Temporal Effects in Motion Behavior for Virtual Reality (VR) Biometrics

Robert Miller* Natasha Kholgade Banerjee† Sean Banerjee‡

Clarkson University, Potsdam, NY

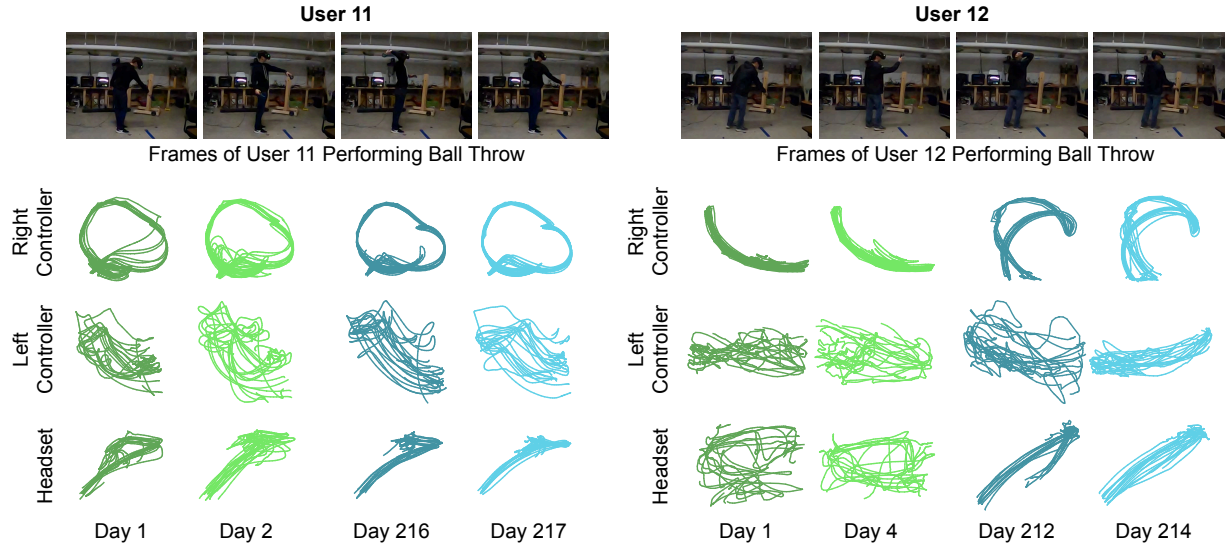

Figure 1: We study how behavior-based authentication and identification in VR is impacted by variability in VR behavior trajectories across short, medium, and long timescales. While users show limited changes in behavior over short timescales, they show differences in the extent to which they maintain consistency over longer timescales. User 11 demonstrates consistent motion trajectories for the VR controllers and headset for a ball-throwing action. User 12 changes their ball-throwing approach from an underhand motion on days 1 and 4 to an overhand throw several months later on days 212 and 214. We show that using long timescale data to train users enhances learning-based identification and authentication using user behavior over multiple timescales.

Abstract

Using the motion behavior of users in virtual reality (VR) as a biometric signature has the potential to enable continuous identification and authentication of users without compromising VR applications if traditional passwords are acquired by malicious agents. Users exhibit natural variabilities in behavior over time that influence their body motions and can alter the trajectories of VR devices such as the headset and the controllers. Behavior variabilities may negatively impact the success rate of VR biometrics. In this work, we evaluate how deep learning approaches to match input and enrollment trajectories are influenced by user behavior variation over varying time scales. We demonstrate that over short timescales on the order of seconds to minutes, no statistically significant relationship is found in the temporal placement of enrollment trajectories and their matches to input trajectories. We find that on medium-scale separation between enrollment and input trajectories, on the order of days to weeks, median accuracy is similar within users who provide input close and distant to enrollment data. Over long timescales on the order of 7 to 18 months, we obtain optimal performance for short and long enrollment/input separations by using training sets from users providing long-timescale data, as these sets encompass coarse and fine-scale changes in behavior.

