VR
IEEE 2022
IEEE CONFERENCE ON VIRTUAL REALITY + 3D USER INTERFACES | 12-16 MARCH 2022 | VIRTUAL EVENT

Virtual Reality Observations: Using Virtual Reality to Augment Lab-Based Shoulder Surfing Research

In Proceedings of the IEEE Conference on Virtual Reality and 3D User Interfaces (VR)

Christchurch, New Zealand, March 2022 (IEEE VR 2022)

Florian Mathis, University of Glasgow, United Kingdom
florian.mathis@glasgow.ac.uk
Joseph O'Hagan, University of Glasgow, United Kingdom
j.ohagan.1@research.gla.ac.uk
Mohamed Khamis, University of Glasgow, United Kingdom
mohamed.khamis@glasgow.ac.uk
Kami Vaniea, University of Edinburgh, United Kingdom
kvaniea@inf.ed.ac.uk

Virtual Reality Observations: Using Virtual Reality to Augment Lab-Based Shoulder Surfing Research

Florian Mathis*
University of Glasgow
University of Edinburgh

Joseph O'Hagan†
University of Glasgow

Mohamed Khamis‡
University of Glasgow

Kami Vaniea§
University of Edinburgh

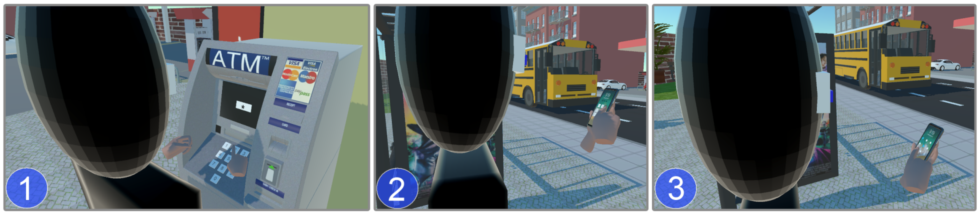

Figure 1: We explore the use of virtual reality (VR) for shoulder surfing research in the authentication research domain. We compare the impact of non-immersive/immersive VR observations on participants' observation performance and behaviour while shoulder surfing authentications. We demonstrate the strengths of VR-based shoulder surfing research by exploring three different authentication scenarios: (❶) automated teller machine (ATM), (❷) smartphone PIN, and (❸) smartphone pattern authentication.

Abstract

Given the difficulties of studying the shoulder surfing resistance of authentication systems in a live setting, researchers often ask study participants to shoulder surf authentications by watching two-dimensional (2D) video recordings of a user authenticating. However, these video recordings do not provide participants with a realistic shoulder surfing experience, creating uncertainty in the value and validity of lab-based shoulder surfing experiments. In this work, we exploit the unique characteristics of virtual reality (VR) and study the use of non-immersive/immersive VR recordings for shoulder surfing research. We conducted a user study (N=18) to explore the strengths and weaknesses of such a VR-based shoulder surfing research approach. Our results suggest that immersive VR observations result in a more realistic shoulder surfing experience, in a significantly higher sense of being part of the authentication environment, in a greater feeling of spatial presence, and in a higher level of involvement than 2D video observations without impacting participants' observation performance. This suggests that studying shoulder surfing in VR is advantageous in many ways compared to currently used approaches, e.g., participants can freely choose their observation angle rather than being limited to a fixed observation angle as done in current methods. We discuss the strengths and weaknesses of using VR for shoulder surfing research and conclude with four recommendations to help researchers decide when (and when not) to employ VR for shoulder surfing research in the authentication research domain.

Keywords: Virtual Reality, Shoulder Surfing, Authentication

1 Introduction

Accessing private data has become a fundamental part of most people's daily life. Examples include, but are not limited to, checking emails on smartphones, accessing the account balance through online banking apps, or withdrawing cash at automated teller machines (ATMs). In many of these situations, users are required to authenticate (e.g., to enter a PIN), which puts them at risk of getting observed (referred to as *shoulder surfing* [23]). Consequently, researchers looked into the shoulder surfing resistance of a large variety of authentication schemes (e.g., [11, 18, 20, 39, 63]). A common approach in human-centred security research is to study such systems' security by inviting participants to the lab, showing them two-dimensional (2D) video recordings, and asking them to guess the observed PIN (e.g., [20, 41]). These recordings show user authentications from pre-defined observation angles, with researchers' intention to simulate a "best-case scenario" for an attacker that shoulder surfs the user. Although 2D video recordings form a suitable baseline for shoulder surfing research [9], it remains unclear a) how (if at all) researchers empirically define the observation perspective, b) if the selected perspective indeed represents a best-case scenario for attackers, and c) if 2D video recordings can provide realistic shoulder surfing experiences. While studying shoulder surfing in a live setting is possible, it is often challenging [82] and in some cases even infeasible. For example, studying shoulder surfing on ATM authentications in the real world is close to impossible due to ethical and legal constraints [19, 77].

As a result, we explore in this work how virtual reality (VR) can support shoulder surfing research by enabling researchers to study shoulder surfing in settings that are challenging to replicate in the lab and infeasible to research in the real world. Ideally, researchers would be able to assess a system's shoulder surfing resistance in a variety of contexts without much effort. Through the use of VR as a research platform, we enable researchers to a) evaluate the shoulder surfing resistance of authentications in situ instead of in lab settings (e.g., [18, 22]), and b) investigate participants' observation strategies in much more detail than what can be achieved in traditional lab settings. To explore the potential of VR for shoulder surfing research on authentication systems, we conducted a lab-based VR user study (N=18). We exposed participants to user authentications in three different contexts: ATM, smartphone PIN, and smartphone pattern authentication. We then ran a comparison of participants' perception when shoulder surfing user authentications using commonly used 2D video recordings (i.e., *2DVO*, our baseline), and non-immersive[1] (i.e., *3DO*) and immersive VR recordings (i.e., *VRO*). Our findings

*e-mail: florian.mathis@glasgow.ac.uk
†e-mail: j.ohagan.1@research.gla.ac.uk
‡e-mail: mohamed.khamis@glasgow.ac.uk
§e-mail: kvaniea@inf.ed.ac.uk

[1] We use the terminology by Freina and Ott [28] where *non-immersive* refers to a computer-based environment that simulates places in the real or imagined worlds. *Immersive* takes the idea even further by providing the perception of being physically present in the non-physical world.

show that there is a significant difference in participants' observation performance between *VRO* and *3DO*. However, inline with Aviv et al.'s findings [9], *2DVO* already provide a suitable baseline measure for shoulder surfing research, especially when assessing a system's resilience against observations. Participants' observation performance is highest in *VRO* (M=93.14%, SD=25.35%), followed by *2DVO* (M=89.35%, SD=30.92%) and *3DO* (M=81.40%, SD=39.01%), with no evidence of a significant difference between *VRO* and *2DVO*. *VRO* resulted in a higher level of sense of being there, in a higher level of spatial presence, and increased participants' involvement and experienced realism compared to *2DVO* and *3DO*. This, together with participants' observation performance, suggests that *VRO* are suitable for shoulder surfing research and are to be preferred in situations where researchers' aim is to a) provide participants with more realistic shoulder surfing experiences and b) study participants' observation strategies in much more detail than what *2DVO* are capable of.

Based on our findings, we contribute four lessons learned, such as accounting for real-world factors (e.g., proxemics [32]) and the importance of introducing participants to novel observation methods, to support researchers in their decision when (and when not) to employ VR as a research method for authentication and shoulder surfing research. In sum, the contribution of our work is three-fold: **(1)** We propose the use of non-immersive and immersive VR observations for shoulder surfing research on authentication systems and explore their strengths and weaknesses. **(2)** We demonstrate through three different authentication scenarios how VR can contribute towards more realistic shoulder surfing research. **(3)** Finally, we discuss our findings in the light of prior works and provide four recommendations to support researchers when leveraging VR for authentication and shoulder surfing investigations.

2 Background and Related Work

To contextualise our work, we review shoulder surfing and authentication research, and works that used VR as a research platform.

2.1 Shoulder Surfing and Authentication Research

The literature of shoulder surfing ranges from works that collected shoulder surfing stories in the wild [23], to more system-focused research that explored the shoulder surfing resistance of novel privacy-protecting (e.g., [15,57]) and security systems (e.g., [20,21,78]). In authentication research, which is considered to be a major theme in human-centred security and privacy [29], most shoulder surfing evaluations rely on either a) two-dimensional video recordings or b) live observations [13,82]. Roth et al. [61] exposed participants to video recordings that showed both the authentication scheme and the user's interactions. De Luca et al. [20] located a camera opposite to their participants and an additional one at participants' back to run post-hoc shoulder surfing evaluations. There is a significant larger body of work that relied on video recordings for shoulder surfing evaluations (e.g., [11,18,39,63]).

