Mihai Bâce*, Alia Saad*, Mohamed Khamis, Stefan Schneegass, and Andreas Bulling

PrivacyScout: Assessing Vulnerability to Shoulder Surfing on Mobile Devices

Abstract: One approach to mitigate shoulder surfing attacks on mobile devices is to detect the presence of a bystander using the phone's front-facing camera. However, a person's face in the camera's field of view does not always indicate an attack. To overcome this limitation, in a novel data collection study (N=16), we analysed the influence of three viewing angles and four distances on the success of shoulder surfing attacks. In contrast to prior works that mainly focused on user authentication, we investigated three common types of content susceptible to shoulder surfing: text, photos, and PIN authentications. We show that the vulnerability of text and photos depends on the observer's location relative to the device, while PIN authentications are vulnerable independent of the observation location. We then present *PrivacyScout* – a novel method that predicts the shoulder-surfing risk based on visual features extracted from the observer's face as captured by the front-facing camera. Finally, evaluations from our data collection study demonstrate our method's feasibility to assess the risk of a shoulder surfing attack more accurately.

Keywords: Privacy, Security, Mobile Devices, Front-facing Camera

DOI 10.56553/popets-2022-0090
Received 2021-11-30; revised 2022-03-15; accepted 2022-03-16.

1 Introduction

Shoulder surfing on mobile devices, i.e. the act of observing content shown on the device screen without users' consent, has been shown to have serious impact on users' privacy, security and even safety [15, 30]. Consequently, a large body of research has investigated means to mitigate shoulder surfing attacks during user authentication, such as using image-based methods [5, 42]), additional input modalities (e.g. multimodal gaze-based authentication [3, 23, 25]), or by physically augmenting the display with a screen protector [34]. While such methods can generally reduce the risk of a loss of sensitive information, they are unable to warn users *when* a shoulder surfing attack is actually happening at a particular moment in time.

A promising approach to detect shoulder surfing on mobile devices in situ is to detect the presence of a bystander using the front-facing camera, which can nowadays capture high-resolution images at increasingly wider fields of view (e.g. the Samsung Galaxy A80 has an ultra wide front-facing camera with 123° field of view [39]). However, prior work [37] assumed that the appearance of a face in the camera's view always implies a successful attack. In practice, however, the bystander may not be able to actually see any content if they are too far away or the observation angle is too steep. In addition, warning users every time a face is detected without assessing the real shoulder surfing risk leads to a "crying wolf" problem [40], which would minimise the system's effectiveness.

To address this limitation, we provide the first detailed analysis of the key characteristics that influence the likelihood of a shoulder surfing attack on mobile devices: the viewing distance, the observation angle, and the type of on-screen content. While the work by Aviv et al. [9] had a similar goal of assessing the potential risk of content observation from different angles, they focused only on authentication and two viewing points. On the other hand, Ali et al. [4] based their work on the distance between the attacker and the phone to determine whether a bystander could read the displayed text. Their work relied on a paper-based visual acuity test [1] to asses readability of text. However, prior work has

Mihai Bâce*: University of Stuttgart, Germany, E-mail: mihai.bace@vis.uni-stuttgart.de
Alia Saad*: University of Duisburg-Essen, Germany, E-mail: alia.saad@uni-due.de
Mohamed Khamis: University of Glasgow, United Kingdom, E-mail: mohamed.khamis@glasgow.ac.uk
Stefan Schneegass: University of Duisburg-Essen, Germany, E-mail: stefan.schneegass@uni-due.de
Andreas Bulling: University of Stuttgart, Germany, E-mail: andreas.bulling@vis.uni-stuttgart.de

* Both authors contributed equally to this research.

shown that the readability of text differed when content was shown on paper vs on a display [27]. In addition, the visual acuity test may not fully capture different characteristics of text comprehension, e.g. picking up words or understanding the context of a conversation. We extend these prior works along two dimensions. First, our work is the first to quantify the success of shoulder surfing attacks that target text, photos, and PINs, all of which have been previously identified as sensitive in day-to-day situations [15]. Second, our data collection study covers a wider range of distances (50 cm, 100 cm, 150 cm, and 200 cm) and observation angles (0°, 30°, and 60°) relative to the device. This allowed us to simulate shoulder surfing of different content types in contexts that match those described by victims of shoulder surfing [15]. For each content type, our novel study allowed us to analyse different quantitative metrics that estimate the likelihood of an attacker to, e.g. retrieve names from a piece of text or observe the facial expressions of people in a photo. In contrast to relying on a standard visual acuity test [1], we relied on feedback from our study participants to extract these metrics that were specifically designed for each content type.

Our analyses reveal a number of interesting findings. For example, while an attacker may easily observe whether a photo was taken indoors or outdoors, or count how many people are present in a photo, they are unlikely to observe finer details, such as facial expressions, when the attack is performed from farther away and at a steeper angle. Furthermore, PIN authentications are generally more vulnerable independent of the distance or observation angle, with a very high chance of success. Using the dataset from our user study, we further investigated the feasibility of building a computational model that can predict the vulnerability of a shoulder surfing attack. We propose PrivacyScout, a novel method that extracts visual features from the attacker's face, as captured by a (wide-angle) front-facing camera of a mobile device and regresses a shoulder surfing risk score – which is a direct measure of the potential risks. The risk assessment score is inspired from and builds on the results of our analysis and is a value between 0 and 3 for the text condition and 0 and 4 for the photo condition – where a higher value means a higher shoulder surfing risk for the mobile device user. A risk assessment score for the authentication scenario was not justified as our findings showed that PINs were equally vulnerable independent of the observation location. We evaluated our method, including variations of it that use different feature subsets, on the dataset we collected in our user study and compared it to several naïve baselines. Results show that, for example, a variant of our method that used all features to regress the risk score for textual content achieved a relative decrease of around 24% in the mean absolute error (MAE): 0.56 (SD=0.38) ours vs. 0.74 (SD=0.58) for the naïve baseline – lower is better. A decrease can also be observed by more accurately predicting the risk score for photos: 0.41 MAE (SD=0.14) for our method vs. 0.56 MAE (SD=0.23) for the best naïve baseline.

Contribution Statement The contributions of our work are two-fold: We present the first detailed analysis on the success of shoulder surfing *beyond* user authentication against commonly attacked content types (PIN, photo, and text) on mobile devices. Our fine-grained analyses (four distances × three observation angles) are based on a 16-participant user study and simulated shoulder surfing attacks in a lab, which is in line with other shoulder surfing studies [7, 9]. Second, we present PrivacyScout, a novel computational method that leverages visual features extracted from an image of the potential attacker to then regress a score that represents the shoulder surfing risk. Results on regressing the shoulder surfing risk score for text and photos show that our method is the best performing in terms of the MAE and can more accurately asses the risk of a shoulder surfing attack. Our findings are significant in that they lay the next steps towards a new generation of mobile UIs that can protect the users' privacy and security in-situ, i.e. while an attack is happening.

2 Related Work

Prior works focused on (1) understanding shoulder surfing and its likelihood and (2) solutions to mitigate it.

2.1 Understanding Shoulder Surfing

Most work on understanding shoulder surfing investigated observing knowledge-based authentication methods, such as PINs, unlock patterns, and passwords. An early survey on shoulder surfing was conducted by Muslukhov et al. [31], where authors investigated user's concerns about unauthorized access threats through interviews and online surveys. In addition, researchers also relied on collective surveys to understand shoulder surfing behaviour and its likelihood. Eiband et al. conducted an online survey (N = 174) to understand how users perceive shoulder surfing events, from both the observer's and user's perspectives [15]. People who have been shoulder surfed reported negative feelings such as anger, embarrassment, and pressure. When asked about the content being observed, responses indicated that text, pictures, games and authentication credentials are the most shoulder surfed content types. In another large online study (N = 1173), authors used controlled video recordings of different authentication patterns and PIN entry methods to understand and compare the vulnerability of each of these methods against observations [7]. The videos featured combinations of different display sizes, phone grips (interaction with thumbs and index), angles (left and right) and distances (near and far). Their findings showed that PINs are less vulnerable to attacks, when compared to unlock patterns. Additionally, observation angles and distances affect shoulder surfing, as the input entry process might be occluded by the user's hand. The same team extended their work by comparing the shoulder surfing videos with actual live simulations [9]. In the simulations, the participants, representing the attackers, stood behind a sitting user (e.g. on their left and right). Although video recordings provide consistency among participants, the success of the observation attacks in replayed videos is not consistent with live settings [47]. Other studies focused on understanding the attacks' susceptibility, due to several reasons such as keyboard layout [41], or graphical passwords [14]. Another online survey was conducted by Harbach et al. [19] in which they investigated users' unlocking behaviours in correspondence to other phone usage interactions and the number of necessary unlock activities, in addition to users' perception of authentication activities posing privacy risks. A recent study by Saad et al. [38] recorded 360-degree staged videos of shoulder surfing scenarios in public transportation, and displayed them to participants in a VR headset. The participants' gaze behaviour indicated a tendency to observe the phone display. The authors concluded that short glances allow observers to disclose the content and impose privacy threats.

Most of the existing solutions were mainly focused on the authentication process, and the observers' ability to recreate the passcodes, while survey responses showed that the authentication is not the only interaction to be concerned about. Investigating how observations differ from one content to the other, how to empirically assess shoulder surfing events and their threat levels contributes to a more profound understanding of the events, and eventually deduce more robust solutions to mitigate them.

2.2 Solutions Mitigating Shoulder Surfing

There is a plethora of solutions designed to prevent shoulder surfers from observing others' displays during authentication, independent of physical hardware, such as privacy filters [34]. Some employed gaze for password entry [18, 29, 36]. Other works combined eye gaze with other approaches such as passwords, PINs, and mid-air gestures, for a more robust multimodal authentication [3, 23, 25]. Moreover, researchers investigated how effective patterns doodling is in overcoming shoulder surfing [50], while others added a force factor while authenticating, as it cannot be perceived by the attackers [28]. Image-based authentication schemes were also investigated as an alternative to traditional approaches [5, 42]. However, fewer works have focused on protecting the privacy of content beyond authentication. Saad et al. investigated communicating the attack events to the user [37]. In this work, the authors used the phone's front camera to detect additional faces in the scene, and notify the user once it happens. Related to this, Zhou et al. [52] designed and evaluated different shoulder surfing awareness and protection methods. This included dimming the screen, switching to greyscale colour, masking the content, and using on-screen cues to indicate shoulder surfing to the user. Similarly, Ali et al. proposed iAlert, an Android application that uses a linear regression model to notify the user with the risk of unauthorised visual access to the display [4]. Their model relied on the distance between the bystander's eyes, to infer the distance and angle between the potential attacker and the phone, derive the readability threat, and accordingly, notify the user that the on-screen text is vulnerable to observation. iAlert's performance was evaluated mainly based on the correlation between the bystanders' ability to read the text and the distance to the phone. The readability threat was based on a paper-based visual acuity test [1]. However, studies showed that that reading comprehension is better on paper than on screen [27]. Other works focused on specific content types. For example, von Zezschwitz et al. distorted sensitive pictures [45], while Eiband et al. customized fonts to be unreadable by bystanders [16]. Both approaches above make the content less clear to observers but understandable by users. Other approaches hide the content completely from shoulder surfers, such as EyeSpot [24] and PrivateReader [35]. Both systems use a filter to distort the on-screen content while the part the user gazes at, as determined via eye tracking, is revealed to the user. Chen et al. introduced HideScreen, a solution that uses the human optical system and its vision properties to prevent bystanders from observing other's screen [12]. Their solution, a visual grid added to the display, blends the displayed content with the background screen, only when viewed from a position outside the user's designated range.