*e-mail: romille@clarkson.edu
†e-mail:nbanerje@clarkson.edu
‡e-mail:sbanerje@clarkson.edu

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction paradigms—Virtual reality; Security and privacy—Security services—Authentication—Biometrics

1 Introduction

Virtual reality (VR) is seeing adoption across a wide gamut of industries including distance education [7, 15, 18, 45, 53, 54], retail [44, 56], personal finance [6, 55], remote teleoperation and driving [19, 25, 36, 37, 47], and healthcare [5, 8, 9, 24, 26, 38, 49]. With the significant quantity of sensitive data that is likely to be generated in widespread adoption of VR, authenticating users' identities becomes paramount. While authentication using traditional credentials has been explored for VR [2, 3, 10–14, 40, 59], the system is compromised if an unintended agent acquires a user's credentials. To combat the vulnerabilities of traditional authentication, a large number of approaches have emerged recently to explore identifying and authenticating users using their behavior in VR [1, 17, 22, 27–29, 31–35, 39, 43]. These approaches define VR behavior as actions or tasks performed during the use of VR applications, e.g., throwing a ball [1, 17, 32–34], pointing [43], shooting an arrow [22], bowling [22], moving their head to music [35], viewing videos [31], filling questionnaires [31], and manipulating objects [27–29, 39], and record the behavior by tracking the motions of the user's interactions with the devices of a VR system. Behavior-based VR authentication research spans authentication within the same VR system [1, 17, 22, 27–29, 31–35, 39, 43], as well as across VR systems [33, 34], with the latter enabling users to access multiple systems without re-providing enrollment data. Behavior biometrics enable unobtrusive and continuous authentication [29, 32].

The major challenge faced in using behavior-based biometrics for

identification and authentication is that human behavior changes over time. Changes span a variety of timescales, ranging, for instance, over short-term variations due to injury, system acclimatization over medium timescales, and evolution of behavior over long timescales due to aging. Changes may also occur if a user operates in a novel environment, if they wear different clothing, or if they move to a new VR system where they may require time to adjust to the physical characteristics of the new system. The impact of behavior change on authentication has been explored in non-VR environments, e.g., desktop [50], mobile [41, 50, 52], and gait from video [30]. Studies in non-VR domains demonstrate that authentication rates fall with increasing temporal differences [41, 50, 52], and with clothing and footwear variations [30]. No prior work exists on evaluating the influence of behavior change in VR. Conclusions from prior studies in non-VR environments cannot be directly transferred to VR. In contrast to desktop and mobile applications, VR environments involve users performing actions that often mimic real-world interactions, so that acclimatization may take lesser time. In contrast to methods that use videos to capture real-world behavior, where clothing variability can impact performance, VR-based biometrics largely correspond to device trajectories that are clothing or footwear invariant. Given that VR emulates real-world behavior unlike desktop/mobile applications but lacks visual diversities that may impact real-world identification, separate studies are needed to study behavior change in VR.

In this paper, we contribute the first work that studies the impact of timescales of human action on success of VR behavior-based authentication and identification using matching algorithms based on deep neural networks. We use the datasets of Miller et al. [33, 34] and Ajit et al. [1] to measure how short, medium, and long timescales affect identification and authentication effectiveness. The Miller et al. dataset contains 41 users performing the a ball-throwing action using 3 VR systems, with 10 trials per session across two sessions per system. Each session is provided on a separate day. The Ajit et al. dataset contains 33 users performing the same ball-throwing action using a single VR system. Similar to the Miller et al. dataset, the Ajit et al. set has users provide 2 sessions on 2 separate days and 10 ball-throwing action trials per session. 16 common users provide using the HTC Vive for both datasets. Following the protocol of Miller et al., we provide analyses for identification and authentication within the same system and across VR systems through three contributions.