Others conducted shoulder surfing research through live observations where participants observed authentications in real time and could choose a viewing position on their own. Zakaria et al. [86] simulated live shoulder surfing by letting participants observe user authentications performed by the experimenter (who acted as "the victim"). Mathis et al. [48] equipped participants with a smartphone to then let them freely move around and record the user's authentication. Saad et al. [62] used 360° real-world videos to better understand users' shoulder surfing gaze behaviour and argued such a virtual scenario "brings us one step closer to the goal of understanding shoulder surfing". Other works investigated the impact of multiple simultaneous shoulder surfers on a system's resistance [40] or proposed a model for modern shoulder surfing where authentications are divided into minimal human observations [83]. Bošnjak and Brumen [13] even argued that shoulder surfing evaluation methods are often "devised based on expert knowledge and general intuition, [but] method design should instead be driven by well-established experimental evaluation" [13]. In many shoulder surfing studies, it has been argued that the systems are evaluated "under optimal conditions for the attacker" [61], that "opting for an expert attack represents a worst-case-scenario that provides a good estimate of the security of an authentication mechanism" [20], and that corresponding threat models assumed a "best case scenario for the attacker" [48]. However, Wiese and Roth [82] recommended to study live observations and that attackers' observation strategies should be taken into account when studying security systems empirically [82]. Aviv et al. [9] went one step further and analysed the claims that video recordings offer a suitable alternative for shoulder surfing research. Although they concluded that 2D video recordings can provide a suitable baseline measure for shoulder surfing [9], they also highlighted the importance of not overclaiming findings of such evaluations as they can, in fact, greatly underestimate the threat of an attacker in a live setting [9].

2.2 VR Studies for Human-centred Research

Several research communities recently began using VR as a research platform for human-centred research. The Human-computer Interaction (HCI) community investigated using VR as a research methodology to evaluate smart artefacts [76] and pedestrian navigation methods [65]. Voit et al.'s comparison of five empirical research methods [76] (i.e., online, VR, augmented reality, lab, in situ) suggested that VR and in situ provide similar insights when evaluating standardised questionnaires such as SUS [14] or AttrakDiff [35]. Weiß et al. [81] showed that alternative empirical research methods (e.g., VR) might be used to infer insights about in situ studies and that the evaluation of situated visualisations is not necessarily dependent on the empirical research method.

In the human-centred security domain, Mathis et al. [47] conducted a replication study to evaluate a real-world authentication scheme in VR. While their work, along with George et al.'s initial comments on using VR as a testbed [30], is the first that validated the potential of VR for human-centred security research, Mathis et al. [47] also argued that their investigation lays only the groundwork. Particularly, that follow-up research is required to validate the use of VR for the broader research field and establish *VR studies* as a complementary research method for real-world investigations. For example, Mathis et al. [47] did not study VR's unique affordances of non-immersive and immersive VR observations for shoulder surfing research. *VR studies* can also be particularly helpful at times where physical spaces are challenging to access or even prohibited (e.g., during a pandemic) [44]. Rebelo et al. [58] argued that VR enables researchers to develop realistic-looking environments that come with greater control of experimental conditions than lab settings and that users' experience can benefit from using VR as a research methodology. Thomas Parsons [53] showed that virtual environments can enhance ecological validity in the clinical, affective, and social neurosciences through evaluation paradigms that combine the experimental control of laboratory measures with emotionally engaging background narratives.

VR studies were also proposed as a new social psychological research tool to overcome the existing problems around control–mundane realism trade-off, lack of replication, and unrepresentative sampling [12]. Fiore et al. [27] proposed *VR studies* in the environmental policy research domain to provide a bridge over the methodological gap between lab and field studies and concluded that VR has the potential to combine the internal validity of controlled lab experiments with the external validity of field experiments.

2.3 Lessons Learned from Prior Work

From the literature, we learned that live shoulder surfing (instead of video recordings) should be preferred when conducting shoulder surfing research (e.g., [8, 82]). However, human-centred security researchers often rely on video recordings due to the difficulties of running these evaluations in real time (e.g., requiring researchers to simulate real-world adversaries [82]). It is worth mentioning that video recordings offer consistency across the entire study sample, which is not necessarily the case in a real-time setting [82]. Prior work showed that VR setups enable researchers to simulate hard-to-reach or safety-critical physical locations in an affordable and effortless way [44, 54]. This is particularly interesting for the human-centred security domain where private and sensitive contexts are often challenging to study [19, 77]. We also noticed that VR has already been successfully applied in several other research domains (e.g., Human-computer Interaction [44, 47, 52, 76], Information Visualisation [81, 85]).

To draw on the success of previous *VR studies* and to close the gap between commonly used 2D video recordings and the often hard to conduct real-time shoulder surfing evaluations, we build upon previous works that used 2D video recordings (e.g., [8, 9, 47]). As such, we investigate the strengths of VR for in situ shoulder surfing research and participants‘ performance when using three-dimensional VR-based observations. While Mathis et al. [47] ran a comparison between 2D videos recorded in VR (*2DVO*, our baseline) and 2D real-world videos, we extend their work by investigating for the first time the impact of 3D non-immersive and immersive VR observations on participants‘ shoulder surfing performance and behaviour. Saad et al. [62] proposed 360° real-world videos for shoulder surfing research, but there is a lack of an evaluation of a) the impact of such recordings on participants‘ performance when observing different authentication schemes in different contexts and b) users‘ observation strategies and their movement behaviour (e.g., positioning, adhering to social proxemics [32]). Furthermore, due to the lack of a baseline condition in the work by Saad et al. [62] (e.g., 2D videos [11, 18, 63]), it remains unclear how participants‘ performance differs in comparison to the use of traditional 2D videos. We fill this gap through an in-depth comparison between three-dimensional VR observations (*3DO* and *VRO*, see Sect. 3.2) and the de facto standard approach (2D Video Observations) to evaluate authentication systems and their resilience against shoulder surfing.

Our work provides promising insights into the use of VR for authentication and shoulder surfing research. It demonstrates how such a research approach enables researchers to study users‘ movement behaviour when observing user authentications in different environments and on different authentication schemes and opens the door for the research community to leverage VR's unique affordances to further advance human-centred security research.

3 Studied Authentication Scenarios: Apparatus and Implementation

We simulated in this work three scenarios that all take place in public spaces: 1) ATM authentication, 2) smartphone PIN authentication, 3) smartphone pattern authentication. We studied these three scenarios due to several reasons: First, a survey by Eiband et al. [23] showed that shoulder surfing is most prominent in public spaces, especially when using smartphones. Second, ATMs are often found in public spaces, are frequently visited by people (e.g., De Luca et al. [19] reported widespread ATM usage), and are particularly challenging to research in the real world [19, 77]. Running a similar study in front of a real-world ATM is close to impossible in the detail required for our research. Furthermore, shoulder surfing forms an important threat vector in authentication research and both studied schemes (i.e., PIN and pattern) form a popular security baseline in the human-centred security field (e.g., for PINs: [8, 20, 31, 39], for patterns: [8, 20, 31]).

To evaluate the suitability of VR-based three-dimensional observations for shoulder surfing research, we first had to collect recordings of users authenticating. We implemented three authentication scenarios using Unity 3D (C#), see Fig. 1. We used a leap motion for the hand tracking [50] and an abstract avatar design that comes with a head, body, legs, eyes, and hands. Note that the abstract avatar‘s dimensions and movements were mapped to a human in the real world. Previous research showed that shoulder surfing studies conducted in virtual environments do not necessarily require highly realistic full-body avatars [45, 47]. Using a more abstract avatar also contributes to making VR studies [44, 49, 51] more accessible to the broader research community [46] as it does not require additional expertise in hardware (e.g., tracking systems) and avatar-building expertise. We used the same avatar (see Fig. 1 and Fig. 3) for all three authentication systems, authentication environments, and observation methods to contribute to high internal validity. To track users' smartphone in the virtual environment, we attached an HTC VIVE tracker to the back of a real smartphone, similar to Amano et al. [7, Figure 5]. We then prepared 2D video recordings and non-immersive/immersive VR recordings for the actual user study (see Sect. 3.2). We enriched participants' shoulder surfing experience with realistic environmental sounds that match the virtual environment (e.g., traffic sounds, birds twittering).

3.1 Authentication Scenarios and Environments

We used a low-polygon styled city package [38], a 3D model of an ATM [1], and a smartphone 3D model [2] that we slightly modified by replacing the lock screen with our authentication schemes (PIN and pattern). For the PIN-based authentication, we used Unity's OnCollisionEnter method which triggers after another object collides (i.e., the user's finger). To implement a realistic pattern-based authentication scheme, we used Unity's Line Renderer component [3] which takes an array of two or more points in 3D space to then draw a straight line between each one. In the smartphone authentication scenarios the UI of the authentication scheme (i.e., the PIN/pattern layout) was only visible for the duration of the authentication. The authentication scheme disappeared as soon as a 4-symbol PIN/pattern was entered. This simulates a real-world smartphone authentication where the user lands on the home screen after authenticating (e.g., when unlocking the device).