Compared to prior work, our PrivacyScout is the first to estimate vulnerability to shoulder surfing in-situ based on real world data on the effectiveness of shoulder surfing. Unlike prior work that focused on a specific usage scenario [24, 35] or content type [16, 45], PrivacyScout is generic and is based on data about the susceptibility of generic text, photos, and authentication to shoulder surfing. These are the content types that are sensitive and were reported to be observed by shoulder surfers [15]. Saad et al. [37] assumed that the appearance of a face in the front-facing camera view means that the user is being shoulder surfed, and accordingly, the user should be notified of the incident. We significantly extend this concept and our understanding of shoulder surfing risks by relying not only on the presence of a face, but also 1) its distance to the screen, 2) the observation angle, 3) face features, and 4) empirical data about the vulnerability of the currently displayed content. This means that our PrivacyScout is able to more accurately determine events where the user is realistically subject to shoulder surfing risk. This, in turn, means that the user would receive less false alarms, which has positive implications on usability [40]. However, notifying the users or mitigating the shoulder surfing incidents is not within the scope of this work. To the best of our knowledge, our work is the first to empirically evaluate and quantify the success of shoulder surfing on different content types from different angles and distances.

3 A Model to Assess the Shoulder Surfing Risk

An ever-increasing number of mobile devices are readily equipped with high-resolution front-facing cameras. Using this camera has a high potential to alleviate the risk of shoulder surfing attacks by identifying people within the camera's field of view who may try to access the users' sensitive or private information. However, not all people captured by the front-facing camera are shoulder surfers. Some might simply be bystanders, or they might be too far away from the device to be able to perceive any on-screen content. Works that rely solely on detecting whether there is a secondary face in the image (e.g. [37]) will produce many false positives, i.e. cases identified as shoulder surfing attacks even if there was no risk for the user. An abundance of false positives results in alert fatigue which in turn makes it less likely users will take the warnings seriously, ultimately leading to less security [40]. Thus, any shoulder surfing assessment tool should alert the user only when a real danger is present. However, it is not clear when a user is particularly vulnerable to shoulder surfing.

To address this, we first conduct a study to understand the theoretical feasibility of a shoulder surfing attack. Our findings are based on a study conducted in a controlled environment, but still represents a best-case scenario for the attacker. Real-world situations might be more challenging due to occlusions caused either by the user, the way the user holds their device, or due to environmental conditions such as reflections on the device's screen which make accessing the screen's content difficult, especially outdoors. Based on the findings from our lab study, we investigated the feasibility of building a computational model that can assess the vulnerability of a shoulder surfing attack. PrivacyScout is a novel method that extracts visual features from the attacker's face, as captured by a wide-angle camera mounted on a mobile device to represent a front-facing camera, and regresses a shoulder surfing risk score – which is a direct measure to assess the potential risks. The risk score is dependent on the content type that is being shown. For shoulder surfing text the score's values range from 0 to 3 while for photos the score's range is between 0 and 4. A detailed description of how each score is calculated is available in section 6. Our assumption is that our system knows the type of content that is being shown on screen, which can be easily retrieved from the application or UI, and uses the corresponding model and score to assess the shoulder surfing risk.

In our threat model, the attacker attempts to shoulder surf the user while sensitive content is displayed on their smartphone. No cameras, mirrors or any other equipment is used. We assume there are no occlusions caused by the user's fingers due to the interaction. We also assume that the attacker is able to circumvent the occlusions caused by the user's head (e.g. by observing from a 30° angle).

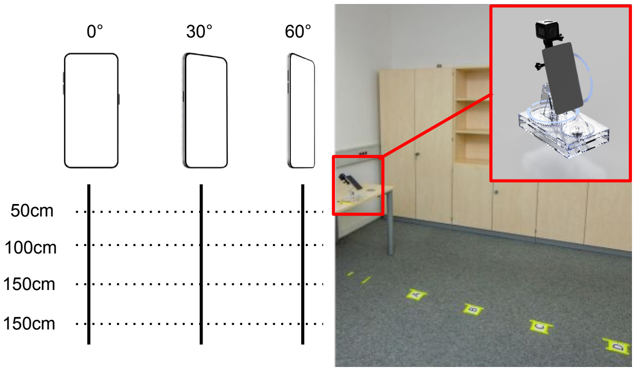

Fig. 1. Study setup: Participants attempted shoulder surfing text, photos, and videos of PIN entry at three different angles: 0°, 30° and 60° and at four different distances from the phone: 50 cm, 100 cm, 150 cm and 200 cm. We implemented a mechanical prototype to rotate the mobile device during the experiment.

4 Key Factors that Influence Shoulder Surfing Success

This study has two aims. The first aim is to collect empirical data on the effectiveness of shoulder surfing in different contexts. Namely, we consider different content types and study how well they can be observed when the attacker is at different distances and angles relative to the device. The analysis of this is reported in section 5. The second aim is to collect a dataset of images of the shoulder surfers' faces while they observe the content. This dataset is to be used subsequently to train a model that estimates the likelihood of a successful shoulder surfing attack. The model is described and evaluated in section 6. We define a successful attack as the event at which a bystander could observe *and* understand the content on the smartphone display. To this end, we designed a study in which participants played the role of a shoulder surfer and attempted to infer content that was on a mobile device screen in multiple shoulder surfing arrangements.

4.1 Study Design

Our study followed a within subjects design and had two independent variables:

IV1 – Shoulder Surfer's Location Relative to the Phone We experimented with several locations for the shoulder surfer relative to the mobile device being observed. We covered different observation angles and distances to the device. Our pilot tests showed that the minimum distance for a shoulder surfer to see the content and not interfere with the user is to stand at 50 cm from the phone, and 200 cm as a maximum distance for a person with normal or corrected-to-normal vision to perceive the content displayed. We chose four different distances between the participant and the phone: 50 cm, 100 cm, 150 cm, and 200 cm. Since shoulder surfing attacks do not exclusively take place when the phone is aligned to the observer [7], we experimented with three *angles* tilting around the y-axis: 0°, +30° and +60° (see Figure 1). These angles were chosen based on pilot tests in which we examined the angles from which content can reasonably be observed. We did not cover -30° and -60° because 1) shoulder surfing performance is expected to be similar across mirrored angles in our study setup where there are no occlusions or screen reflections, and 2) including them would have made the study excessively long, increasing the risk of skewing results due to participant fatigue. This resulted in twelve different locations (4 distances x 3 angles). The order of locations was counterbalanced using a Latin-square. Pre-defined locations were marked on the floor to facilitate guiding participants to the locations. To determine the orientation of the phone as it faces the user (the pitch axis), we ran a pilot test in which 20 participants reported the angles of their phone's tilt around the x-axis during every-day use. To this end, our pilot study participants used a web tool that shows the orientation angles in terms of Tait-Bryan angles (alpha, beta and gamma) [13] and reported the values to us. Using the mean beta value, we set the phone pitch angle to 40°.

IV2 – Content Type: At each location, we showed text, photos, and videos simulating PIN entry, as seen in Figure 2. These content types were chosen because they are the sensitive content types that are most commonly shoulder surfed [15]. For the text, we selected twelve phrases from the film review dataset by Pang et al. [33], designed for use in sentiment-analysis [32]. Text font size was 15 pts and was displayed on a 6.5" display and resolution: 1560 x 720 pixels. As for the photos, we selected twelve photos from the MPII human pose dataset [6]. The selected photos contained 2-4 people performing different activities. In all the selected photos, faces of the subjects were visible. Lastly, we recorded twelve videos of 4-digit Personal Identification Numbers (PINs) entry for authentication. The videos were recordings of the experimental assistant unlocking their mobile device. The duration of the entire video varied between 5 to 7 seconds. However, the actual pin entry was faster, around 2 seconds, simulating natural PIN entry behaviour. The phone was placed on a vertical stand, the PINs did not

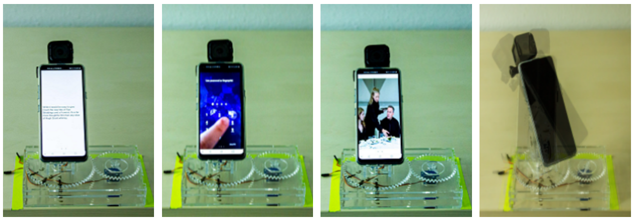

Fig. 2. Examples of different contents types on phone display: (left to right): text, PIN, photo. We used the mechanical prototype we implemented to rotate the phone on the yaw axis, to simulate shoulder surfing at 30° and 60°.

contain any duplicate digits, and the index finger was seen in the entry process (Figure 2). While the textual and photo contents were static and remained on display till the participant provided their input, the close-up video of a user's finger while entering a 4-digit PIN is played once, mimicking the real-life situation of a one-time PIN entry.

4.2 Apparatus

We built a mechanical prototype that holds a smartphone and allows rotating it about the Yaw axis (see Figure 2). This was done to control the viewing angles to simulate scenarios where the shoulder surfer is observing at the angles listed in subsection 4.1. The smartphone used was an LG smartphone (6.5" display, 1560px × 720px). As shown in the figure, in addition to holding the phone, the prototype also supports a GoPro Hero 4 session [21], with an UltraWide 170° Field of View (FOV), placed on top of the smartphone, maintaining an angle identical to the phone's. This camera was used to overcome the limited FOV of smartphone front-facing cameras (77° for the LG smartphone).

4.3 Participants

We recruited 16 participants (15 Males, 1 Female), aged between 22 and 34 years (M = 26.25, SD = 3.82) through mailing lists and word of mouth. Their heights ranged between 169 cm and 202 cm (M = 184.37 cm, SD = 8.21 cm). All participants reported having a normal or corrected-to-normal vision. Participation in the user study was voluntary.