1. To measure impact over short timescales on the order of seconds to minutes, we study the relationship between temporal placements of enrollment trajectories and their matches to input trajectories. Our study enables analyzing whether user behaviors undergo evolution on short time scales from the start of usage to continued interaction, and whether within-session evolutionary behavior impacts trajectory matching. For repeatable actions such as ball-throwing, we find no significant effect of temporal placement of enrollment trajectories in identifying matches for input trajectories.
2. To measure impact over medium timescales on the order of days, we study identification accuracy for two clusters of users in the dataset of Miller et al. We define clusters by identifying users whose days between enrollment and input sessions fall at or below a day threshold for one cluster and above the threshold for the other cluster. We find that median accuracy for users with temporal separation between sessions below and above the threshold is similar.
3. To measure impact over long timescales on the order of weeks and months, we study identification and authentication success for the 16 common users across the datasets of Miller et al. [33] and Ajit et al. [1]. We observe that long time periods introduce coarse alterations to some users, and that users demonstrate inter-user differences in trajectory modification for headsets and controllers. Variable alterations in user interactions with the devices cause matching algorithms trained using short temporal separation in enrollment and input on the order of days to generate low success at matching input data provided several months after the enrollment data. We find that matching algorithms trained to recognize coarse and fine-grained behavior differences using distant enrollment/input separation provide improved success for input data that is close and distant from the enrollment data. We obtain highest success by providing evidence from close and distant enrollment/input separations.

Overall, our results show that user behavior in VR exhibits low variability over short and medium timescales. We find that security mechanisms that use training data on large-scale changes in user behaviors over long time periods can provide high success in behavior-based VR security. We have made all code and data for our work public at `https://git.io/J9GIW`.

2 Related Work

2.1 Behavior-Based Authentication in VR

Early work in behavior-based techniques for VR authentication focused on using head motions for head-mounted wearables such as Google Glass [20, 46, 57] and Google Cardboard [35], with head movement patterns acquired using on-board inertial measurement units in response to music [20], image presentations [46], and presentations of VR targets [35]. Eye blinks have also been used in conjunction with head motion for authentication in Google Glass [46]. Since older head-mounted wearables such as Glass and Cardboard lack involvement of user hands, they are less interactive in comparison to recent VR systems that provide immersive experiences by tracking hand behavior via controllers or vision systems.

The incorporation of hand controllers in VR systems enables users to perform a wide range of interactions in VR such as picking, pointing, swiping, throwing, and turning, and to integrate multiple actions for higher level tasks such as driving, teleoperating, or playing a sport in VR. With increased degrees of freedom offered by the motions of the hand in comparison to the head, recent approaches to VR authentication have explored integrating features from hand and head trajectories to improve performance of VR authentication over early work on using head motion alone. Pfeuffer et al. [43] perform identification for activities such as pointing, grabbing, walking, and typing using a set of hand-chosen features computed from the device trajectories. The features represent the movement patterns of each VR device, pairs of devices, distances between devices, and distances and angles to a target. They use feature selection to determine the best-performing features, and obtain a maximum accuracy of 63.55% for the pointing action using random forests. Kupin et al. [17] perform identification for 14 right-handed users throwing a ball using an HTC Vive when using the motion of the dominant hand controller alone as a biometric. They use the nearest neighbor distance between trajectories to match input trajectories against an enrollment set, and provide a maximum accuracy of 92.86%. Ajit et al. [1] train a perceptron model to obtain the optimal weights for aggregating input-to-enrollment distances between position and orientation features from the headset and the hand controllers of the HTC Vive. They evaluate 21 device-feature combinations and demonstrate a maximum accuracy of 93.03% on a dataset of 33 right-handed users using the best performing features. Miller et al. [33] augment the features for the perceptron in Ajit et al. to include velocity, angular velocity, and trigger grab/release. They show results of identification performance within the same system and across systems for a dataset of 41 users who provide data on the HTC Vive, Oculus Quest, and HTC Vive Cosmos. In their cross-system results, input data is provided on one VR system, and enrollment on a separate VR system. They show accuracies of 91% for within-system identification using the Quest and Cosmos, and 97% using the Vive. Cross-system accuracies are lower between 58% and 85%. The lower cross-system accuracies indicate that system-to-system differences may influence user behavior when users switch systems.