3.2 Authentication Recordings

Two-dimensional video recordings (our baseline) are typically recorded from pre-defined observation angles with the aim to provide attackers with a best-case scenario, i.e., a clear sight on a mobile device's screen and input (e.g., [11, 39, 63]). We used VR capture [60] to create such 2D video recordings of both the user's input and the authentication scheme. Fig. 1 shows the three authentication systems participants observed. We used an observation position that presents participants with a “best case scenario”. The observation perspective for the 2D video recordings has been determined through pilot tests. For the three-dimensional recordings, we built upon Ultimate Replay [74], a state-based replay system that records the scene using “snapshots“ at regular intervals that reconstruct the scene during playback. We implemented additional scripts to track mesh changes and to keep track of the different states of Unity's Line Renderer component. Participants then experienced the authentications ($\sim$ 2 - 3.5 seconds, similar to previous PIN/pattern-based research [6, 18]) using state-of-the-art *2D Video Observations*, *3D Observations*, and *VR Observations*.

2D Video Observations (*2DVO*, baseline). Our baseline depicts the scenario where both the user's input and the authentication scheme were recorded using an angle that provides a shoulder surfer with a “best-case” scenario, similar to how prior shoulder surfing evaluations were conducted (e.g., [11, 39, 63, 78]). Participants performed their shoulder surfing observations on video recordings on a computer screen and could not manipulate the observation position and orientation. Note that we recorded the authentications through virtual cameras in the virtual environment. Previous work showed that shoulder surfers' observation performance on VR-based two-dimensional video material matches to a great extent with findings from a video-based real-world shoulder surfing study [47].

3D Observations (*3DO*, non-immersive). Participants' initial observation view was positioned so that the camera points towards the user's back. We did this to ensure that our participants come up with individual observation strategies and are required to change their position and perspective. The initial position did not provide them with a clear line of sight on the authentication scheme. Participants navigated in the environment using a traditional mouse-keyboard configuration, which we borrowed from previous work on direct manipulations in non-immersive VR environments (e.g., [25,59]). Participants used the keyboard to simulate walking (i.e., translation along the x/y/z-axis) and the mouse to simulate head movements (i.e., rotations along the x/y/z-axis), and watched the authentications on a traditional computer monitor after setting up their preferred observation position/orientation. Participants were not restricted to physical real-world conditions. We aimed to investigate if participants exploit the unique affordances of such a 3D observational approach in a virtual environment (e.g., being independent of gravitational force).

VR Observations (*VRO*, immersive). Participants were wearing a VR headset (i.e., HTC VIVE) and could freely move around and change their observation perspective and position as they wished. This depicts a scenario which is closest to in situ observations where a bystander can freely move around in a physical space and shoulder surf a user authenticating.

4 METHODOLOGY

We conducted a series of 1.5 hour in-the-lab investigations where participants (in the role of observers) observed overall 648 authentications (18 participants × 12 PINs/patterns × 3 authentication scenarios). We reached out to potential participants using social media postings and word of mouth (outside of a university environment). We recruited a sample of 18 participants (5 male, 13 female). Participants were on average 32.44 years (min=18, max=61, SD=12.22). All participants reported that they have used an ATM before and that they own a smartphone that they use on a daily basis. Slightly more than half of our participants (N=11) mentioned that they have used VR before. Participants observed authentications in all three authentication scenarios: 1) 4-digit PIN entries on an ATM, 2) 4-digit PIN entries on a smartphone, and 3) 4-symbol pattern entries on a smartphone. All participants went through all three observation methods (within-subject design). Conditions were counter-balanced using a Latin Square. As independent variables, we had the **observation type** (three levels: *2DVO* (our baseline), *3DO*, *VRO*), and the **threat model** (two levels: single-view and repeated-view observations, both threat models are frequently used when evaluating a system's security [39,41,47]). While in single-view observations participants could observe the user authenticating only once, in repeated-view observations participants could replay the authentication. The type of attack was alternating, similar to [41]. We had four dependent variables: **Observation Performance:** Participants' observation performance, the number of successful PIN/pattern guesses. **Levenshtein Distance:** the minimum number of single-digit edits between participants' best guess and the correct PIN/pattern, which is commonly used in shoulder surfing research (e.g., [4,21,31]). **Sense of Presence:** Participants' sense of presence experienced when using the different observation methods, measured using the standard IPQ questionnaire [66]. **Perceived workload:** Participants' perceived workload when using the different observation methods, measured using the NASA-TLX questionnaire [33].

Demographic questions (including age, gender, VR experience) were asked using Qualtrics [56]. We used additional in-VR questionnaires [26] to measure participants' perceived workload (NASA-TLX [34]) and presence (IPQ [66]). We did this to ensure a consistent VR experience and not break participants' focus [55].

4.1 Study Procedure

We first explained a) the different authentication scenarios and authentication schemes, b) the different observation methods, and c) what participants' task is (i.e., observing 4-digit PIN authentications). In advance of the observation task, participants went through an example authentication (e.g., "1234" PIN entry). We did this to familiarise them with the observation methods and the authentication schemes. Participants then started with the first observation method (e.g., *2DVO*) and observed four authentications for each authentication context. Participants were not allowed to clip through the virtual avatar in *3DO* and *VRO* as this would not be possible in the real world. However, we did not restrict them from positioning themselves in, for example, front of the virtual avatar because a) this could happen in the real world as well (e.g., standing at a bus station) and b) we aimed to investigate if participants make use of proxemics [32] (e.g., do they maintain a certain social distance to the user authenticating? are they aware that such observations are likely noticeable by the user authenticating?). For each observation, participants could provide up to three PIN/pattern guesses. Participants then filled in the NASA-TLX [34] and the IPQ questionnaire [66]. We concluded with semi-structured interviews (available in Appendix A in our supplementary material) about participants' perceived performance and their observation experience when using the different observation methods.

4.2 Ethical Considerations and Compensations

Our research has been reviewed and approved by the College of Science and Engineering Ethics Committee at the University of Glasgow. The study was conducted in Austria due to COVID-19. Participants were paid €15 (€10/h) and took part in a lottery to win additional €15. Participants were made aware in advance of the study that chances of winning increases with the number of successfully observed PINs/patterns. We did this to motivate them to perform well in their shoulder surfing task (similar to [41,48]).

5 RESULTS

We first report participants' observation performance, represented through the *percentages of successful observations* and the *mean Levenshtein distances*. We then report participants' sense of presence and perceived workload when using *2DVO*, *3DO*, and *VRO*. Finally, we provide a qualitative analysis of the semi-structured interviews along participants' observation strategies. Unless otherwise stated, we performed an aligned rank transformation on our data to correct for violations of normalcy using ART by Wobbrock et al. [84] and ART-C [24] for post-hoc pairwise comparisons. We report η_p^2 (*partial eta square*) as an effect size statistic for our ART analysis (0.01 = small, 0.06 = medium, 0.14 = large [16,17]). Appendix C & D in our supplementary material provide a full overview of the F-ratios, together with effect sizes, means, and stdevs.

5.1 Observation Performance and Levenshtein Distance

Participants' observations in *VRO* resulted in overall more successful observations (M=93.14%, SD=25.34%) than in *2DVO* (M=89.35%, SD=30.92%) and *3DO* (M=81.40%, SD=39.01%). We calculated the mean Levenshtein distances between participants' best guess and the correct PIN/pattern to proceed with a statistical analysis and to gain better insights into how close participants' guesses are to the entered PINs/patterns.

ATM Authentication: Participants' observation performance was M=94.44% (SD=15.94%) for *2DVO*, M=83.33% (SD=23.90%) for *3DO*, and M=95.59% (SD=14.40%) for *VRO*. There was a significant effect of observation method (F(1,83) = 4.584, $p < 0.05$, η_p^2 = 0.10) and threat model (F(1,83) = 4.526, $p < 0.05$, η_p^2 = 0.05) on

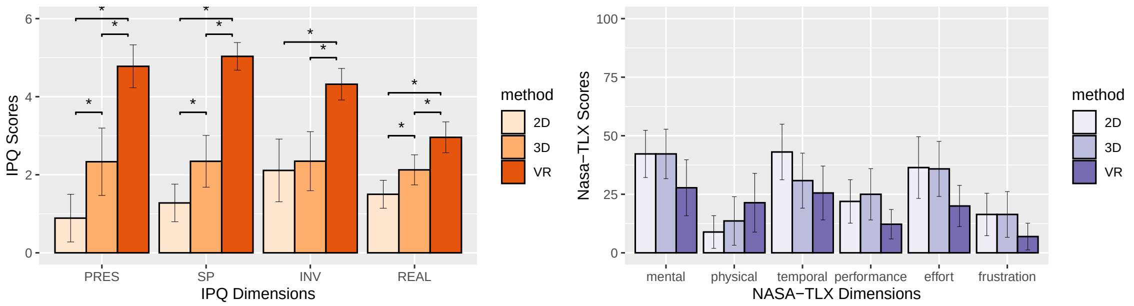

Figure 2: *VRO* led to a significantly higher sense of being there, higher spatial presence, higher involvement, and higher experienced realism than *2DVO* and *3DO*. There were no statistically significant differences in participants' perceived workload when using the different observation methods. Error bars denote the 95% confidence interval (CI).

participants' guesses and their distance to the correct PIN. There was also an interaction effect between threat model × observation method (F(1,83) = 3.319, $p < 0.05$, $\eta_p^2 = 0.07$). Post-hoc Bonferroni adjusted analysis did not confirm the interaction effect, with all pairwise-comparisons being not significant ($p > 0.05$). Follow-up analysis on the main effect of observation method revealed that participants' guesses on ATM authentications were closer to the correct PIN when using *VRO* (M=0.074, SD=0.250) and *2DVO* (M=0.097, SD=0.288) compared to *3DO* (M=0.278, SD=0.470) ($p < 0.05$).