4.4 Procedure

Upon arriving in our lab, we explained the study to the participant, and asked them to sign a consent form and fill a demographics questionnaire. Participants then had to scan a QR code using their personal mobile devices to access an online form in which they can indicate their guesses after each shoulder surfing event. The participant then traversed the different experimental conditions through 12 shoulder surfing events. In each event, the experimenter remotely guided the participant to the location they should stand in, which was marked on the floor, e.g. "Please stand in the spot that is marked A, and look towards the phone, your current location is A1", where A1 corresponds to a distance of 50 cm and an observation angle of 0°. All participants stood exactly at these predefined points, independent of their heights, unlike the work of Aviv et al. [9], where participant did not stand in the exact positions, depending on their heights. The content was then displayed on the phone using a PowerPoint presentation that was remotely controlled by the experimenter. Participants were asked to observe the content, and respond to questions in the online form they accessed via their personal phones earlier. The experimenter clearly explained that the answers provided by the participants should correspond to their perception (what do you see?), not to the actual content (what is actually there?). Participants had no incentive to lie, as there was no reward for correctly guessing the content. After observing text, participants were asked whether they can see individual words from the text (Yes/No), whether they understand what the text is about (Yes/No), and were asked to specify the names mentioned in the text (open text). As for the photos, participants were asked whether they can tell the photo is in an indoor or an outdoor setting (indoor/outdoor), how many people they see in the photo (open text), what the activity of the people in the photo was (open text), and whether they can see the facial expressions of the people in the photo (Yes/No). When observing videos of PIN entry, participants were asked to indicate the 4-digit PIN they had just seen (open text), and the Yes/No question: "If you cannot enter the full PIN, could you make out parts of the PIN?". We opted for yes/no questions rather than open text ones for content elements that are subjective. For example, had we asked "What are the facial expressions of the people in the photo", responses may have been very different across participants (e.g. grinning vs smiling vs laughing). For the purpose of building a risk assessment model, we assume it is a sufficient threat to privacy if the shoulder surfer can confirm seeing the facial expression, regardless of what they subjectively think the facial expression is.

The participants were asked to answer the questions carefully and to take as much time as they need. In each shoulder surfing event, we logged two photos: one using the GoPro and one using the front-facing smartphone camera. Samples of the photos from each angle and distance are shown in Figure 10. Social distancing regulations and COVID-19 precautions measures were abided to as per the current regulations at the time of running the study. Our study conformed with the ethics regulations of our university.

4.5 Analysing Responses to Open-ended Questions

To analyse the participants' responses, two researchers independently rated the accuracy of the responses provided by participants to the open-ended questions as either: Successful attack, partially successful attack, unsuccessful attack. For example, responses that specify only one of two names mentioned in the shoulder surfed text, or identified a secondary rather than a primary activity by people in a photo, were considered partially correct. The researchers then discussed any inconsistent ratings to agree on a final rating. This method is inspired by previous work on analysing guesses after observing obfuscated photos [46]. Analysing responses to Yes/No questions was trivial. As for the accuracy of shoulder surfing PINs, we measured the Levenshtein distance between the participant's guess and the original PIN, as we detail in section 5. This was inspired by a plethora of work that utilised the Levenshtein distance to evaluate the accuracy of shoulder surfing attacks [26, 44].

5 Assessing Vulnerability to Shoulder Surfing on Mobile Devices

We report the findings of our user study (section 4) according to the three different content types (PINs, photos, and text) used to assess the vulnerability to shoulder surfing on mobile devices.

5.1 PINs

In the PIN condition, study participants that played the role of an attacker were asked to reproduce the PIN that was shown on screen at different distances and viewing angles. In addition, if they were not able to fully observe the PIN, participants were asked to report whether they could make out parts of the PIN, thus also incurring a potential risk. Based on these two questions, we derived three different evaluation metrics. The *edit distance* (or Levenshtein distance) represents the similarity between the actual, ground truth PIN and the one observed by the study participants. It is a distance metric between two input strings that associates a cost to single-character edits: insertion, deletion, substitution, and also transposition. As such, the edit distance is a metric that indicates how close the PIN observed by the attacker is to the actual one. A score of 0 indicates that the two PINs are identical, while a score of 4 (the PIN's length) means the attacker cannot observe any of the digits. Overall, from all the observations independent of location (a combination of distance and angle), the mean edit distance was 0.73 (SD=1.24).

Figure 3 shows the results per observation location, which is a combination of distance and angle, averaged per study participant. We conducted a two-way repeated measures ANOVA to analyse the effects of distance and angle on the average edit distance. Our analysis did not reveal any statistically significant results, neither of the main effects of *distance* and *angle* nor interaction effects. For the complete statistical results, please consult the appendix (subsection A.1).

5.2 Photos

To assess the vulnerability of photos during shoulder surfing attacks on mobile devices, we calculated five different evaluation metrics: *success rate indoors vs. outdoors*, *success rate count people*, *success rate facial expressions*, *success rate activity*, and *partial success rate activity.* We first evaluated the mean probabilities across study participants independent of observation location. Overall, the mean probability for successfully observing whether the photo was indoors or outdoors was 98.96% (SD=10.15%). The mean probability for correctly counting how many people were in the photo was 96.35% (SD=18.74%). The mean probability for observing the facial expressions of the people in the photo was 51% (SD=50%). The mean probability for correctly observing what people were doing, i.e. their activity, was

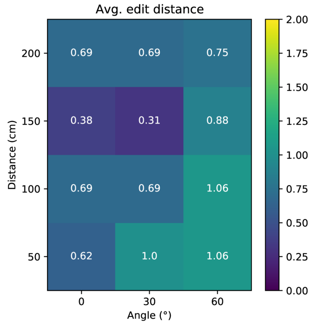

Fig. 3. Analysis of shoulder surfing vulnerability on mobile devices for the PIN condition by location (distance and angle). Average edit distance between actual, ground-truth PIN and the observed PIN. The minimum edit distance is 0, indicating that the attacker guessed all digits in the correct position, whereas the highest is 4, indicating that none of the guessed digits are correct. The lower the value, the more successful the shoulder surfing attack. The differences between the different angles and distances were not statistically significant. The results suggest that attacks are highly successful from all angles/distances.

78.65% (SD=40.98%). Finally, the probability for partially observing what the people were doing was 14.06% (SD=26%).

Figure 4 shows the average probabilities for each location and evaluation metric. We then analysed the influence of the main effects, distance and angle, and possible interaction effects, distance*angle, on each of the five evaluation metrics. Using a two-way repeated measures ANOVA, we found that the influence of distance on *success rate facial expression* ($F_{3,45} = 38.10$, $p < 0.01$, $\eta_p^2 = 0.72$) and *success rate activity* ($F_{3,45} = 4.10$, $p = 0.011$, $\eta_p^2 = 0.21$) was statistically significant. The influence of distance was not significant on the remaining metrics. The angle played a significant role on *success rate count people* ($F_{2,30} = 5.74$, $p < 0.01$, $\eta_p^2 = 0.28$) and *success rate facial expressions* ($F_{2,30} = 33.42$, $p < 0.01$, $\eta_p^2 = 0.69$). We did not find any statistically significant influence of angle on the other three metrics. Looking at the interaction effects between distance and angle, we did not find any statistically significant influence on neither of the five metrics. For the complete statistical results, please see the appendix (subsection A.2).

Figure 5 shows the influence of the distance on the evaluation metrics on which results were statistically significant: *success rate facial expressions* and *success rate activity*. Pairwise post-hoc Tukey HSD tests showed that the differences between all the distances except 150 cm and 200 cm ($p = 0.21$) and their influence on *success rate facial expressions* were statistically significant at the $p < 0.05$ level. Regarding the *success rate activity*, pairwise test showed that only the difference between 50 cm vs. 200 cm ($p = 0.01$) and 100 cm vs. 200 cm ($p = 0.02$) were significant at the $p < 0.05$ level.

Figure 6 shows the influence of the angle on *success rate count people* and *success rate facial expressions* – the metrics we found to be significantly influenced by the angle. On *success rate count people*, post-hoc Tukey HSD tests showed differences between 0° vs. 60° and between 30° vs. 60° ($p < 0.05$). On *success rate facial expressions*, we found pairwise differences between 0° vs. 30° and between 0° vs. 60° ($p < 0.05$).

5.3 Text

We assessed the vulnerability to shoulder surfing of textual content on mobile devices using four evaluation metrics: *success rate recognise words*, *success rate understand context*, *success rate retrieve names*, and *partial success rate retrieve names*. We first evaluated the mean probabilities across study participants independent of observation location. Overall, study participants were able to recognise words shown on screen with a mean probability of 71.35% (SD=45.21%). On average, 57.81% (SD=49.39%) were able to understand the general theme or context of the text shown on the mobile device's screen. In 57.81% (SD=49.39%) of the total cases, study participants were able to correctly recognise all the names present in the text. Participants partially retrieved names from the text in 6.25% (SD=24.21%) of the cases.

Figure 7 shows an overview of the four evaluation metrics according to the location, which is a combination of distance and viewing angle relative to the mobile device. We analysed the influence of the main effects, distance and angle, and possible interaction effects, distance*angle, on each of the four evaluation metrics. A two-way repeated measures ANOVA showed that the distance significantly influenced *success rate words* ($F_{3,45} = 40.87$, $p < 0.01$, $\eta_p^2 = 0.73$), *success rate understand* ($F_{3,45} = 72.03$, $p < 0.01$, $\eta_p^2 = 0.83$), and *success rate names* ($F_{3,45} = 36.55$, $p < 0.01$, $\eta_p^2 = 0.71$). Similarly to the distance, the angle also had a significant influence on the same factors: *success rate words* ($F_{2,30} = 50.45$, $p < 0.01$, $\eta_p^2 = 0.77$), *success rate understand* ($F_{2,30} = 79.55$, $p < 0.01$, $\eta_p^2 = 0.84$), and *success rate names* ($F_{2,30} = 30.27$, $p < 0.01$, $\eta_p^2 = 0.67$). Besides simple main effects, we also identified interaction effects of distance*angle on the same three factors as follows: *success rate words* ($F_{6,90} = 4.97$, $p < 0.01$, $\eta_p^2 = 0.25$), *success rate understand* ($F_{6,90} = 3.94$, $p < 0.01$, $\eta_p^2 = 0.21$),

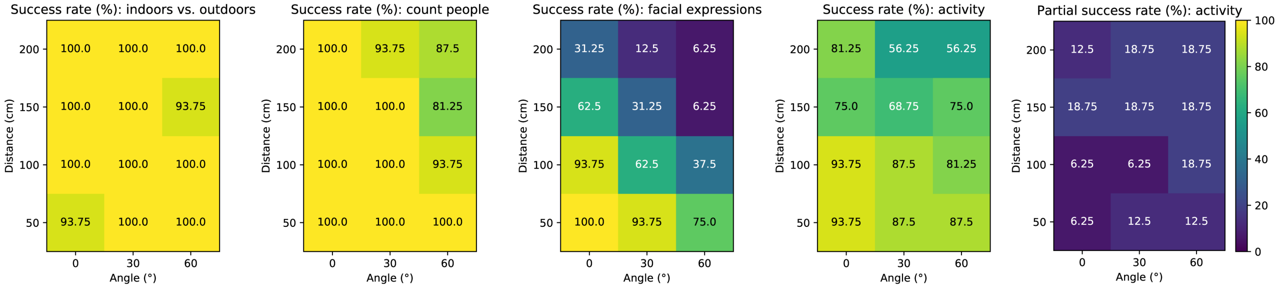

Fig. 4. Analysis of shoulder surfing vulnerability on mobile devices for the photo condition by location (distance and angle). From left to right: (1) The probability of correctly observing whether the photo was indoors or outdoors, (2) the probability of correctly counting how many people were in the photo, (3) the probability of observing the facial expressions of the people in the photo, (4) the probability of correctly observing what the people in the photo were doing, and (5) the probability of only partially observing the activity.