Olade et al. [39] assess identification performance for grabbing, rotation, and dropping activities on a dataset of 25 users using eye location and position and orientation of the headset and hand controllers as features. They evaluate a variety of classifiers such as decision trees, discriminant analysis, support vector machines, logistic regression, *k* nearest neighbors, naive Bayes, and ensemble classification, with and without principal component analysis for dimensionality reduction. They find that *k* nearest neighbors provides the maximum accuracy of 98.6%. Miller et al. [31] demonstrate identification results for a dataset of 511 users using the HTC Vive to watch five 360° videos and answer multiple choice questions on the videos. They evaluate *k* nearest neighbors, random forests, and gradient boosting machines as classifiers, and provide a maximum accuracy of 95% using random forests. The 511 users dataset of Miller et al. [31] is the largest single VR system dataset. However, the study incorporates limited full-body movement found in traditional VR experiences as users remain standing when watching the videos and answering questions. More recent approaches to VR authentication have moved toward using deep learning owing to the success of neural networks at learning arbitrary decision boundaries. Mathis et al. [27] evaluate a variety of network architectures and device combinations to perform identification on a dataset of 23 users using an HTC Vive to enter a PIN number through digits on a Rubik's-like cube. They demonstrate highest and second-highest accuracies of 98.91% and 98.55% using a fully convolutional network and ResNet respectively. Liebers et al. [22] evaluate multi-layer perceptrons and long short-term memory recurrent neural networks (RNNs) for archery and bowling activities performed by 16 users using the Oculus Quest. They demonstrate highest accuracy of 90% for RNNs on the archery activity when height normalization is performed. Miller et al. [34] train Siamese neural networks that facilitate higher cross-system success in comparison to their prior work [33] by learning to characterize inter-system differences. They obtain within-system identification accuracies of 99.75%, 99.51%, and 98.04% for the Quest, Vive, and Cosmos, and cross-system identification accuracies of between 87.82% to 98.53%, with an average improvement of 29.78% over their prior work.

None of the prior work in VR biometrics so far provides a comprehensive analysis of temporal influence on behavior-based authentication. Several approaches [27, 31, 39] analyze performance using enrollment and input data that has been collected in succession where the difference in temporal spacing is within minutes, as opposed to the longer timescales of days, weeks, or months that are more likely to exist between enrollment and input data in a realistic use case. Other approaches do evaluate authentication or identification performance using data over separate days with the explicit goal of assessing authentication over a longer timespan of days or weeks [1,17,23,33,34,43]. Some approaches provide demonstrations of real-time continuous authentication against enrollment collected weeks or months in the past [28, 29, 32]. However, all prior work lacks fine-grained breakdown of performance on short, medium, and long timescales ranging over seconds to minutes, days, and weeks to months. In this work, we bridge the lack of knowledge on the impact of action timescales on continued security by contributing a detailed study of performance on varying timescales.

2.2 Impact of Variability on VR Behavior Biometrics

While prior work in assessing impact of human behavior variability on use of behavior biometrics in VR is lacking, similar work exists for behavior-based authentication in other domains. Much of the work remains focused in modeling and addressing visual changes due to clothing and body carriage [4, 21, 30, 58] for gait-based user recognition through video. Matovski et al. [30] demonstrate that visual variabilities such as clothing and accessory differences play a higher role in lowering recognition in comparison to short or medium temporal separations between data provision. Visual variations in clothing play no role in defining the user's behavior in a VR environment, and clothing is unlikely to affect interactions unless the clothing significantly weights down the user. Visual variations *can* affect the accuracy of VR systems that use vision-based tracking algorithms, however, these effects have less to do with the user's behavior than with the degree to which vision-based tracking algorithms are trained to be invariant to clothing change. While change in carriage are likely to introduce variability, prior work on addressing carriage change in gait-based recognition are unlikely to translate to continuous authentication in VR. Most gait-based work focuses largely on repetitive walking cycles and lacks a full coverage of the dense array of interactions that users perform in VR environments.