Smartphone PIN Authentication: Participants' observation performance was M=77.78% (SD=30.34%) for *2DVO*, M=69.44% (SD=36.41%) for *3DO*, and M=83.82% (SD=26.74%) for *VRO*. There was a significant effect of observation method (F(1,83) = 4.95, $p < 0.05$, $\eta_p^2 = 0.11$) and threat model (F(1,83) = 6.69, $p < 0.05$, $\eta_p^2 = 0.07$) on the mean Levenshtein distance. Participants' guesses in *VRO* were closer to the correct PIN (M=0.265, SD=0.448) than in *3DO* (M=0.648, SD=0.867) ($p < 0.05$). There were no significant differences between the other pairs (*2DVO*: M=0.403, SD=0.685).

Pattern Smartphone Authentication: Participants' observation performance was M=95.83% (SD=14.02%) for *2DVO*, M=91.67% (SD=22.36%) for *3DO*, and M=100.00% (SD=0.00%) for *VRO*. There was a significant effect of observation method (F(1,83) = 3.21, $p < 0.05$, $\eta_p^2 = 0.07$) and threat model (F(1,83) = 25.53, $p < 0.05$, $\eta_p^2 = 0.24$) on the mean Levenshtein distance. Participants' guesses in *VRO* were closer to the correct pattern (M=0.00, SD=0.00) than in *3DO* (M=0.139, SD=0.371) ($p < 0.05$). There were no significant differences between the other pairs (*2DVO*: M=0.083, SD=0.305).

Summary: Observation Performance

The Levenshtein distances confirmed the differences in participants' observation performance between *VRO* and *3DO*, but not between *VRO* and *2DVO*. *VRO* resulted in the most accurate observations, followed by *2DVO*.

5.2 Sense of Presence (IPQ)

There was a significant effect of observation method on the overall IPQ scores (F(2,34) = 71.429, $p < 0.05$, $\eta_p^2 = 0.81$). Post-hoc analysis confirmed that the sense of presence was significantly higher in *VRO* (M=4.22, SD=1.76) than in *3DO* (M=2.28, SD=1.93) and *2DVO* (M=1.55, SD=1.77) ($p < 0.05$). The difference between *3DO* and *2DVO* was also significant ($p < 0.05$). Fig. 2 shows an overview of the results, featuring the subscales 1) sense of being there (PRES), 2) spatial presence (SP), 3) involvement (INV), 4) experienced realism (REALISM). We followed up with a more nuanced analysis on the level of each subscale.

Sense of being there. The observation methods elicited statistically significant changes in participants' sense of being (F(2,34) = 31.932, $p < 0.05$, $\eta_p^2 = 0.65$). Post-hoc analysis revealed a statistically significant lower sense of being in *2DVO* (M=0.88, SD=1.45) and in *3DO* (M=2.33, SD=2.14) compared to *VRO* (M=4.78, SD=1.55) ($p < 0.05$). The difference between *2DVO* and *3DO* was also statistically significant ($p < 0.05$).

Spatial presence. Participants' experienced spatial presence differed statistically significantly between the different observation methods (F(2,34) = 59.61, $p < 0.05$, $\eta_p^2 = 0.78$). Post-hoc analysis revealed statistically significant differences in participants' spatial presence in *2DVO* (M=1.28, SD=1.48) and in *3DO* (M=2.34, SD=1.99) compared to *VRO* (M=5.03, SD=1.18) ($p < 0.05$). The difference between *2DVO* and *3DO* was also significant ($p < 0.05$).

Involvement. Participants' experienced involvement was statistically significantly different in the different observation methods (F(2,34) = 20.592, $p < 0.05$, $\eta_p^2 = 0.55$). Post-hoc analysis revealed statistically significant differences in *2DVO* (M=2.11, SD=2.15) and in *3DO* (M=2.35, SD=1.91) compared to *VRO* (M=4.32, SD=1.46) ($p < 0.05$). There is no evidence that participants' experienced involvement differed statistically between *2DVO* and *3DO*.

Experienced Realism. Participants' experienced realism was statistically significantly different between the different observation methods (F(2,34) = 23.944, $p < 0.05$, $\eta_p^2 = 0.58$). Post-hoc analysis revealed statistically significant differences in participants' experienced realism in *2DVO* (M=1.50, SD=1.64) and in *3DO* (M=2.13, SD=1.83) compared to *VRO* (M=2.96, SD=1.98) ($p < 0.05$). The difference between *2DVO* and *3DO* was also significant ($p < 0.05$).

Summary: Sense of Presence

VRO led to a significant higher sense of being part of the virtual environment, to a higher spatial presence, and to a higher feeling of involvement and experienced realism than *2DVO* and *3DO*.

5.3 Perceived Workload (NASA-TLX)

Shapiro-Wilk tests of normality indicated that participants' perceived workload when experiencing the different observation methods follows a normal distribution on the level of each observation method. Therefore, we did not perform an aligned rank transformation. Mauchly's test of sphericity indicated that the assumption of sphericity had not been violated, $\chi^2(2) = 3.255$, p=0.196. Participants' perceived workload was statistically significantly different between the observation methods, F(2,34) = 4.715, $p < 0.05$, $\eta_p^2 = 0.217$, but post-hoc analysis with Bonferroni adjustment did not confirm the significant differences ($p > 0.05$). The mean values of participants' perceived workload are M=28.15 (SD=15.77) for *2DVO*, M=27.31 (SD=14.61) for *3DO*, and M=18.98 (SD=17.62) for *VRO*. Fig. 2 shows the mean NASA-TLX values for each dimension.

Summary: Perceived Workload

There is no evidence that *VRO* or *3DO* led to a higher workload than *2DVO*, suggesting that participants' differences in perceived workload when using *2DVO*, *VRO*, and *3DO* are negligible.

5.4 Semi-structured Interviews

We concluded our study with semi-structured interviews to a) shed more light on participants' perception and performance when using the different observation methods and b) better understand their perceived differences to shoulder surfing in the wild. We transcribed the interview data and split participants' statements into meaningful excerpts. This process resulted in overall N=292 participant statements, which we then systematically clustered using an affinity diagram. The initial clustering was performed by the lead researcher. A second researcher then performed an independent review of the clustering and added tags to clusters that required another iteration. Both researchers then met to discuss the clustering and to resolve any discussion points that came up during the review process. Through this process, we identified five themes: 1) Observation Methods' Unique Characteristics, 2) *VRO* for More Realistic Shoulder Surfing Experiences, 3) Lab vs Real-World Observations, 4) The Differences Between the Authentication Scenarios, and 5) General Comments. Below, we discuss those that are particularly relevant for the scope of our research in more detail. Reporting the number of participants who shared certain opinions would be inaccurate due to the use of a semi-structured interview approach and the study's exploratory nature. Thus, we do not include frequencies. Quotes are translated from German to English where necessary.

5.4.1 Observation Methods' Unique Characteristics

We noticed that *VRO* contributed to a close-to-reality looking over someone's shoulder experience. Although *3DO* provided participants with a more realistic shoulder surfing experience than *2DVO*, the mouse-keyboard interaction impacted participants' observation performance. Consequently, the "plug-and-play" characteristic of *2DVO* resulted in observations being easier than *3DO*. P11 mentioned that in *VRO* "[they] could position [themselves] in a way how they wanted it and it was super easy to select the position; this was more difficult with keyboard/mouse" (P11). Others mentioned that in *VRO* "[you] just need to walk to a specific position" (P17). Regarding *3DO*, participants mentioned that their experience was closer-to-reality than *2DVO* because "it felt more like that [they] really want to look over someone's shoulder" (P15). P7 mentioned that "they could experiment a bit like in the real world where you can observe [the authentication] from different perspectives." (P7). Although the lack of manipulations was raised by some participants in *2DVO*, there was a general consensus that it was easier to observe authentications in *2DVO* than in *3DO*. Participants mentioned that the observation position + angle provided them with a clear line of sight and that their only task was to watch the authentication recording. In fact, some participants mentioned they found the videos more realistic because they used *VRO* and *3DO* "in a way to really abuse them" (P9), resulting in some unusual observation positions.