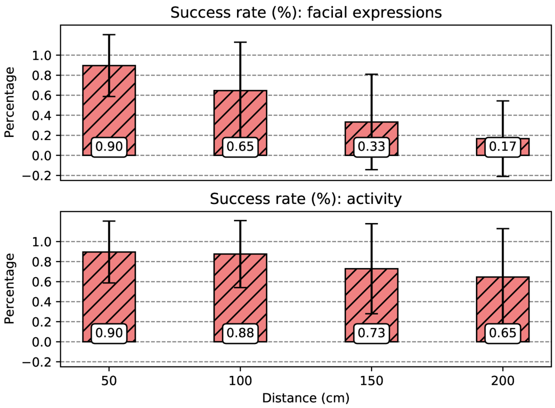

Fig. 5. The influence of *distance* on the two evaluation metrics *success rate facial expressions* (top) and *success rate activity* (bottom). The bars represent the mean and the error bars the standard deviation of all the samples at a specific distance.

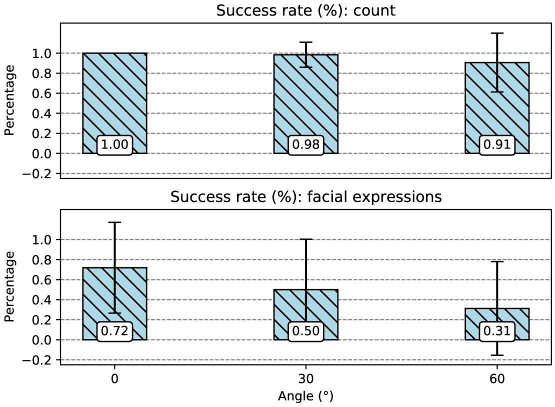

Fig. 6. The influence of the observation angle on *success rate count people* (top) and *success rate facial expressions* (bottom). The bars represent the mean and the error bars the standard deviation of all the samples at a specific angle.

and *success rate names* ($F_{6,90} = 2.69$, $p = 0.02 < 0.05$, $\eta_p^2 = 0.15$). We did not find any statistically significant influence of distance, angle , or distance*angle on *partial success rate names*. Complete statistical results are available in the appendix (subsection A.3).

We further analysed the influence of distance on the three evaluation metrics that were significantly influenced by it (Figure 8). Pairwise post-hoc Tukey HSD tests show that most differences were statistically significant at the $p < 0.01$ level except for 50 cm vs. 100 cm ($p = 0.38$) - this on neither of the three evaluation metrics, *success rate words* - $p = 0.38$, *success rate understand* - $p = 0.14$, and *success rate names* - $p = 0.33$). In addition, the difference between 150 cm and 200 cm was not significant on *success rate names* ($p = 0.08$).

Figure 9 shows the influence of the observation angle on the three evaluation metrics: *success rate words*, *success rate understand*, and *success rate names*. Pairwise post-hoc Tukey HSD tests revealed that the difference between 0° and 60°, and between 30° and 60° were significant at the $p < 0.01$ level on all metrics. We did not find the difference between 0° and 30° to be statistically significant on neither of the metrics (*success rate words* - $p = 0.15$, *success rate understand* - $p = 0.60$, and *success rate names* - $p = 0.20$)

6 PrivacyScout: A Method to Assess Vulnerability to Shoulder Surfing on Mobile Devices Using Front-Facing Cameras

Inspired by our analysis in section 5, we investigated the feasibility of assessing the shoulder surfing risk using the

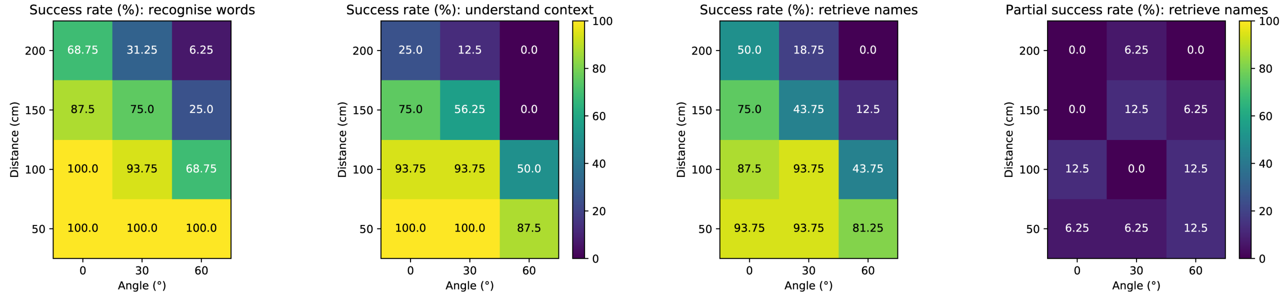

Fig. 7. Analysis of shoulder surfing vulnerability on mobile devices for the text condition by location (distance and angle). From left to right: (1) the probability of correctly recognising words from a piece of text, (2) the probability of understanding the general theme or context of the text, (3) the probability of correctly retrieving all the names from the text, and (4) the probability of partially retrieving *some* names from the text.

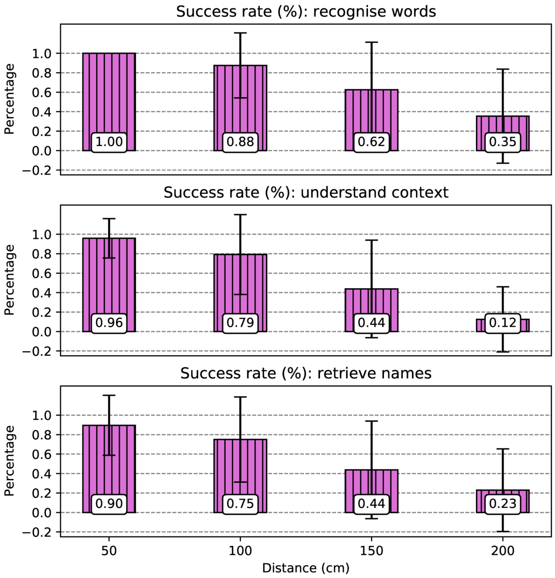

Fig. 8. The influence of the observation *distance* on three evaluation metrics. From top to bottom: (top) the probability of recognising words, (middle) the probability of correctly understanding the general theme or context, and (bottom) the probability to correctly retrieve names from the text. The bars represent the mean and the error bars the standard deviation of all the samples at a specific distance.

front-facing camera of current mobile devices as well as the potential for wide-angle cameras, in our case using a GoPro Hero 4 Session – the same cameras that were used during the data collection study (section 4).

6.1 Method Overview

We explored and developed PrivacyScout, a novel method that can predict the risk of a shoulder surfing attack according to the camera's view and the type of content that is shown on screen. PrivacyScout extracts features from an image of the attacker, which are then used to regress a risk score. We formulated this task as a regression problem, where we built and tested a model for two of the three content types studied in section 5. We opted to study only *text* and *photos* as the likelihood of a shoulder surfing attack was shown to vary for these content types not only with distance but also observation angle relative to the mobile device. We excluded the *PIN* condition, since our analysis revealed that PINs are highly vulnerable independent of the two factors.

Our method takes as input an image captured with the front-facing camera. We then used the open-source, publicly available framework OpenFace [11] to detect the person's face and extract a number of features. OpenFace provides basic information of the detected face (e.g. the confidence value), 2D and 3D facial landmarks [10, 49], the gaze direction vectors and angles as obtained using a gaze estimation model [48], or the user's head pose expressed as three values for head rotation and three for head translation.

Using OpenFace, we extracted the following face image features:

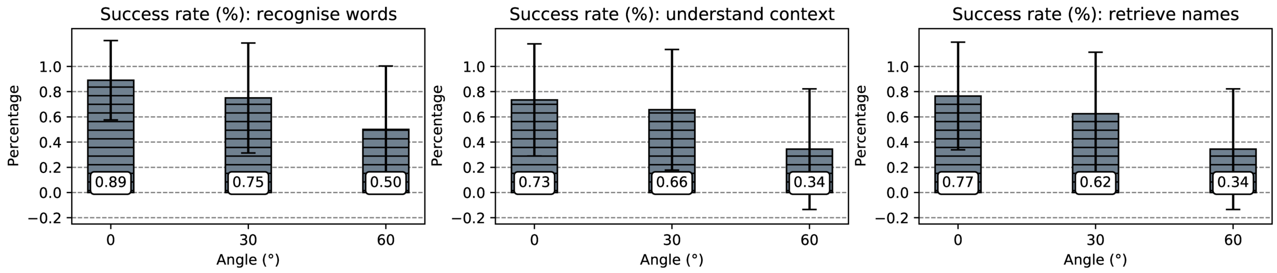

Fig. 9. The influence of the observation *angle* on three evaluation metrics. From left to right: (1) the probability of recognising words, (2) the probability of correctly understanding the general theme or context, and (3) the probability to correctly retrieve names from the text. The bars represent the mean and the error bars the standard deviation of all the samples at a specific angle.