Despite the importance of modeling temporal effects of behavior variability on behavior biometrics use, work for desktop and mobile environments is limited. Syed et al. [50] characterize user habituation to a password entry task on a keyboard in terms of the time it takes a user to enter the password. They demonstrate that for complex credentials, users take a longer time in the first few trials, with later trials demonstrating shorter entry times and reduced variability over successive entries. They use variable sizes of training datasets to demonstrate that rate of reduction in equal error rate (EER) for authentication using keystroke dynamics is affected not solely by decreased training data, but also by the presence of habituation, i.e., that keystrokes too far in the past from the current entry point may represent divergent behavior. Syed et al. [51] confirm the findings of habituation in delay and hold time of Syed et al. [50] by analyzing successive windows of keystroke entry patterns and observing fewer variations in key combinations used in later entries. The work of Palaskar et al. [41] and Syed et al. [52] demonstrates similar findings for gesture-based authentication in touch devices such as smartphones, where authentication EER increases significantly when training and testing samples are distanced by more than 600 strokes. They show that retraining classifiers using data closer in time to the test samples reduces EER. Unlike desktop or smartphone environments, the high cross-session security performance demonstrated in VR environments [1, 17, 23, 33, 34, 43] indicates that VR applications structured to leverage user interactions similar to those in the real world may require lesser habituation time.

3 Dataset Analysis

We use the 41-subject dataset of Miller et al. [33, 34] and the 33-subject dataset of Ajit et al. [1] to analyze the temporal effect in VR biometrics. The datasets were collected upon Institutional Review Board (IRB) approval from faculty, staff, and students at a small rural university with approximately 3,500 students. Subjects ranged in age from 18 to 38 years old with varying degrees of experience in VR. Subjects were not provided any financial compensation for providing data for either study. The Miller et al. dataset [33,34] consists of users providing data using multiple VR systems—an HTC Vive, HTC Vive Cosmos, and Oculus Quest—enabling analysis within and across VR systems. The Ajit et al. [1] dataset consists of users providing data using a single VR system, i.e., the HTC Vive. The task in both datasets is identical and consists of a user throwing a ball at a virtual target 10 times on two separate sessions per system. During each session and throw, the physical characteristics and locations of the ball, the target, and the pedestal holding the ball remained constant. Each user session is separated by a minimum of 24 hours. As shown in Table 1, the difference between sessions varies between 1 and 25 days. Of the 41 subjects in the Miller et al. [33, 34] dataset, 16 subjects provided data in the Ajit et al. [1] dataset using the HTC Vive. For the purpose of this paper we consider the 16 subjects as having prior experience in the VR application. For the 16 subjects, on average 344.56 ± 166.30 days separate their first session in the Ajit et al. [1] study and the first session in the Miller et al. [33, 34] study with a maximum of 561 days and a minimum of 214 days. Of

Metric	VA1/VA2 [1]	Q1/Q2 [33,34]	Q2/V1 [33,34]	V1/V2 [33,34]	V2/C1 [33,34]	C1/C2 [33,34]
Mean	4.03 ± 2.38	1.17 ± 0.80	2.27 ± 2.36	3.00 ± 2.07	8.15 ± 8.61	2.05 ± 1.60
Maximum	8.00	6.00	8.00	7.00	25.00	6.00
Minimum	1.00	1.00	1.00	1.00	1.00	1.00

Table 1: Temporal difference between sessions for the 33 subject Ajit et al. [1] dataset and the 41 subject Miller et al. [33, 34] dataset. In the header, V = HTC Vive, Q = Oculus Quest, C = HTC Vive Cosmos, A = data from Ajit et al., 1 and 2 designate the capture session.

the 16 common users across the Ajit et al. and Miller et al. dataset, 6 users provided data in April 2018 and 10 provided data in March 2019 for the Ajit et al. dataset. All 16 subjects provided data in October 2019 for the Miller et al. dataset. As a result, for 6 of the 16 users the average difference between the first Miller et al. and first Ajit et al. session is 552.33 ± 5.39 days and 219.90 ± 5.69 for the remaining 10 subjects.