5.4.2 *VRO* for More Realistic Shoulder Surfing Experiences

In *VRO*, P3 voiced that "the [real] environment would be completely irrelevant; it does not matter if [they are] in a basement, in an attic, outside, or at the sea" (P3), and that they did not feel like being part of an experiment. Others mentioned that "with the VR headset [they] moved within the environment and it felt on a physical way more realistic" (P4). For *3DO*, participants voiced that they did not feel being part of the environment to the same extent as in *VRO* because of the presence of reality and that they were "aware of everything that surrounded [them] in the reality" (P15). P3 explained this based on the fact that they were "sitting in front of the PC and could see stuff on the left and right side that is not related to the [authentication scheme]" (P3). For *2DVO*, participants voiced that their task was only to "watch" the authentications and that they were "very conscious that there is a technical device between [them] and the environment" (P4). The overall qualitative feedback suggests that there are two extremes: While *VRO* contributed towards a reasonably realistic in situ shoulder surfing experience, *2DVO* and *3DO* were considered to be observations from "another world".

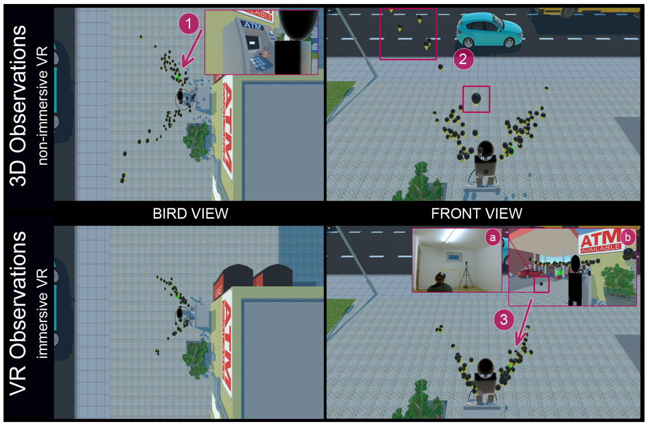

Figure 3: ❶ shows the reference position + orientation of *2DVO*. Participants made use of the absence of physical constraints in *3DO* (see ❷). From the immersive VR observations we noticed that social factors (e.g., the proximity to the user authenticating) lose relevance in such a virtual environment, which we discuss further in Sect. 6.4. ❸ shows a *VR observation* in which the participant pretended to tie their shoes while performing the observation (a); (b) shows the observation position through another perspective.

5.4.3 Lab vs Real-World Observations

Participants reported that they would perform real-world observations similarly as done using *VRO*, e.g., "I can imagine that [real-world observations] work exactly how I did it in VR" (P12). However, across all participants the message was that they would respect the social distances to the user more in the real world. P9 mentioned that "[they] would probably stay further away and do it less conspicuously" (P9). Others voiced that they completely ignored the social factor during the study and "only optimised [their] viewing point" (P10). P4 added that in the real world "there would be other people [and that they] would probably feel being observed" (P4). P13 voiced that in the ATM scenario "the user who withdraws cash probably already acts precautiously – so you would realise when someone stays that close to you." (P13).

In summary, we noticed that while *VRO* contributed to more realistic shoulder surfing experiences than *2DVO*, participants mentioned that users would sense if someone is close to them. In our study, participants did not necessarily consider the social factor (i.e., proxemics [32]) in their observations (see participants' tracked observation positions in Fig. 3, visualised through black dots), which arguably takes on an important role in real-world observations [15].

6 Discussion

We explored how the use of VR can contribute to advanced shoulder surfing research. We found that *VRO* provided participants with a reasonably realistic shoulder surfing experience without negatively impacting their shoulder surfing performance (see Sect. 5.1). Our study showed that *VRO* contribute to a significant higher sense of being in the environment, a greater feeling of spatial presence, a higher level of involvement, and a higher experienced realism than *2DVO* (baseline). While this is an expected finding with the benefits of immersive VR in terms of presence being well-known to the VR community (e.g., [73, 75]), the advantages of *VRO* over *2DVO* are particularly interesting for shoulder surfing research. Our findings imply that previous shoulder surfing studies using 2D videos were not necessarily capable of providing participants with a close-to-reality shoulder surfing experience; therefore, impacting the often desired high ecological validity of usable security research studies [46]. Despite the advantages of *VRO*, our results suggest that *2DVO* are sufficient to assess a system's resilience against observations (see Sect. 5.1). This confirms Aviv et al.'s findings when comparing 2D video recordings with live observations [9]. In all three authentication contexts, there is no evidence that *VRO* were more accurate than *2DVO*. Below, we discuss the impact of *3DO* on shoulder surfing experiments together with participants' observation behaviour in more detail. Participants' observation behaviour was similar across the authentication scenarios. Therefore, we moved the smartphone PIN/pattern visualisations to Appendix B in our supplementary material and discuss participants' observation behaviour on ATM authentications in more detail in Sect. 6.1.

6.1 VR-based Observation Methods: A Blessing and a Curse for Shoulder Surfing Research

From participants' shoulder surfing behaviour (see Fig. 3), we noticed that in *3DO* participants made use of the unique characteristics of non-immersive VR. This is apparent in our study as follows: In *3DO*, participants positioned themselves in several different positions, many of which are challenging to reach in *VRO* due to physical constraints. Although some of these positions seem to be unrealistic at first glance, such observations can indeed happen in the real world using, for example, drones equipped with cameras [79] or surveillance cameras on the corner of a building. In our study, some participants linked their observations to other real-world actions. P7 brought up the example of observing ATM input in an unobtrusive way while tying shoes (Fig. 3-3a). As such, VR-based shoulder surfing studies using *VRO* and *3DO* enable researchers to study different observation strategies in much more detail what can be achieved with traditional *2DVO*.

While our findings suggest that a VR-based research approach can provide researchers with insights into participants' observation strategies, doing this is not necessarily in favour of a critical security evaluation at times where the observation method deviates from a realistic observation (e.g., mouse-keyboard manipulations in *3DO*). Fig. 3 and the qualitative feedback suggest that participants made use of the affordances of *3DO* (e.g., being physically independent), but using *2DVO* and *VRO* led to more accurate observations (see Sect. 5.1). This means that at times where VR-based observation methods are introduced for authentication research (e.g., *3DO*) and the shoulder surfing resilience of a system is at the centre of the investigation, participant-defined observation positions can greatly overestimate a system's resilience against observations. Taking *3DO* and ATM authentication as an example, someone could conclude that observations on ATM authentications are successful in "only" 83.33% observations, while both the de facto standard evaluation approach (i.e., *2DVO* [20,47]) and *VRO* resulted in noticeable more successful observations (*2DVO*: 94.44%, *VRO*: 95.59%). Therefore, researchers risk being mislead into thinking that the system is more resilient against observations than it actually is.

6.2 VR Observation Methods and Their Use Cases

The literature discussed how participants' lack of experience can lead to an under-estimation of risk [82] and emphasised the importance of participants' familiarity with the authentication methods (e.g., [18, 20, 39, 43, 48]). Building upon these discussions, we argue that participants' experience is particularly important when researchers introduce novel observation methods for shoulder surfing research. As evidenced by our semi-structured interviews, *VRO* were perceived as highly realistic. However, the interaction with alternative methods, which differ from participants' real-world observation experiences (e.g., mouse-keyboard manipulations, *3DO*), can have a negative impact on shoulder surfing evaluations and corresponding security conclusions of authentication systems. Still, in cases where the focus is more on an exploratory shoulder surfing evaluation such as studying participants' observation behaviour and their observation strategies, shoulder surfing methods such as *3DO* can be particularly helpful because they enable researchers to study situations that are challenging to research using other means.

6.3 VR Studies and Research in the Wild

It is important to acknowledge that *VR studies* should not, at any point, replace traditional real-world lab or field studies, but rather complement them [44,47,49]. As put by Mäkelä et al. [44], "VR field studies situate between lab studies and real-world field studies, being closer to field studies in ecological validity, and closer to lab studies with regards to their required effort". Virtual simulations make it "easy to experiment with different physical display configurations, e.g., layouts, shapes, sizes and locations" [44]. In a similar vein, we showed how VR enables researchers to study human shoulder surfing on authentication schemes in several contexts without much additional effort. Studying all three authentication contexts in the wild would require a significant amount of additional hardware (e.g., tracking sensors, cameras) and is often infeasible to do due to the nature of private and sensitive contexts [19]. While the usable security community often expects in the wild research to increase the generalisability and the ecological validity of research findings [46], it has been argued that "we [as a community] just need to be a little bit more open to what sort of solutions/evaluations we are expecting out of [something] that has not actually been deployed in the real world." [46]. *VR studies* [44,49] can be particularly helpful to further contribute to more realistic authentication research and studies of that type can be particularly promising when researchers aim to run a large number of consecutive experiments. It is often easier to maintain such virtual environments and make adjustments (e.g., change lighting conditions, replace authentication systems). Virtual artefacts are also easier to store, reuse, deploy, and share because they do not require physical storage space [44, 46].