1. **face center x and y**: The position of the face's centre in pixels calculated as the average of the four eye landmarks corresponding to the corners of the two eyes and two landmarks corresponding to the corners of the mouth, similar to prior work on gaze estimation [51].
2. **midpoint x and y**: The midpoint between the two eyes, inspired by prior work from Ali et al. [4].
3. **face width** and **height**: They are calculated as the absolute difference between the minimum and maximum x or y coordinate of all the 68, 2D facial landmarks provided by OpenFace. The values are both expressed in pixels.
4. **face size**: Calculated as the width multiplied by height, expressed also in pixels.
5. **distance between eyes**: The distance between the centre of the two eyes, as proposed by Ali et al. [4].
6. **pose Tx, Ty, Tz**, and **pose Rx, Ry, Rz**: The estimated head pose relative to the camera expressed as six values, three for head translation and three for head rotation. The head pose rotation is expressed in radians, while the head translation in millimetres.
7. **gaze angle x and y**: The gaze direction in the horizontal and vertical direction in world coordinates, expressed in radians. These are average values for both eyes.

We used these features to train support vector regressors (SVRs).

6.2 On Detectability of Faces Using the Front-Facing Camera

A prerequisite to any method that relies on the front-facing camera of a mobile device is the ability to detect the users' and/or the attacker's face in the captured image. As shown in prior work, in mobile settings, it is often the case that the face and the facial landmarks are either outside the camera's field of view or only partially visible [22]. Because of this, in this experiment, we evaluated the face detection method that is available in the OpenFace [11] framework on both images captured with the phone's front facing camera and those captured with the GoPro, which has a wider field of view. A reasonable assumption is that front-facing cameras of future mobile devices will have wider lenses. Figure 10 shows sample images from one study participant who agreed to publicly share their data.

For the images captured with the GoPro, OpenFace was able to detect and extract features from 142 photos out of 192 (12 locations × 16 participants). For images captured with the mobile phone, OpenFace was able to detect a face and extract features from 67 photos out of 168 (12 locations × 14 participants – for 2 participants of the 16, we were unable to use their data due to an issue with the data recording software). Figure 11 shows a detailed breakdown of the percentage of images in which a face can be detected for the GoPro and for the mobile phone's camera. Our analysis shows that, using images from the GoPro, a face can always be detected if the person is up to 100 cm away independent of the observation angle. In contrast, using the phone's camera, a potential attacker's face could reliably be detected only when the observation angle was 0° or when the distance was 50 cm and the observation angle was 30° (around 78% detection rate).

Because of the limited field of view of **current** front-facing cameras on mobile devices, in the evaluations that follow, we only use the data captured by the GoPro. Current-generation mobile devices such as the Apple iPhone 12 already come equipped with wide angle lenses. Therefore, we believe that such technological advancements are also possible for the front-facing camera.

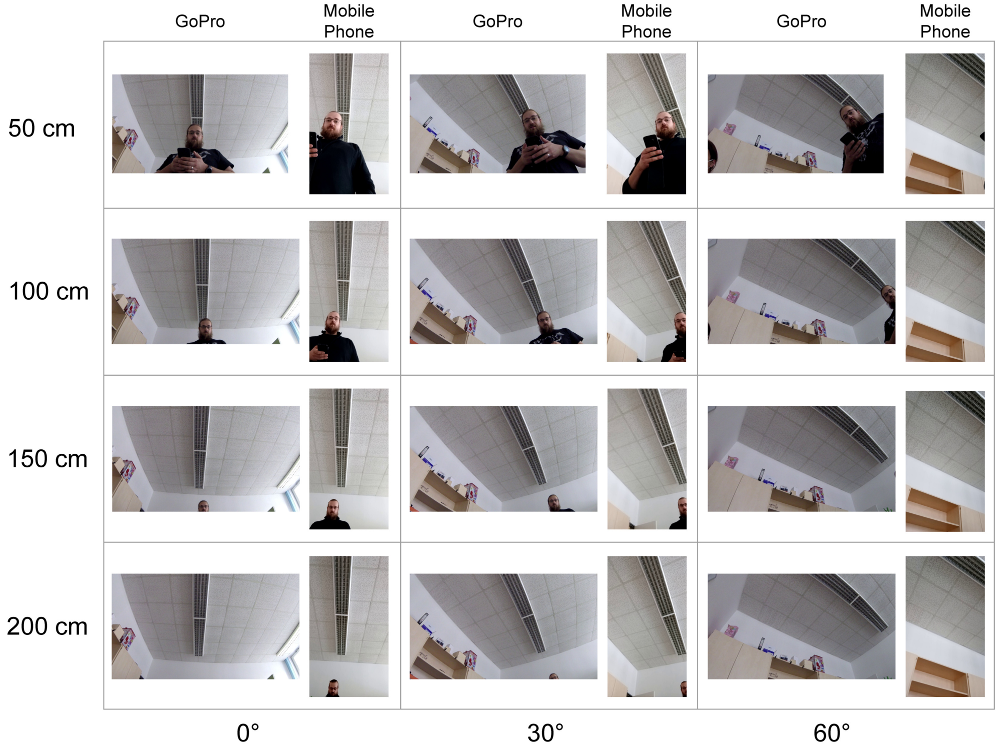

Fig. 10. Sample images from one of the participants who agreed to publicly share their data. Images captured with both the GoPro and the front-facing camera of a mobile phone at the four distances (50, 100, 150, and 200 cm) and three viewing angles relative to the device (0°, 30°, and 60°).

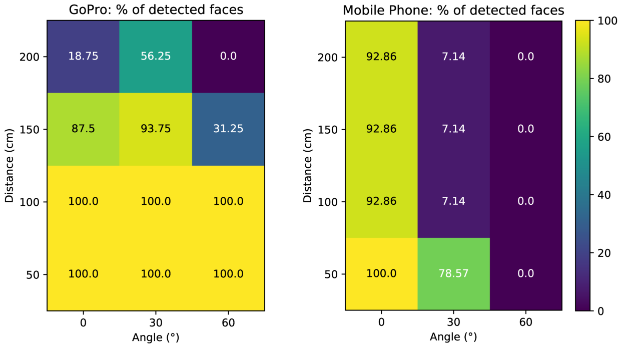

Fig. 11. The percentage of images from the user study in which OpenFace [11] could detect faces and extract facial features by distance and observation angle.

6.3 Evaluation Results

To build our regression models, one for *text* and one for *photo*, we first had to derive a score that quantifies the shoulder surfing risk. Based on the analysis from section 5, we derive each score as the sum of the individual characteristics for each content type. Concretely:

$$\begin{aligned} score_{text} = {} & success\_rate_{words} + success\_rate_{context} \\ & + success\_rate_{names} \end{aligned} \tag{1}$$

where $success\_rate_{words}$ represents the score for correctly recognising words in a piece of text, $success\_rate_{context}$ is the score for correctly understanding the context, and $success\_rate_{names}$ is the score for correctly retrieving names from the text. Each individual score has a value of either 0 or 1, so the minimum overall value for $score_{text}$ is 0 and the maximum is 3.

Similarly, we defined a score for the photo condition:

$$\begin{aligned} score_{photo} = {} & success\_rate_{indoors\ vs.\ outdoors} \\ & + success\_rate_{count\ people} \\ & + success\_rate_{facial\ expressions} \\ & + success\_rate_{activities} \end{aligned} \tag{2}$$

where $success\_rate_{indoors\ vs.\ outdoors}$ is the score for correctly recognising if the photo was taken indoors or outdoors, $success\_rate_{count\ people}$ is the score for correctly counting how many people were in the photo, $success\_rate_{facial\ expressions}$ is the score for being able to see the persons' facial expressions, and $success\_rate_{activities}$ is the score for correctly recognising what the people were doing (i.e. their activity). Each individual score has a value of either 0 or 1, so the minimum overall value for $score_{photo}$ is 0 and the maximum is 4.

Using the features described in subsection 6.1, we derived several feature sets, trained multiple regression models, and evaluated their performance. In addition, we also implemented three naïve baseline methods. All methods were trained in a leave-one-subject-out cross validation, which means that each model was trained on the data from N-1 participants and evaluated on the remaining one. We report the mean absolute error (MAE), which is commonly used in regression tasks. All these models were trained only on the images collected with the GoPro since in many of the images captured using the mobile phone, no face can be detected.

We implemented different versions of PrivacyScout that use different feature sets.

- Face position: SVR that uses only the face position (x and y) in pixels as features.
- Face position and size: SVR that uses the face position (x and y) and face size in pixels as features.
- Head pose: SVR that uses the six values of the head pose (three head rotation, three head translation) as features.
- Gaze angles: SVR that uses the two gaze angles as features.
- All: SVR that uses all of the above features.

In addition, we implemented six baselines: three inspired by prior work [4] and three naïve baselines.

- Midpoint between eyes: SVR that uses the reference midpoint between the eyes (x and y) in pixels as features.
- Distance between eyes: SVR that uses the distance between the eyes in pixels as feature.
- Both: SVR that uses both the midpoint (x and y) and the distance between eyes as features.
- Naïve mean: A method that predicts the mean value from the training set.
- Naïve median: A method that predicts the median value from the training set.
- Naïve constant: A method that always predicts a constant value. We pick the maximum value for both conditions, i.e. 4 for photo and 3 for text.

Table 1 shows the results of the evaluation of PrivacyScout when trained for the text condition. The best performing method is the one that uses all the features as provided by OpenFace with a mean MAE of 0.56 (SD=0.38) vs. 0.65 (SD=0.37) for *Face position*, 0.62 (SD=0.37) for *Face position and size*, 0.75 (SD=0.40) for *Head pose*, and 0.60 (SD=0.38) for Gaze angles. The three baselines achieved 0.65 (SD=0.37) for *Midpoint between eyes*, 0.70 (SD=0.39) for *Distance between eyes*, and 0.61 (SD=0.39) for *Both.* Results from *Face position* and *Midpoint between eyes* are identical since the features, i.e. the x and y positions, are highly correlated (r>0.99). The naïve methods achieved a mean MAE of 0.74 (SD=0.58) for both *median* and *constant*, and 1.0 (SD=0.28) for *mean.* The median and the constant naïve baselines are the same because, in this case, the median value was the same as the constant we set, which was 3 for text – the same as the maximum value of the score. A one-way repeated measures ANOVA showed that the difference between conditions was significant ($F(10, 150) = 8.66$, $p < 0.01$, $\eta_p^2 = 0.37$). Post-hoc tests did not reveal any statistically significant differences. This might be due to the limited amount of data used in the evaluation.

Table 1. Evaluation results from leave-one-subject-out cross validation for regression of the shoulder surfing risk score for **text**. The risk score is a value between 0 and 3. The best performing method is underlined.