4 Experiments and Results

We use the Siamese network architecture from Miller et al. [34] to conduct our studies on assessing temporal effects over short, medium, and long timescales on identification and authentication. Data from the input session is fed on one limb of the network, and data from the enrollment library is provided at the other limb. The data consists of the position and orientation features of the time trajectories from the three devices, i.e., the headset and hand controllers. The network is trained to compare the data from the two limbs and output a match score, with a high score indicating that the corresponding users are identical. Prior to feeding to the network, we normalize the trajectory positions to be mean-centered and scale normalized, thereby reducing the contribution of static physical parameters such as height from the ground, and enabling the study of dynamic behavior changes related to action performance.

We use the following network hyperparameters for all studies.

- We use a batch size of 128 to speed up training as opposed to Miller et al. who use a batch size of 64.
- We perform training optimization using a cyclic learning rate. Cyclic rates, where the learning rate varies between a minimum and maximum bound, have been shown to improve classification in standard image recognition tasks [48].

Since we use a single set of hyperparameters and do not perform hyperparameter tuning, we do not employ a validation set during training. Miller et al. demonstrate that using position and orientation features from all devices, i.e., the headset and the two hand controllers, generally provides highest accuracies and lowest EERs. We use the same set of features in our approach.

We provide results for within and cross-system identification and authentication for the short and medium timescale analyses using the dataset of Miller et al., and for the long timescales study using the 16 users in common with Miller et al. and Ajit et al. For the Miller et al. dataset, we evaluate same-system pairings where enrollment and input data come from the 2 sessions of a single system. We label the same-system pairings as Q1/Q2 for the first and second sessions of the Quest, V1/V2 for the first and second sessions of the Vive, and C1/C2 for the first and second sessions of the Cosmos. We evaluate cross-system pairings where enrollment comes from an earlier system and input from a later system. According to the dataset, Quest data was provided earlier for each user, followed by Vive, and Cosmos data was provided last. This results in cross-system pairings of Q1/V1, Q2/V2, Q2/V1, and Q2/V2 for Quest and Vive, V1/C1, V2/C2, V2/C1, and V2/C2 for Vive and Cosmos, and Q1/C1, Q2/C2, Q2/C1, and Q2/C2 for Quest and Cosmos.

We employ a leave-one-user-out approach for analyzing network performance. For each left out user, we build training pairs by linking throws from the enrollment system and session with throws from the input system and session across all training users. For the Siamese network, enrollment/input pairs that come from the same user are given a 0 distance label, while those that come from different users are given a label of 1 to represent a high distance. During testing, we use each trained Siamese network to produce distances from the test user's input throw to the enrollment throws of all users, i.e., training and test users. Using all users enables us to determine performance by matching against a comprehensive library of user data. We evaluate identification accuracy by determining if the smallest distance belongs to the correct user. We evaluate the authentication EER by obtaining false accept rates (FARs) and false reject rates (FRRs) upon varying the threshold against which the match distance is compared for accepting the user as genuine. We obtain EER as the FRR where FRR and FAR are identical. We average accuracies and EERs over all users involved in each analysis.

4.1 Analysis of Behavior over Short Timescales

We study the impact of temporal location of an enrollment throw on identification over a short timescale, by considering system pairings where users provide input data no more than one day after the enrollment data, i.e., no more than up to 24 hours. We perform our analysis on same-system pairings, i.e., Q1/Q2, V1/V2, and C1/C2. 38, 18, and 24 users provide enrollment and input data within a day for Quest, Vive, and Cosmos. We perform leave-one-user-out train and test within each subset of 38, 18, and 24 users. Given the small sample size, we use the non-parametric Friedman's test to determine if there is a significant difference in the temporal location of the enrollment match for each user's input throw. We obtain p-values of 0.3213, 0.5947, and 0.1106 for Quest, Vive, and Cosmos indicating no statistically significant effect of temporal location of enrollment throw for each input throw, indicating that user behavior variabilities over short timescales have limited influence on identification or authentication success. Since the Friedman's test did not indicate any significance, i.e. our p-values were not significant, we do not conduct any post-hoc tests for pairwise differences.