We believe that VR replications are particularly promising for usable security and privacy research when the targeted real-world space is not available, which is not unlikely when conducting research in relatively sensitive and private contexts (e.g., studying ATM authentication behaviour [19] or security systems at airports [64]).

6.4 Lessons Learned and Recommendations

We outline four lessons learned and recommendations to support and guide researchers in future VR-based shoulder surfing studies.

Recommendation #1: Account For Real-World Factors if They are of Relevance and Consider How the Corresponding Research Findings Transfer to the Real World. The use of VR can greatly advance shoulder surfing research by enabling researchers to get insights into participants' observation strategies. However, results from such *VR studies* also highly depend on how well reality is emulated (e.g., proxemics [32], additional bystanders [40]). We encourage researchers to control for proxemics [32] in virtual environments if social factors are of relevance to the research question. Contrary to prior work that found users' perception of personal space in the real world is similar to that in a virtual environment [10, 36], we noticed that at times where participants optimise their shoulder surfing observations, social factors and the proximity to the user authenticating lose relevance and may even be ignored by participants. There are several directions where future work is called. For example, we encourage future work to consider detection mechanisms that inform participants during their observations when they are in the user's field of view. In cases where the user authenticating would be aware of an observation, participants may want to reconsider their observation position to perform less conspicuous observations (as reported by P9 and P10 in Sect. 5.4.3). At this point, it is important to consider the existing community discussions when aiming for close-to-reality shoulder surfing behaviour in virtual environments. Slater [69] argued that the effect of both "place illusion" and "plausibility illusion" (PI) can contribute to realistic behaviour in virtual environments and that improved visual realism can enhance realistic behavioural responses [70]. Skarbez et al. [67, 68] argued that PI is "essentially the extent to which a scenario complies with a user's expectations". As put by Weber et al. [80], "there is only little research about the effects of perceived realism in VR and the conducted studies generally show that higher realism goes along with stronger presence". It is important to note that the effect of perceived realism in VR is often relatively small [80] and that a high level of realism does not necessarily imply strong presence [37].

We demonstrated how VR increases participants' perceived shoulder surfing realism, but it is important to keep in mind that hinting at similar behaviour to the real world is, due to the the introduced challenges when conducting security and privacy research in the wild [19, 46], often only possible using qualitative research methods (as done in Sect. 5.4 or in [19, 23]). Conducting similar shoulder surfing studies in the real world (e.g., in different private and sensitive contexts) would go beyond what is ethically and legally possible.

Recommendation #2: Consider How Participants Can Best Be Familiarised With VR Observation Methods. Participants' lack of experience w.r.t. novel shoulder surfing methods can significantly impact their experience, preference, and performance when observing authentications. Even traditional input systems (e.g., mouse-keyboard manipulations) can have a negative impact on participants' experience and performance. Consequently, it is important to introduce participants to novel (VR-based) shoulder surfing methods prior to the data collection as their lack of experience can significantly impact the outcome of a system's shoulder surfing evaluation (e.g., see Sect. 5.1).

Recommendation #3: Consider a VR-Based Shoulder Surfing Approach When the Aim is to Contribute Towards Reasonably "Realistic" Shoulder Surfing Experiences, but Keep *2DVO* as a Baseline Measure. As evidenced through our participants' qualitative feedback and the IPQ scores (see Sect. 5.4 and Sect. 5.2), *VRO* leads to more realistic shoulder surfing experiments compared to using *2DVO*. However, traditional *2DVO* already provide a suitable baseline measure for a system's resilience against observations [9]. While novel shoulder surfing methods (e.g., *3DO*, *VRO*) may be used to contribute towards more realistic shoulder surfing experiences and increase participants' sense of being part of the shoulder surfing environment, they do not necessarily outperform traditional *2DVO*. It is important to set clear expectations and identify at the beginning of the research whether or not it is useful to employ a VR-based research approach when studying shoulder surfing. In situations where investigations in the wild are infeasible, VR-based shoulder surfing research can be particularly promising, but to make results more tangible, and to support replication studies and comparisons to prior works, we recommend to keep state-of-the-art 2D video observations (i.e., *2DVO*) as a baseline condition.

Recommendation #4: Use VR to Study Shoulder Surfing in Contexts that are Challenging to Access in the Real World. VR-based shoulder surfing studies are not an alternative to real-world research, but rather complement and advance lab studies by enabling researchers to study scenarios that are otherwise challenging to access (e.g., ATM authentication [18, 19, 22]). In such situations, using VR for human-centred shoulder surfing research can be particularly valuable as such a research approach does not require having physical access to private and sensitive contexts and gives researchers more control of the study environments (e.g., high internal validity, more consistency across participants). Virtual environments are often more affordable and faster to build, deploy, and evaluate than corresponding real-world scenarios [44]. The use of VR as a testbed for human-centred research can be particularly promising at times where pandemics (e.g., COVID-19) significantly impact the safety and well-being of people. While our initial investigation of using VR for shoulder surfing research on authentication systems took place in the lab, we encourage future work to look at more distributed research approaches [49]. While remote (virtual/augmented reality) experiments introduce practical and ethical concerns [71], they can, as put by Steed et al. [72], "continue to forge forward with experimental work".

7 Future Research Directions

We explored the strengths and weaknesses of 3D VR recordings for shoulder surfing research, which we compared to state-of-the-art shoulder surfing evaluations using 2D video recordings. We were particularly interested in participants' shoulder surfing behaviour and how participants exploit VR's unique affordances when performing observation attacks on user authentications. However, we did not account for the many additional factors (e.g., shoulder surfing users when interacting with different devices such as tablets [57], or situations in which shoulder surfing defense strategies are applied [42]). We leave this to future work. Similar to the work by Aviv et al. [8] we did not study text-based authentication, mainly because traditional PIN and pattern authentications are the most commonly used baselines measures in shoulder surfing and authentication research (e.g., [20, 31, 39]). Future research may apply 3D VR recordings for the evaluation of multimodal authentication schemes (e.g., gaze + touch/mid-air [5, 39]). Furthermore, we used a non-vivid environment (e.g., no additional bystanders) to immerse participants into different authentication scenarios. We did this because one key factor of shoulder surfing research on authentication systems is to provide participants (in the role of observers) with a best-case scenario when observing authentications (e.g., [11, 39, 63, 78]). More vivid contexts may led to an even more realistic atmosphere, which forms an interesting future research direction. Finally, a photorealistic VR environment may further increase the visual realism of such a virtual environment. However, recording such sensitive and private contexts as studied in our work is often infeasible to do in the wild. For example, creating 360° real-world recordings as done in the work by Saad et al. [62] introduces ethical and legal challenges in the context of ATM authentication. Such recordings are also limited to what is actually possible to stage/record in the real world. Virtual replications are particularly promising at this point because they provide researchers with more flexibility in changing parts of the environment [44] and enable researchers to study scenarios that are challenging (or even impossible) to access in the real world.

8 Conclusion

We introduced non-immersive and immersive VR observations to advance lab-based shoulder surfing research. We demonstrated how VR and its unique affordances can be applied in the human-centred security research domain to study shoulder surfing in different authentication scenarios. We showed that immersive VR recordings provide participants with a reasonably realistic human shoulder surfing experience without impacting their observation performance compared to commonly used 2D video recordings. Through our investigation of using VR for shoulder surfing research, we hope to contribute to more realistic human-centred security research in the long run and encourage future work to find ways to further improve lab-based usable security and privacy research using VR.

Acknowledgments

We thank all participants for taking part in our research. We also thank all reviewers whose comments significantly improved the paper. This publication was supported by the University of Edinburgh and the University of Glasgow jointly funded PhD studentships, and partially by the 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).

REFERENCES

[1] 3d atm model, 2019. https://free3d.com/3d-model/atm-57251.html, accessed 04 November 2021.

[2] 3d smartphone model, 2021. https://free3d.com/3d-model/iphonex-113534.html, accessed 04 November 2021.

[3] U. 3D. User manual, 2021. https://docs.unity3d.com/Manual/class-LineRenderer.html, accessed 04 November 2021.

[4] Y. Abdelrahman, M. Khamis, S. Schneegass, and F. Alt. Stay cool! understanding thermal attacks on mobile-based user authentication. In *Proc. of the 2017 CHI Conf. on Human Factors in Computing Systems*, CHI '17. ACM, New York, NY, USA, 2017.