Method	Mean MAE	Std. dev. of the MAE
Face position	**0.65**	**0.37**
Face position and size	**0.62**	**0.37**
Head pose	**0.75**	**0.40**
Gaze angles	**0.60**	**0.38**
All	**0.56**	**0.38**
Midpoint between eyes	**0.65**	**0.37**
Distance between eyes	**0.70**	**0.39**
Both	**0.61**	**0.39**
Naïve mean	**1.00**	**0.28**
Naïve median	**0.74**	**0.58**
Naïve constant	**0.74**	**0.58**

Table 2 shows the results of the evaluation of PrivacyScout when trained for the photo condition. The best performing method is the one that uses the *Gaze angles* with a mean MAE of 0.41 (SD=0.14) vs. 0.42 (SD=0.13)

Table 2. Evaluation results from leave-one-subject-out cross validation for regression of the shoulder surfing risk score for **photos**. The risk score is a value between 0 and 4. The best performing method is underlined.

Method	Mean MAE	Std. dev. of the MAE
Face position	0.42	0.13
Face position and size	0.42	0.13
Head pose	0.44	0.12
Gaze angles	0.41	0.14
All	0.42	0.13
Midpoint between eyes	0.42	0.13
Distance between eyes	0.46	0.16
Both	0.42	0.13
Naïve mean	0.61	0.08
Naïve median	0.56	0.23
Naïve constant	0.56	0.23

for *Face position*, 0.42 (SD=0.13) for *Face position and size*, 0.44 (SD=0.12) for *Head pose*, and 0.42 (SD=0.13) for *All*. The three baselines achieved 0.42 (SD=0.13) for *Midpoint between eyes*, 0.46 (SD=0.16) for *Distance between eyes*, and 0.42 (SD=0.13) for *Both*. However, all methods appear to perform similarly. The naïve methods achieved a mean MAE of 0.56 (SD=0.23) for both *median* and *constant*, and 0.61 (SD=0.08) for *mean*. The median and the constant naïve baselines are, just like in the previous experiment, the same since the median value was the same as the constant we set, which was 4 for photo – the same as the maximum value of the score. A one-way repeated measures ANOVA shows that the difference between conditions was significant ($F(10, 150) = 6.53$, $p < 0.01$, $\eta_p^2 = 0.30$). Post-hoc Tukey HSD tests showed significant differences between *Face position*, *Face position and size*, *Gaze angles*, *All*, *Midpoint between eyes*, and *Both* vs. *Naïve mean*. All other pairwise differences were not statistically significant at $p < 0.05$.

Robustness to Limited Training Data. Our previous evaluations relied on images from all the distances and viewing angles to train models. However, in practice, it is possible that images of bystanders at 0° are difficult to capture due to the user blocking the camera's field of view. Therefore, we also conducted a preliminary evaluation of all our models trained on images only from the 30° and 60° angles from all distances. We thus reduced the training set and kept the test set the same, i.e. including all distances and viewing angles.

Results show that the methods in both the photo and text condition suffer from a decrease in performance, e.g. for the text condition, the performance of the *All* method decreased from 0.56 (SD=0.38) to 0.69 (SD=0.30). For the photo condition, the performance of the *Gaze angles* method decreased from 0.41 (SD=0.14) to 0.56 (SD=0.14). Detailed results of this evaluation are available in the appendix (Appendix B: Table 3 and Table 4). We hypothesise that this difference is also due to the limited amount of training data available. For the previous evaluation, each leave-one-subject-out cross-validation fold had, on average, 133.12 (SD=1.17) images for training and 8.88 (SD=1.17) for testing. In this evaluation, by removing the 0° condition, we were able to use, on average, 87.19 (SD=0.73) images for training.

7 Discussion

7.1 The Factors That Influence Shoulder Surfing on Mobile Devices

In this work, we first started by analysing whether the location of the observer has an influence on the success of a shoulder surfing attack. Our analysis (section 5) covers three content types that have been shown to be the most likely to be attacked in everyday situations [15]: text messages, photos of people, and PIN authentications. PINs are still some of the most widely used methods for user authentication [2]. A significant body of research has tried to develop not only ways to better understand when such attacks occur [7], but also methods to mitigate them [28]. Previous similar work, such as iAlert, focused on detecting a bystander, and analysing their position to infer observation attacks' likelihood, and notify the user accordingly [4]. On the other hand, Hidescreen analysed human vision to create a grid that creates a blurry display for people outside the main user's viewing range [12]. In contrast to prior work, our method makes a significant step towards better understanding whether the distance to and the observation angle relative to the mobile device play a key role in shoulder surfing PINs, in normal setups and without display modifications. Our analysis showed that **PINs are highly vulnerable independent of the observer's location**. The average edit distance between the real, ground truth PIN and the observed one was 0.73 (SD=1.24). This means that, on average, less than one operation was necessary to retrieve the correct PIN. We believe that PINs are more vulnerable since observers may follow the users' finger and, having a mental model of the digits' layout, they can easily decode what the PIN was even without fully observing it. As an extension to our work, it would be interesting to study whether other authentication methods (e.g. pattern based [7, 8]) are also as vulnerable as PINs. In addition, real-world situations may present different challenges than lab settings. For example, due to the way users hold their device or occlusions, the success rate of a shoulder surfing attack might be different.

When looking at shoulder surfing attacks targeting text and photos of people, our results (Figure 4, Figure 7) showed that the observers' location played a significant role but only in certain situations. Correctly observing how many people were present in a photo or whether they were indoors or outdoors was possible with a high average probability across locations and distances (~96% and ~99%). On the other hand, correctly observing finer details such as facial expressions was possible with a high probability when physically close to the device (less than 100 cm) and at most at a 30° angle. The distance to the device also plays a role when observing **what** people are doing (Figure 5). The analysis of textual content showed even more interesting results. Correctly recognising words, understanding the general context, or retrieving names (Figure 7) depends on the observers' distance (Figure 8) and angle relative to the mobile device (Figure 9).

Based on our analyses, we provide evidence that shows the potential for a (real-time) shoulder surfing risk assessment tool. Our work demonstrates that **not all locations are equal, and neither are the content types**. Consequently, our work directly shows it is possible to differentiate bystanders from actual shoulder surfers, in particular for text and photographic content.

7.2 Towards Real-Time Assessment of Shoulder Surfing Risk

Guided by our analysis on the factors that influence shoulder surfing on mobile devices, we proposed PrivacyScout– a novel method to assess the shoulder surfing risk using visual features extracted from the observer's face captured by the (potentially wide-angle) camera of a mobile device. Table 1 and Table 2 show the results of our methods using different feature sets: from simple statistics of the user's face to more complex head pose and/or gaze estimates.

For the text condition, Table 1 shows that the best performing method was the one that used *all* the features. This induces a higher computational cost as the method will have to first detect the face, then extract the facial landmarks, estimate the users' head pose, and, finally, estimate the users' gaze direction. In contrast, for the *photo* condition, all methods performed similarly. Therefore, using the feature set that only requires the x and y position of the face will have the lowest computational cost. Such a method can already run in real time on current mobile devices [20].

Another complementary dimension to computational cost is energy efficiency. Methods to run and process the live feed of a camera will deplete the battery of the phone quickly. Therefore, future work should also investigate when to assess the shoulder surfing risk or when to optimally use the camera through e.g. opportunistic sensing [43].

7.3 Applications of PrivacyScout

We envision multiple ways in which PrivacyScout can be employed. Mobile applications that deal with sensitive data, such as mobile banking apps or health apps, may integrate our models to determine if a shoulder surfing attack is likely. Alternatively, PrivacyScout can be integrated into operating systems to function as a service across all mobile applications. Once a shoulder surfer is detected, the application may then deploy privacy protection mechanisms from prior work by, for example, blurring or hiding the sensitive content [24, 37, 45, 52].

Alternatively, PrivacyScout could warn the user of the threat to privacy. This can be done, for example, by showing a warning message or an alert icon [37, 52]. However, warnings should not be excessive lest they result in habituation [40], which in turn leads to users dismissing warnings without paying attention, putting them at risk. We discuss the usability and social implications further in the next section.

7.4 Usability and Social Implications of Shoulder Surfing Mitigation

It is important to consider the usability and social implications of consistent warning. Previous work showed that reacting in an overt way to a shoulder surfing situation may damage the relationship between the user and the shoulder surfer [17], who may be friends, family members or partners [15, 30]. In fact, a study by Farzand et al. [17] showed that users prefer to use cover methods for mitigating shoulder surfing, and invest time and effort in making sure the shoulder surfer does not know that they were caught. It is important to note that shoulder surfers are not necessarily the user's enemies; shoulder surfing is in many cases unintentional or done out of boredom rather than being motivated by malicious intentions [15]. Some recorded shoulder surfing incidents even involved children shoulder surfing their older siblings or parents out of curiosity rather than malicious intentions [17].

The usable security community has emphasised the importance of usability and social acceptability in security and privacy enhancing systems. A security system that is not user-centred will either be misused or avoided by users, resulting in less overall security. By predicting the shoulder surfing risk more accurately, our PrivacyScout advances the usability of state-of-the-art shoulder surfing detection methods [37] that relied on the presence of a face to assume shoulder surfing risk.

7.5 Limitations and Future Work

Our work makes significant strides towards better assessing and quantifying shoulder surfing risk on mobile devices. However, our work also has some limitations.

First, shoulder surfing attacks are influenced by a number of factors, such as the position of the mobile phone user (e.g. standing vs. sitting), the position of the shoulder surfer, their heights, the way users might hold their mobile device, the size of the users' hands, and even the size of the screen. As such, all these factors provide an endless number of possibilities for the assessment of shoulder surfing risks. In our work, we made several assumptions that we validated in a pilot study (e.g. the orientation of the phone) and provided quantitative evidence that the observation angle and distance from the mobile device influences the success of shoulder surfing.

Second, our evaluation of PrivacyScout assumed an ideal scenario for the attacker, without any occlusions caused by the mobile phone users' head, torso, or fingers. This experimental setup also allowed us to collect a diverse training set for our models covering all distances and angles. By simulating the presence of a user and removing images from the 0° angle, our method's performance deteriorated (subsection 6.3). Future work should investigate ways to increase the robustness of such methods when only limited training data is available. In addition, in real-world situations, the shoulder surfer's face may be partially occluded and, as such, there is a need for methods to also detect partially visible faces [22]. Nevertheless, our results present a solid foundation to methodically assess shoulder surfing risks.

Third, any method to assess the shoulder surfing risk through the front-facing camera is only as good as what it can capture. Our approach cannot capture people outside the camera's view, but this is not a limitation that we can address. However, our demonstration of PrivacyScout relied on a wide-angle GoPro that had a field of view of 170°. This is in contrast to current mobile phone front-facing cameras that are limited and are not suitable for this task (Figure 11). Nevertheless, our analyses highlight that, while *current* front-facing cameras have a limited field of view which makes it difficult to detect potential shoulder surfers (e.g. due to occlusion), there is significant potential for *wide-angle front-facing cameras* to enhance the user's privacy and determine whether the presence of a user's face in the camera view is a shoulder surfing risk or not.

