m.com/mws/media/1254232O/global-visual-hacking-experiment-study-summary.pdf

[13] Markus Jakobsson and Ruj Akavipat. 2011. Rethinking passwords to adapt to constrained keyboards. http://www.markus-jakobsson.com/fastwords.pdf

[14] Mark Keith, Benjamin Shao, and Paul Steinbart. 2009. A Behavioral Analysis of Passphrase Design and Effectiveness. *Journal of the Association for Information Systems* 10, 2, Article 2 (2009). https://aisel.aisnet.org/jais/vol10/iss2/2

[15] Mark Keith, Benjamin Shao, and Paul John Steinbart. 2007. The usability of passphrases for authentication: An empirical field study. *International Journal of Human-Computer Studies* 65, 1 (2007), 17 – 28. https://doi.org/10.1016/j.ijhcs.2006.08.005 Information security in the knowledge economy.

[16] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. 2016. GazeTouchPass: Multimodal Authentication Using Gaze and Touch on Mobile Devices. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, New York, NY, USA, 2156–2164. https://doi.org/10.1145/2851581.2892314

[17] Mohamed Khamis, Malin Eiband, Martin Zürn, and Heinrich Hussmann. 2018. EyeSpot: Leveraging Gaze to Protect Private Text Content on Mobile Devices from Shoulder Surfing. *Multimodal Technologies and Interaction* 2, 3, Article 45 (2018). https://doi.org/10.3390/mti2030045

[18] Mohamed Khamis, Mariam Hassib, Emanuel von Zezschwitz, Andreas Bulling, and Florian Alt. 2017. GazeTouchPIN: Protecting Sensitive Data on Mobile Devices Using Secure Multimodal Authentication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI 2017)*. ACM, New York, NY, USA, 446–450. https://doi.org/10.1145/3136755.3136809

[19] Hikari Kinjo and Joan Gay Snodgrass. 2000. Does the Generation Effect Occur for Pictures? *The American Journal of Psychology* 113, 1 (2000), 95–121. http://www.jstor.org/stable/1423462

[20] Katharina Krombholz, Thomas Hupperich, and Thorsten Holz. 2016. Use the Force: Evaluating Force-Sensitive Authentication for Mobile Devices. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 207–219. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/krombholz

[21] William Melicher, Darya Kurilova, Sean M. Segreti, Pranshu Kalvani, Richard Shay, Blase Ur, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, and Michelle L. Mazurek. 2016. Usability and Security of Text Passwords on Mobile Devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 527–539. https://doi.org/10.1145/2858036.2858384

[22] Karen Renaud. 2009. Guidelines for Designing Graphical Authentication Mechanism Interfaces. *Int. J. Inf. Comput. Secur.* 3, 1 (June 2009), 60–85. https://doi.org/10.1504/IJICS.2009.026621

[23] Florian Schaub, Ruben Deyhle, and Michael Weber. 2012. Password Entry Usability and Shoulder Surfing Susceptibility on Different Smartphone Platforms. In *Proceedings of the 11th International Conference on Mobile and Ubiquitous Multimedia (MUM '12)*. ACM, New York, NY, USA, Article 13, 10 pages. https://doi.org/10.1145/2406367.2406384

[24] Tobias Seitz, Manuel Hartmann, Jakob Pfab, and Samuel Souque. 2017. Do Differences in Password Policies Prevent Password Reuse ?. In *CHI '17 Extended Abstracts on Human Factors in Computing Systems*. ACM, Denver, CO, USA. https://doi.org/10.1145/3027063.3053100

[25] Emanuel von Zezschwitz, Alexander De Luca, Bruno Brunkow, and Heinrich Hussmann. 2015. SwiPIN: Fast and Secure PIN-Entry on Smartphones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 1403–1406. https://doi.org/10.1145/2702123.2702212

[26] Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2014. Honey, I Shrunk the Keys: Influences of Mobile Devices on Password Composition and Authentication Performance. In *Proceedings of the 8th Nordic Conference on Human-Computer Interaction: Fun, Fast, Foundational (NordiCHI '14)*. ACM, New York, NY, USA, 461–470. https://doi.org/10.1145/2639189.2639218

[27] Emanuel von Zezschwitz, Paul Dunphy, and Alexander De Luca. 2013. Patterns in the Wild: A Field Study of the Usability of Pattern and Pin-based Authentication on Mobile Devices. In *Proceedings of the 15th International Conference on Human-computer Interaction with Mobile Devices and Services (MobileHCI '13)*. ACM, New York, NY, USA, 261–270. https://doi.org/10.1145/2493190.2493231

[28] Emanuel von Zezschwitz, Sigrid Ebbinghaus, Heinrich Hussmann, and Alexander De Luca. 2016. You Can't Watch This!: Privacy-Respectful Photo Browsing on Smartphones. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 4320–4324. https://doi.org/10.1145/2858036.2858120

[29] Yulong Yang, Gradeigh D. Clark, Janne Lindqvist, and Antti Oulasvirta. 2016. Free-Form Gesture Authentication in the Wild. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (CHI '16)*. ACM, New York, NY, USA, 3722–3735. https://doi.org/10.1145/2858036.2858270

[30] Yulong Yang, Janne Lindqvist, and Antti Oulasvirta. 2014. Text Entry Method Affects Password Security. In *The LASER Workshop: Learning from Authoritative Security Experiment Results (LASER 2014)*. USENIX Association, Arlington, VA. https://www.usenix.org/conference/laser2014/program/agenda/presentation/yang