4.2 Analysis of Behavior over Medium Timescales

As part of our study of behavior over medium timescales, we analyze how identification and authentication is influenced by temporal difference on the order of days to weeks between enrollment and input data. Per user, we analyze the time difference between enrollment and input sessions for the system pairings under consideration, and we identify two clusters of users, one of whose time differences fall at or below a threshold t, and one above the threshold. We select t to be the value that enabled clusters across all system pairings to be balanced, i.e., where the maximum difference between cluster counts across all system pairings was no more than 20% of the total number of users. We find t to be 10 days for cross-system input and enrollment pairs. For within-system pairs, we analyze performance for the Vive for which we identify t to be 3 days. We do not report same-system results for the Quest and Cosmos, as we find that most users provide data within 2 days for these systems, and there is no threshold that ensures the cluster counts are balanced. Similarly, we find that the separation between the Quest and Vive sessions is small for most users, leading to no adequate threshold with balanced cluster counts, and therefore Quest/Vive pairings are excluded from the medium-scale analysis. We perform leave-one-user-out training within each cluster by leaving the input and enrollment data for each user and using input and enrollment pairs from remaining users in

Cluster	E/I	Q1/C1	Q1/C2	Q2/C1	Q2/C2	V1/C1	V1/C1	V2/C1	V2/C2	V1/V2
E/I Separation $\leq t$	#Users	21	16	21	16	25	19	25	25	22
	Accuracies	96.19	89.38	96.67	80.62	98.80	90.53	98.40	94.80	100.00
	EERs	0.95	2.79	0.57	3.21	0.38	3.86	1.12	1.64	0.05
E/I Separation $> t$	#Users	20	25	20	25	16	22	16	16	19
	Accuracies	97.00	100.00	93.50	99.20	91.25	98.18	98.75	99.38	100.00
	EERs	0.97	0.33	1.45	0.14	1.85	0.51	0.98	0.60	0.23

Table 2: Accuracies and EERs as percentages for various system pairs for identification and authentication in clusters of users with enrollment/input (E/I) separation of at or below t days and more than t days. The value of t is 3 days for V1/V2 and 10 days for the remaining pairings.

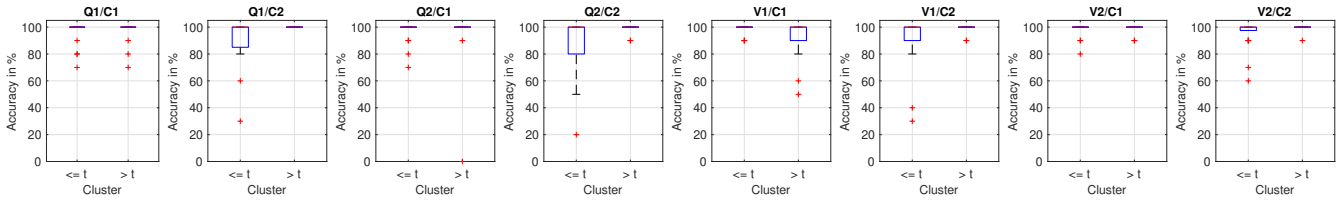

Figure 2: Boxplot of accuracies for cross-system within cluster training. Outliers influence the average accuracies listed in Table 2. We observe a median accuracy of 100% for clusters spaced less than t days and greater than t days, where t=10.

the cluster for training. We report test results by using the trained network to obtain distances between the left-out user's input throws and enrollment throws of all users in the left-out user's cluster, and computing success metrics. The analyses enable assessment of how authentication with higher enrollment/input temporal separation performs against authentication with lower temporal separation.