Last but not least, our study has imbalanced demographics. We argue that while gender does not play a significant role in observation attacks, the heights of both the user and attacker might. Modelling the interplay between the height of the user, the attacker, and the device's orientation (which influences where the face appears in the camera's view) is complex and left for future work. Nevertheless, despite not analysing all factors influencing shoulder surfing, our work demonstrated significant differences in other key characteristics such as content type, observation distance, or viewing angle.

8 Conclusion

We studied the influence of multiple viewing angles and distances on the success of a shoulder surfing attack on mobile devices. Evaluating the most likely types of content to be shoulder surfed, i.e. text messages, photos, or PINs, we showed that PIN authentications are most vulnerable while the success probability of an attack on text and photos depends on the observer's location relative to the mobile device. We then proposed PrivacyScout, a novel computational method that can regress a shoulder surfing risk score using visual features extracted from the (wide-angle) front-facing camera of a mobile device that captures the observer's face. Overall, our analyses and method represent the next steps towards real-time privacy- and security-aware UIs and systems that can better alert users of shoulder surfing.

Acknowledgement

This project is partly funded by the Swiss National Science Foundation (SNSF) Early Postdoc. Mobility Fellowship (199991), the Deutsche Forschungsgemeinschaft (DFG) German Research Foundation (425869382), part of Priority Program SPP2199 Scalable Interaction Paradigms for Pervasive Computing Environments, EPSRC (EP/V008870/1) and the PETRAS National Centre of Excellence for IoT Systems Cybersecurity, which is also funded by the EPSRC (EP/S035362/1), and the the European Research Council (ERC; grant agreement 801708).

References

[1] Vision activity – eye chart. https://www.teachengineering.org/content/cub_/activities/cub_human/cub_human_lesson06_activity1_eyechart.pdf, 2022. last accessed on March 15, 2022..

[2] Yomna Abdelrahman, Mohamed Khamis, Stefan Schneegass, and Florian Alt. *Stay Cool! Understanding Thermal Attacks on Mobile-Based User Authentication*, page 3751–3763. Association for Computing Machinery, New York, NY, USA, 2017.

[3] Yasmeen Abdrabou, Mohamed Khamis, Rana Mohamed Eisa, Sherif Ismail, and Amrl Elmougy. Just gaze and wave: exploring the use of gaze and gestures for shoulder-surfing resilient authentication. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications*, pages 1–10, 2019.

[4] Mohammed Eunus Ali, Anika Anwar, Ishrat Ahmed, Tanzima Hashem, Lars Kulik, and Egemen Tanin. 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, page 1–4, New York, NY, USA, 2014. Association for Computing Machinery.

[5] Florian Alt, Stefan Schneegass, Alireza Sahami Shirazi, Mariam Hassib, and Andreas Bulling. Graphical passwords in the wild: Understanding how users choose pictures and passwords in image-based authentication schemes. In *Proceedings of the 17th International Conference on Human-Computer Interaction with Mobile Devices and Services*, pages 316–322, 2015.

[6] Mykhaylo Andriluka, Leonid Pishchulin, Peter Gehler, and Bernt Schiele. 2d human pose estimation: New benchmark and state of the art analysis. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[7] Adam J. Aviv, John T. Davin, Flynn Wolf, and Ravi Kuber. Towards baselines for shoulder surfing on mobile authentication. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, ACSAC 2017, page 486–498, New York, NY, USA, 2017. Association for Computing Machinery.

[8] Adam J Aviv and Dane Fichter. Understanding visual perceptions of usability and security of android's graphical password pattern. In *Proceedings of the 30th Annual Computer Security Applications Conference*, pages 286–295, 2014.

[9] Adam J. Aviv, Flynn Wolf, and Ravi Kuber. Comparing video based shoulder surfing with live simulation. In *Proceedings of the 34th Annual Computer Security Applications Conference*, ACSAC '18, page 453–466, New York, NY, USA, 2018. Association for Computing Machinery.

[10] T. Baltrusaitis, P. Robinson, and Louis-Philippe Morency. Constrained local neural fields for robust facial landmark detection in the wild. *2013 IEEE International Conference on Computer Vision Workshops*, pages 354–361, 2013.

[11] T. Baltrusaitis, Amir Zadeh, Y. Lim, and Louis-Philippe Morency. Openface 2.0: Facial behavior analysis toolkit. *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 59–66, 2018.

[12] Chun-Yu Chen, Bo-Yao Lin, Junding Wang, and Kang G Shin. Keep others from peeking at your mobile device screen! In *The 25th Annual International Conference on Mobile Computing and Networking*, pages 1–16, 2019.

[13] Online Course. Mobile hci unit 2. https://mobilehci-unit2-orientation.glitch.me/, 2021. last accessed on March 15, 2022..

[14] Paul Dunphy, Andreas P Heiner, and N Asokan. A closer look at recognition-based graphical passwords on mobile devices. In *Proceedings of the Sixth Symposium on Usable Privacy and Security*, pages 1–12, 2010.

[15] Malin Eiband, Mohamed Khamis, Emanuel von Zezschwitz, Heinrich Hussmann, and Florian Alt. 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, page 4254–4265, New York, NY, USA, 2017. Association for Computing Machinery.

[16] Malin Eiband, Emanuel von Zezschwitz, Daniel Buschek, and Heinrich Hussmann. My scrawl hides it all: Protecting text messages against shoulder surfing with handwritten fonts. In *Proc. of CHI EA '16*, CHI EA '16, pages 2041–2048, New York, NY, USA, 2016. ACM.

[17] Habiba Farzand, Kinshuk Bhardwaj, Karola Marky, and Mohamed Khamis. The interplay between personal relationships & shoulder surfing mitigation. In *Proceedings of the Mensch und Computer 2021 (MuC '21)*, 2021.

[18] Alain Forget, Sonia Chiasson, and Robert Biddle. Shoulder-surfing resistance with eye-gaze entry in cued-recall graphical passwords. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1107–1110, 2010.

[19] Marian Harbach, Emanuel Von Zezschwitz, Andreas Fichtner, Alexander De Luca, and Matthew Smith. {It's} a hard lock life: A field study of smartphone ({Un) Locking} behavior and risk perception. In *10th symposium on usable privacy and security (SOUPS 2014)*, pages 213–230, 2014.

[20] Apple Inc. Apple developer documentation. https://apple.co/3d132A0, 2021. last accessed on March 15, 2022..

[21] GoPro Inc. Go pro. https://gopro.com/, 2021. last accessed on March 15, 2022..

[22] M. Khamis, Anita Baier, N. Henze, Florian Alt, and A. Bulling. Understanding face and eye visibility in front-facing cameras of smartphones used in the wild. *Proceedings*

of the 2018 CHI Conference on Human Factors in Computing Systems, 2018.

[23] Mohamed Khamis, Florian Alt, Mariam Hassib, Emanuel von Zezschwitz, Regina Hasholzner, and Andreas Bulling. Gazetouchpass: Multimodal authentication using gaze and touch on mobile devices. In *Proceedings of the 34th Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, New York, NY, USA, 2016. ACM.

[24] Mohamed Khamis, Malin Eiband, Martin Zürn, and Heinrich Hussmann. Eyespot: Leveraging gaze to protect private text content on mobile devices from shoulder surfing. *Multimodal Technologies and Interaction*, 2(3), 2018.

[25] Mohamed Khamis, Mariam Hassib, Emanuel von Zezschwitz, Andreas Bulling, and Florian Alt. Gazetouchpin: Protecting sensitive data on mobile devices using secure multimodal authentication. In *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, ICMI 2017, New York, NY, USA, 2017. ACM.

[26] Mohamed Khamis, Tobias Seitz, Leonhard Mertl, Alice Nguyen, Mario Schneller, and Zhe Li. Passquerade: Improving error correction of text passwords on mobile devices by using graphic filters for password masking. In *Proceedings of the 37th Annual ACM Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. ACM.

[27] Yiren Kong, Young Sik Seo, and Ling Zhai. Comparison of reading performance on screen and on paper: A meta-analysis. *Computers & Education*, 123:138–149, 2018.

[28] Katharina Krombholz, Thomas Hupperich, and Thorsten Holz. May the force be with you: The future of force-sensitive authentication. *IEEE Internet Computing*, 21(3):64–69, 2017.

[29] Manu Kumar, Tal Garfinkel, Dan Boneh, and Terry Winograd. Reducing shoulder-surfing by using gaze-based password entry. In *Proceedings of the 3rd symposium on Usable privacy and security*, pages 13–19, 2007.

[30] Diogo Marques, Tiago Guerreiro, Luis Carriço, Ivan Beschastnikh, and Konstantin Beznosov. *Vulnerability & Blame: Making Sense of Unauthorized Access to Smartphones*, page 1–13. Association for Computing Machinery, New York, NY, USA, 2019.

[31] Ildar Muslukhov, Yazan Boshmaf, Cynthia Kuo, Jonathan Lester, and Konstantin Beznosov. Know your enemy: the risk of unauthorized access in smartphones by insiders. In *Proceedings of the 15th international conference on Human-computer interaction with mobile devices and services*, pages 271–280, 2013.

[32] Bo Pang and Lillian Lee. Movie review data. http://www.cs.cornell.edu/people/pabo/movie-review-data/, 2021. last accessed on March 15, 2022..

[33] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of EMNLP*, pages 79–86, 2002.

[34] George Probst. Analysis of the effects of privacy filter use on horizontal deviations in posture of vdt operators. Master's thesis, Virginia Polytechnic Institute and State University, 2000.

[35] Kirill Ragozin, Yun Suen Pai, Olivier Augereau, Koichi Kise, Jochen Kerdels, and Kai Kunze. Private reader: Using eye tracking to improve reading privacy in public spaces. In *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19, New York, NY, USA, 2019. Association for Computing Machinery.

[36] Vijay Rajanna, Seth Polsley, Paul Taele, and Tracy Hammond. A gaze gesture-based user authentication system to counter shoulder-surfing attacks. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 1978–1986, 2017.

[37] Alia Saad, Michael Chukwu, and Stefan Schneegass. Communicating shoulder surfing attacks to users. In *Proceedings of the 17th International Conference on Mobile and Ubiquitous Multimedia*, MUM 2018, page 147–152, New York, NY, USA, 2018. Association for Computing Machinery.