[5] Y. Abdrabou, M. Khamis, R. M. Eisa, S. Ismael, and A. Elmougy. Engage: Resisting shoulder surfing using novel gaze gestures authentication. In *Proc. of the 17th International Conf. on Mobile and Ubiquitous Multimedia*. ACM, New York, NY, USA, 2018.

[6] Y. Abdrabou, M. Khamis, R. M. Eisa, S. Ismail, and A. Elmougy. Just gaze and wave: Exploring the use of gaze and gestures for shoulder-surfing resilient authentication. In *Proc. of the ACM Symp. on Eye Tracking Research & Applications*. ACM, New York, NY, USA, 2019.

[7] T. Amano, S. Kajita, H. Yamaguchi, T. Higashino, and M. Takai. Smartphone applications testbed using virtual reality. In *Proc. of the 15th EAI International Conf. on Mobile and Ubiquitous Systems: Computing, Networking and Services*, MobiQuitous '18. ACM, New York, NY, USA, 2018.

[8] A. J. Aviv, J. T. Davin, F. Wolf, and R. Kuber. Towards baselines for shoulder surfing on mobile authentication. In *Proc. of the 33rd Annual Computer Security Applications Conference*, ACSAC 2017. ACM, New York, NY, USA, 2017.

[9] A. J. Aviv, F. Wolf, and R. Kuber. Comparing video based shoulder surfing with live simulation. In *Proc. of the Computer Security Applications Conf.*, ACSAC '18. ACM, New York, NY, USA, 2018.

[10] J. N. Bailenson, J. Blascovich, A. C. Beall, and J. M. Loomis. Equilibrium theory revisited: Mutual gaze and personal space in virtual environments. *Presence*, 2001.

[11] A. Bianchi, I. Oakley, and D. S. Kwon. Spinlock: A single-cue haptic and audio pin input technique for authentication. In *Haptic and Audio Interaction Design*. Springer, Berlin, Heidelberg, 2011.

[12] J. Blascovich, J. Loomis, A. C. Beall, K. R. Swinth, C. L. Hoyt, and J. N. Bailenson. Immersive virtual environment technology as a methodological tool for social psychology. *Psychological Inquiry*, 2002.

[13] L. Bošnjak and B. Brumen. Shoulder surfing experiments: A systematic literature review. *Computers & Security*, 2020.

[14] J. Brooke. Sus: a ”quick and dirty” usability. 1996.

[15] F. Brudy, D. Ledo, S. Greenberg, and A. Butz. Is Anyone Looking? Mitigating Shoulder Surfing on Public Displays through Awareness and Protection. In *Proc. of The International Symposium on Pervasive Displays*, PerDis '14. ACM, New York, NY, USA, 2014.

[16] J. Cohen. Eta-squared and partial eta-squared in fixed factor anova designs. *Educational and Psychological Measurement*, 1973.

[17] J. Cohen. *Statistical power analysis for the behavioral sciences*. Academic press, 2013.

[18] A. De Luca, K. Hertzschuch, and H. Hussmann. Colorpin: Securing pin entry through indirect input. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '10. ACM, New York, NY, USA, 2010.

[19] A. De Luca, M. Langheinrich, and H. Hussmann. Towards understanding atm security: A field study of real world atm use. In *Proc. of the 6th Symposium on Usable Privacy and Security*, SOUPS '10. ACM, New York, NY, USA, 2010.

[20] A. De Luca, E. von Zezschwitz, N. D. H. Nguyen, M.-E. Maurer, E. Rubegni, M. P. Scipioni, and M. Langheinrich. Back-of-device authentication on smartphones. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '13. ACM, New York, NY, USA, 2013.

[21] A. De Luca, E. von Zezschwitz, L. Pichler, and H. Hussmann. Using Fake Cursors to Secure On-Screen Password Entry. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '13. ACM, New York, NY, USA, 2013.

[22] P. Dunphy, A. Fitch, and P. Olivier. Gaze-contingent passwords at the atm. In *Communication by Gaze Interaction (COGAIN)*, 2008.

[23] M. Eiband, M. Khamis, E. von Zezschwitz, H. Hussmann, and F. Alt. Understanding shoulder surfing in the wild: Stories from users and observers. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '17. ACM, New York, NY, USA, 2017.

[24] L. A. Elkin, M. Kay, J. J. Higgins, and J. O. Wobbrock. An aligned rank transform procedure for multifactor contrast tests, 2021.

[25] U. Erra, D. Malandrino, and L. Pepe. Virtual reality interfaces for interacting with three-dimensional graphs. *International Journal of Human–Computer Interaction*, 2019.

[26] M. Feick, N. Kleer, A. Tang, and A. Krüger. The virtual reality questionnaire toolkit. UIST Adjunct. ACM, New York, NY, USA, 2020.

[27] S. M. Fiore, G. W. Harrison, C. E. Hughes, and E. E. Rutström. Virtual experiments and environmental policy. *Environmental Economics and Management*, 2009.

[28] L. Freina and M. Ott. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The international scientific Conf. elearning and software for education*, 2015.

[29] S. Garfinkel and H. R. Lipford. Usable security: History, themes, and challenges. *Synthesis Lectures on Information Security, Privacy, and Trust*, 2014.

[30] C. George, M. Khamis, D. Buschek, and H. Hussmann. Investigating the third dimension for authentication in immersive virtual reality and in the real world. In *2019 IEEE Conf. on Virtual Reality and 3D User Interfaces (VR)*, March 2019.

[31] C. George, M. Khamis, E. von Zezschwitz, M. Burger, H. Schmidt, F. Alt, and H. Hussmann. Seamless and secure vr: Adapting and evaluating established authentication systems for virtual reality. In *Network and Distributed System Security Symposium (NDSS 2017)*, USEC '17. NDSS, February 2017.

[32] E. T. Hall. *The hidden dimension*. Garden City, NY: Doubleday, 1966.

[33] S. Hart and L. Staveland. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Human mental workload*, 1988.

[34] S. G. Hart. Nasa-task load index (nasa-tlx); 20 years later. In *Proc. of the human factors and ergonomics society annual meeting*. Sage publications Sage CA: Los Angeles, CA, 2006.

[35] M. Hassenzahl, M. Burmester, and F. Koller. Attrakdiff: A questionnaire to measure perceived hedonic and pragmatic quality. In *Mensch & Computer*, 2003.

[36] H. Hecht, R. Welsch, J. Viehoff, and M. R. Longo. The shape of personal space. *Acta Psychologica*, 2019.

[37] M. Hofer, T. Hartmann, A. Eden, R. Ratan, and L. Hahn. The role of plausibility in the experience of spatial presence in virtual environments. *Frontiers in Virtual Reality*, 2020.

[38] T. hundred fifty-five (255) pixel studios. City package, 2021. https://assetstore.unity.com/packages/3d/environments/urban/city-package-107224, accessed 04 November 2021.

[39] M. Khamis, F. Alt, M. Hassib, E. von Zezschwitz, R. Hasholzner, and A. Bulling. Gazetouchpass: Multimodal authentication using gaze and touch on mobile devices. In *Proc. of the 34th Annual ACM Conf. Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16. ACM, New York, NY, USA, 2016.

[40] M. Khamis, L. Bandelow, S. Schick, D. Casadevall, A. Bulling, and F. Alt. They are all after you: Investigating the viability of a threat model that involves multiple shoulder surfers. In *Proc. of the 16th International Conf. on Mobile and Ubiquitous Multimedia*, MUM '17. ACM, New York, NY, USA, 2017.

[41] M. Khamis, L. Trotter, V. Mäkelä, E. v. Zezschwitz, J. Le, A. Bulling, and F. Alt. Cueauth: Comparing touch, mid-air gestures, and gaze for cue-based authentication on situated displays. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, Dec. 2018.

[42] H. Khan, U. Hengartner, and D. Vogel. Evaluating attack and defense strategies for smartphone pin shoulder surfing. In *Proc. of the 2018 CHI Conf. on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2018.

[43] L. Kraus, R. Schmidt, M. Walch, F. Schaub, and S. Möller. On the use of emojis in mobile authentication. In S. De Capitani di Vimercati and F. Martinelli, eds., *ICT Systems Security and Privacy Protection*. Springer International Publishing, Cham, 2017.

[44] V. Mäkelä, S. R. R. Rivu, S. Alsherif, M. Khamis, C. Xiao, L. M. Borchert, A. Schmidt, and F. Alt. Virtual Field Studies: Conducting Studies on Public Displays in Virtual Reality. In *Proc. of the 38th Annual ACM Conf. on Human Factors in Computing Systems*, CHI '20. ACM, New York, NY, USA, 2020.