[38] Alia Saad, Jonathan Liebers, Uwe Gruenefeld, Florian Alt, and Stefan Schneegass. Understanding Bystanders' Tendency to Shoulder Surf Smartphones Using 360-degree Videos in Virtual Reality. In *Proceedings of the 23rd International Conference on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '21, New York, NY, USA, 2021. ACM.

[39] Samsung. Samsung galaxy a80. https://www.samsung.com/global/galaxy/galaxy-a80/specs/, 2021. last accessed on March 15, 2022..

[40] Angela Sasse. Scaring and bullying people into security won't work. *IEEE Security Privacy*, 13(3):80–83, 2015.

[41] Florian Schaub, Ruben Deyhle, and Michael Weber. 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, New York, NY, USA, 2012. Association for Computing Machinery.

[42] Stefan Schneegass, Frank Steimle, Andreas Bulling, Florian Alt, and Albrecht Schmidt. Smudgesafe: Geometric image transformations for smudge-resistant user authentication. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 775–786, 2014.

[43] Minho Shin, Cory Cornelius, D. Peebles, Apu Kapadia, D. Kotz, and Nikos Triandopoulos. Anonysense: A system for anonymous opportunistic sensing. *Pervasive Mob. Comput.*, 7:16–30, 2011.

[44] Emanuel von Zezschwitz, Alexander De Luca, Bruno Brunkow, and Heinrich Hussmann. *SwiPIN: Fast and Secure PIN-Entry on Smartphones*, page 1403–1406. Association for Computing Machinery, New York, NY, USA, 2015.

[45] Emanuel von Zezschwitz, Sigrid Ebbinghaus, Heinrich Hussmann, and Alexander De Luca. 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, pages 4320–4324, New York, NY, USA, 2016. ACM.

[46] Emanuel von Zezschwitz, Sigrid Ebbinghaus, Heinrich Hussmann, and Alexander De Luca. *You Can't Watch This! Privacy-Respectful Photo Browsing on Smartphones*, page 4320–4324. Association for Computing Machinery, New York, NY, USA, 2016.

[47] Oliver Wiese and Volker Roth. Pitfalls of shoulder surfing studies. In *NDSS workshop on usable security*, pages 1–6, 2015.
[48] E. Wood, T. Baltrusaitis, Xucong Zhang, Yusuke Sugano, P. Robinson, and A. Bulling. Rendering of eyes for eye-shape registration and gaze estimation. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 3756–3764, 2015.
[49] Amir Zadeh, T. Baltrusaitis, and Louis-Philippe Morency. Convolutional experts constrained local model for facial landmark detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2051–2059, 2017.
[50] Nur Haryani Zakaria, David Griffiths, Sacha Brostoff, and Jeff Yan. Shoulder surfing defence for recall-based graphical passwords. In *Proceedings of the seventh symposium on usable privacy and security*, pages 1–12, 2011.
[51] Xucong Zhang, Yusuke Sugano, and A. Bulling. Revisiting data normalization for appearance-based gaze estimation. *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications*, 2018.
[52] Huiyuan Zhou, Khalid Tearo, Aniruddha Waje, Elham Alghamdi, Thamara Alves, Vinicius Ferreira, Kirstie Hawkey, and Derek Reilly. Enhancing mobile content privacy with proxemics aware notifications and protection. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, page 1362–1373, New York, NY, USA, 2016. Association for Computing Machinery.

A Appendix

In this section, we present the complete results for our statistical analysis from section 5.

A.1 Statistical Results PINs

We conducted a two-way repeated measures ANOVA to analyse the effects of distance and angle on the *average edit distance*. Our analysis did not reveal any statistically significant results, neither of the main effects of distance and angle nor interaction effects.

Distance:

- average edit distance: $F_{3,45} = 0.65$, $p = 0.59$, $\eta_p^2 = 0.04$

Angle:

- average edit distance: $F_{2,30} = 1.34$, $p = 0.28$, $\eta_p^2 = 0.08$

Distance*Angle:

- average edit distance: $F_{6,90} = 0.30$, $p = 0.93$, $\eta_p^2 = 0.02$

A.2 Statistical Results Photos

Using a two-way repeated measure ANOVA, we analysed the influence of distance, angle, and distance*angle on each of the five different evaluation metrics: *success rate indoors vs. outdoors*, *success rate count people*, *success rate facial expressions*, *success rate activity*, and *partial success rate activity*. Statistically significant results are highlighted in **bold**.

Distance:

- success rate indoors vs. outdoors: $F_{3,45} = 0.65$, $p = 0.59$, $\eta_p^2 = 0.04$
- success rate count people: $F_{3,45} = 1.31$, $p = 0.28$, $\eta_p^2 = 0.08$
- **success rate facial expressions**: $F_{3,45} = 38.10$, $p < 0.01$, $\eta_p^2 = 0.72$
- **success rate activity**: $F_{3,45} = 4.10$, $p = 0.011$, $\eta_p^2 = 0.21$
- partial success rate activity: $F_{3,45} = 0.73$, $p = 0.54$, $\eta_p^2 = 0.05$)

Angle:

- success rate indoors vs. outdoors: $F_{2,30} = 0.48$, $p = 0.62$, $\eta_p^2 = 0.03$
- **success rate count people**: $F_{2,30} = 5.74$, $p < 0.01$, $\eta_p^2 = 0.28$)
- **success rate facial expressions**: $F_{2,30} = 33.42$, $p < 0.01$, $\eta_p^2 = 0.69$
- success rate activity: $F_{2,30} = 0.90$, $p = 0.42$, $\eta_p^2 = 0.06$
- partial success rate activity: $F_{2,30} = 0.35$, $p = 0.71$, $\eta_p^2 = 0.02$

Distance*Angle:

- success rate indoors vs. outdoors: $F_{6,90} = 1.18$, $p = 0.32$, $\eta_p^2 = 0.07$
- success rate count people: $F_{6,90} = 1.0$, $p = 0.43$, $\eta_p^2 = 0.06$
- success rate facial expressions: $F_{6,90} = 0.98$, $p = 0.45$, $\eta_p^2 = 0.06$
- success rate activity: $F_{6,90} = 0.48$, $p = 0.82$, $\eta_p^2 = 0.03$
- partial success rate activity: $F_{6,90} = 0.18$, $p = 0.98$, $\eta_p^2 = 0.01$

A.3 Statistical Results Text

Using a two-way repeated measure ANOVA, we analysed the influence of the main effects, distance and angle, and possible interaction effects, distance*angle, on each of the four evaluation metrics: *success rate recognise words*, *success rate understand context*, *success rate retrieve names*, and *partial success rate retrieve names.* Statistically significant results are highlighted in **bold**.

Distance:

- **success rate recognise words**: $F_{3,45} = 40.87$, $p < 0.01$, $\eta_p^2 = 0.73$
- **success rate understand context**: $F_{3,45} = 72.03$, $p < 0.01$, $\eta_p^2 = 0.83$
- **success rate retrieve names**: $F_{3,45} = 36.55$, $p < 0.01$, $\eta_p^2 = 0.71$
- partial success rate retrieve names: $F_{3,45} = 0.74$, $p = 0.54$, $\eta_p^2 = 0.05$

Angle:

- **success rate recognise words**: $F_{2,30} = 50.45$, $p < 0.01$, $\eta_p^2 = 0.77$
- **success rate understand context**: $F_{2,30} = 79.55$, $p < 0.01$, $\eta_p^2 = 0.84$
- **success rate retrieve names**: $F_{2,30} = 30.27$, $p < 0.01$, $\eta_p^2 = 0.67$
- partial success rate retrieve names: $F_{2,30} = 0.18$, $p = 0.84$, $\eta_p^2 = 0.01$

Distance*Angle:

- **success rate recognise words**: $F_{6,90} = 4.97$, $p < 0.01$, $\eta_p^2 = 0.25$
- **success rate understand context**: $F_{6,90} = 3.94$, $p < 0.01$, $\eta_p^2 = 0.21$
- **success rate retrieve names**: $F_{6,90} = 2.69$, $p = 0.02 < 0.05$, $\eta_p^2 = 0.15$
- partial success rate retrieve names: $F_{6,90} = 1.03$, $p = 0.41$, $\eta_p^2 = 0.06$

B Additional Method Evaluation Results

Table 3 shows the results of the evaluation of PrivacyScout when trained for the text condition using only images from the 30° and 60° viewing angle. A one-way repeated measures ANOVA showed that the difference between conditions was significant ($F(10, 150) = 4.09$, $p < 0.01$, $\eta_p^2 = 0.21$). Post-hoc tests did not reveal any statistically significant differences at the $p < 0.05$ level. This might be due to the limited amount of data used in the evaluation.

Table 3. Evaluation results from leave-one-subject-out cross validation for regression of the shoulder surfing risk score for **text**. Models trained using images from the 30° and 60° viewing angle (no images from 0°). The risk score is a value between 0 and 3. The best performing method is underlined.

Method	Mean MAE	Std. dev. of the MAE
Face position	**0.75**	**0.28**
Face position and size	**0.72**	**0.27**
Head pose	**0.79**	**0.34**
Gaze angles	**0.74**	**0.30**
All	**0.69**	**0.30**
Baseline: Midpoint	**0.74**	**0.28**
Baseline: Distance	**0.72**	**0.32**
Baseline: Both	**0.71**	**0.28**
Naïve mean	**1.00**	**0.28**
Naïve median	**0.74**	**0.58**
Naïve constant	**0.74**	**0.58**

Table 4 shows the results of the evaluation of PrivacyScout when trained for the photo condition using only images from the 30° and 60° viewing angle. A one-way repeated measures ANOVA showed that the difference between conditions was significant ($F(10, 150) = 4.94$, $p < 0.01$, $\eta_p^2 = 0.25$). Post-hoc tests revealed pairwise differences between *Face position and size* and *Head pose* vs. *Naïve median* ($p < 0.05$). The differences between all the other pairs were not statistically significant at the $p < 0.05$ level.

Table 4. Evaluation results from leave-one-subject-out cross validation for regression of the shoulder surfing risk score for **photos**. Models trained using images from the 30° and 60° conditions (no images from 0°). The risk score is a value between 0 and 4. The best performing method is underlined.

Method	Mean MAE	Std. dev. of the MAE
Face position	**0.48**	**0.11**
Face position and size	**0.47**	**0.11**
Head pose	**0.47**	**0.12**
Gaze angles	**0.56**	**0.14**
All	**0.53**	**0.13**
Baseline: Midpoint	**0.48**	**0.11**
Baseline: Distance	**0.50**	**0.13**
Baseline: Both	**0.48**	**0.11**
Naïve mean	**0.62**	**0.09**
Naïve median	**0.64**	**0.16**
Naïve constant	**0.56**	**0.23**