[45] F. Mathis, K. Vaniea, and M. Khamis. Observing virtual avatars: The impact of avatars' fidelity on identifying interactions. In *Proc. of the 24th International Conf. on Academic Mindtrek*, AcademicMindtrek '21. ACM, New York, NY, USA, 2021.

[46] F. Mathis, K. Vaniea, and M. Khamis. Prototyping usable privacy and security systems: Insights from experts. *International Journal of Human–Computer Interaction*, 2021.

[47] F. Mathis, K. Vaniea, and M. Khamis. Replicueauth: Validating the use of a lab-based virtual reality setup for evaluating authentication systems. In *Proc. of the 39th Annual ACM Conf. on Human Factors in Computing Systems*, CHI '21. ACM, New York, NY, USA, 2021.

[48] F. Mathis, J. H. Williamson, K. Vaniea, and M. Khamis. Fast and secure authentication in virtual reality using coordinated 3d manipulation and pointing. *ACM Trans. Comput.-Hum. Interact.*, Jan. 2021.

[49] F. Mathis, X. Zhang, J. O'Hagan, D. Medeiros, P. Saeghe, M. McGill, S. Brewster, and M. Khamis. Remote xr studies: The golden future of hci research? In *CHI 2021 Workshop on XR Remote Research*, 2021.

[50] L. Motion. Leap motion, 2019. accessed 04 November 2021.

[51] J. O'Hagan and J. R. Williamson. Reality aware vr headsets. In *Proc. of the 9TH ACM International Symposium on Pervasive Displays*, PerDis '20. ACM, New York, NY, USA, 2020.

[52] J. O'Hagan, J. R. Williamson, M. McGill, and M. Khamis. Safety, power imbalances, ethics and proxy sex: Surveying in-the-wild interactions between vr users and bystanders. In *2021 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 2021.

[53] T. D. Parsons. Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in Human Neuroscience*, 2015.

[54] S. Pedram, R. Skarbez, S. Palmisano, M. Farrelly, and P. Perez. Lessons learned from immersive and desktop vr training of mines rescuers. *Frontiers in Virtual Reality*, 2021.

[55] S. Putze, D. Alexandrovsky, F. Putze, S. Höffner, J. D. Smeddinck, and R. Malaka. Breaking the experience: Effects of questionnaires in vr user studies. In *Proc. of the 2020 CHI Conf. on Human Factors in Computing Systems*, CHI '20. ACM, New York, NY, USA, 2020.

[56] Qualtrics. Qualtrics, 2005. accessed 04 November 2021.

[57] K. Ragozin, Y. S. Pai, O. Augereau, K. Kise, J. Kerdels, and K. Kunze. Private Reader: Using Eye Tracking to Improve Reading Privacy in Public Spaces. In *Proc. of the 21st International Conf. on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '19. ACM, New York, NY, USA, 2019.

[58] F. Rebelo, P. Noriega, E. Duarte, and M. Soares. Using virtual reality to assess user experience. *Human Factors*, 2012.

[59] G. Robertson, S. Card, and J. Mackinlay. Three views of virtual reality: nonimmersive virtual reality. *Computer*, 1993.

[60] RockVR. Vr capture, 2021. https://assetstore.unity.com/packages/tools/video/vr-capture-75654, accessed 04 November 2021.

[61] V. Roth, K. Richter, and R. Freidinger. A pin-entry method resilient against shoulder surfing. In *Proc. of the 11th ACM Conf. on Computer and Communications Security*. ACM, New York, NY, USA, 2004.

[62] A. Saad, J. Liebers, U. Gruenefeld, F. Alt, and S. Schneegass. Understanding bystanders' tendency to shoulder surf smartphones using 360-degree videos in virtual reality. 2018.

[63] H. Sasamoto, N. Christin, and E. Hayashi. Undercover: Authentication usable in front of prying eyes. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*. ACM, New York, NY, USA, 2008.

[64] M. A. Sasse. Red-eye blink, bendy shuffle, and the yuck factor: A user experience of biometric airport systems. *IEEE Security Privacy*, 2007.

[65] G.-L. Savino, N. Emanuel, S. Kowalzik, F. Kroll, M. C. Lange, M. Laudan, R. Leder, Z. Liang, D. Markhabayeva, M. Schmeißer, N. Schütz, C. Stellmacher, Z. Xu, K. Bub, T. Kluss, J. Maldonado, E. Kruijff, and J. Schöning. Comparing pedestrian navigation methods in virtual reality and real life. In *2019 International Conf. on Multimodal Interaction*, ICMI '19. ACM, New York, NY, USA, 2019.

[66] T. Schubert, F. Friedmann, and H. Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 2001.

[67] R. Skarbez, F. P. Brooks, Jr., and M. C. Whitton. A survey of presence and related concepts. *ACM Comput. Surv.*, Nov. 2017.

[68] R. Skarbez, J. Gabbard, D. A. Bowman, T. Ogle, and T. Tucker. Virtual replicas of real places: Experimental investigations. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[69] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2009.

[70] M. Slater, P. Khanna, J. Mortensen, and I. Yu. Visual realism enhances realistic response in an immersive virtual environment. *IEEE computer graphics and applications*, 2009.

[71] A. Steed, S. Frlston, M. M. Lopez, J. Drummond, Y. Pan, and D. Swapp. An 'in the wild' experiment on presence and embodiment using consumer virtual reality equipment. *IEEE Transactions on Visualization and Computer Graphics*, 2016.

[72] A. Steed, F. Ortega, A. Williams, E. Kruijff, W. Stuerzlinger, A. Batmaz, A. Won, E. Rosenberg, A. Simeone, and A. Hayes. Evaluating immersive experiences during covid-19 and beyond. 2020.

[73] A. Steed and R. Schroeder. Collaboration in immersive and non-immersive virtual environments. In *Immersed in Media*. 2015.

[74] TI. Ultimatereplay, 2021. https://assetstore.unity.com/packages/tools/camera/ultimate-replay-2-0-178602, accessed 04 November 2021.

[75] S. Ventura, E. Brivio, G. Riva, and R. M. Baños. Immersive versus non-immersive experience: Exploring the feasibility of memory assessment through 360 technology. *Frontiers in psychology*, 2019.

[76] A. Voit, S. Mayer, V. Schwind, and N. Henze. Online, VR, AR, Lab, and In-Situ: Comparison of Research Methods to Evaluate Smart Artifacts. In *Proc. of the 2019 CHI Conf. on Human Factors in Computing Systems*, CHI '19. ACM, New York, NY, USA, 2019.

[77] M. Volkamer, A. Gutmann, K. Renaud, P. Gerber, and P. Mayer. Replication study: A cross-country field observation study of real world {PIN} usage at atms and in various electronic payment scenarios. In *Symposium on Usable Privacy and Security (SOUPS)*, 2018.

[78] E. von Zezschwitz, A. De Luca, B. Brunkow, and H. Hussmann. Swipin: Fast and secure pin-entry on smartphones. In *Proc. of the 33rd Annual ACM Conf. on Human Factors in Computing Systems*, CHI '15. ACM, New York, NY, USA, 2015.

[79] Y. Wang, H. Xia, Y. Yao, and Y. Huang. Flying eyes and hidden controllers: A qualitative study of people's privacy perceptions of civilian drones in the us. *Proc. on Privacy Enhancing Tech.*, 2016.

[80] S. Weber, D. Weibel, and F. W. Mast. How to get there when you are there already? defining presence in virtual reality and the importance of perceived realism. *Frontiers in Psychology*, 2021.

[81] M. Weiß, K. Angerbauer, A. Voit, M. Schwarzl, M. Sedlmair, and S. Mayer. Revisited: Comparison of empirical methods to evaluate visualizations supporting crafting and assembly purposes. *IEEE Transactions on Visualization and Computer Graphics*, 2020.

[82] O. Wiese and V. Roth. Pitfalls of shoulder surfing studies. In *NDSS Workshop on Usable Security*, 2015.

[83] O. Wiese and V. Roth. See you next time: A model for modern shoulder surfers. In *Proc. of the 18th International Conf. on Human-Computer Interaction with Mobile Devices and Services*, MobileHCI '16. ACM, New York, NY, USA, 2016.

[84] J. O. Wobbrock, L. Findlater, D. Gergle, and J. J. Higgins. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proc. of the SIGCHI Conf. on Human Factors in Computing Systems*, CHI '11. ACM, New York, NY, USA, 2011.

[85] E. Wu, M. Piekenbrock, T. Nakumura, and H. Koike. Spinpong - virtual reality table tennis skill acquisition using visual, haptic and temporal cues. *IEEE Transactions on Visualization and Computer Graphics*, 2021.

[86] N. H. Zakaria, D. Griffiths, S. Brostoff, and J. Yan. Shoulder surfing defence for recall-based graphical passwords. In *Proc. of the 7th Symp. on Usable Privacy and Security*, SOUPS '11. ACM, New York, NY, USA, 2011.