Table 2 provides average accuracies and EERs for all system pairings analyzed in this work for both clusters. We find that in most cases, average accuracies are high. We observe that for the lower-separation cluster, identification accuracy is higher and EER is lower than for the higher-separation cluster in the case of pairings Q2/C1 and V1/C1. For all other pairings, average success for the lower-separation cluster is lower, which appears anomalous considering that one may expect shorter timescales to introduce lesser behavior modification. However, the average scores are misleading as they are skewed by a few outlier users for whom the network malperforms. Figure 2 shows box-plots of per-user accuracies for each enrollment/input system pairing across the two clusters. As shown by the figure, the median accuracy is 100.00% for lower- and higher-separation pairings. While a spread is observed for Q1/C2, Q2/C2, and V1/C2, the lower quartile value is no lower than around 80%. The anomalous behavior observed in average performance is due to a few outlier users for whom the network malperforms for most system pairings.

One confounding factor for the higher-separation cluster is related to prior experience with the VR ball-throwing task. As discussed in the analysis of behavior over long timescales in Subsection 4.3, 16 users in the Miller et al. dataset provided data as part of the Ajit et al. study. 14 of these 16 users have enrollment/input temporal separations across system pairings that induce the users to be largely assigned to the higher-separation cluster. We find that the 14 users demonstrate an average accuracy of 96.88% over all cross-system pairings analyzed in Table 2. The remaining 2 of the 16 users assigned to the lower-separation cluster demonstrate an average accuracy of 97.50%. The average overall accuracy for the remaining 23 users in the lower-separation cluster is 95.22%. We use a Wilcoxon Signed Rank test to determine if the accuracy for users with prior knowledge were significantly different from those without. Our statistical test shows the differences are not significant with a p-value of 0.2716.

4.3 Analysis of Behavior over Long Timescales

We determine the extent to which temporal spacing on the order of months affects authentication performance by analyzing the 16 users in common between the datasets of Miller et al. and Ajit et al. For each user, there exist two Vive sessions in both datasets. The average temporal difference between the two sessions of Ajit et al., referred to as VA1 and VA2, is is 4.13 ± 2.80 days and and 2.38 ± 2.19 days between the Vive sessions of Miller et al., referred to as V1 and V2. The average difference between the first session of Ajit et al., i.e., VA1, and the first session of Miller et al., i.e., V1 is 344.56 ± 166.30 days.

Fine- and Coarse-grained Behavior Changes. Figure 3 shows device trajectories users 5, 3, 1, and 16 from the Miller et al. dataset who also provided data for Ajit et al. As the figure demonstrates, some users exhibit coarse change in behavior between the Ajit et al. and Miller et al. captures, while other users exhibit fine change. Users also exhibit varying degrees of change per device. Users 5 and 3 show fine-grained differences in the right controller trajectories across the two captures. User 5 shows a moderate difference in the left controller and headset motions. The headset motion of user 3 shows a somewhat moderate change as well, whereas their left controller demonstrates a high variation. With user 1, stark changes are observed for the left controller and a moderate change as a wider arc in the right controller trajectory. For user 16, we see variation in the throwing style with the right hand controller between Ajit et al. and Miller et al. Our study demonstrates that coarse and gradual variations impact the ability of matching networks to learn overall behavior patterns within and across users.

To assess the impact of temporal distance on identification and authentication, we use the leave-one-user-out approach to train Networks 1 through 4 using training data with varying levels of enrollment/input separation. With each network, we perform a variety of tests with summary results shown in Table 3. We conduct significance testing to evaluate the tests with respect to appropriate counterparts. Since our sample size is 16 subjects, we use the non-parametric Friedman's test to determine if the difference in accuracy for the networks is significant. We obtain a p-value $< 2.2 \times 10^{-16}$ indicating at least one set of network pairs are significantly different. Since the Friedman's test does not indicate which pairs are different, we conduct a post-hoc analysis using the Conover Test with Bonferroni correction for multiple comparisons. To measure effect size for

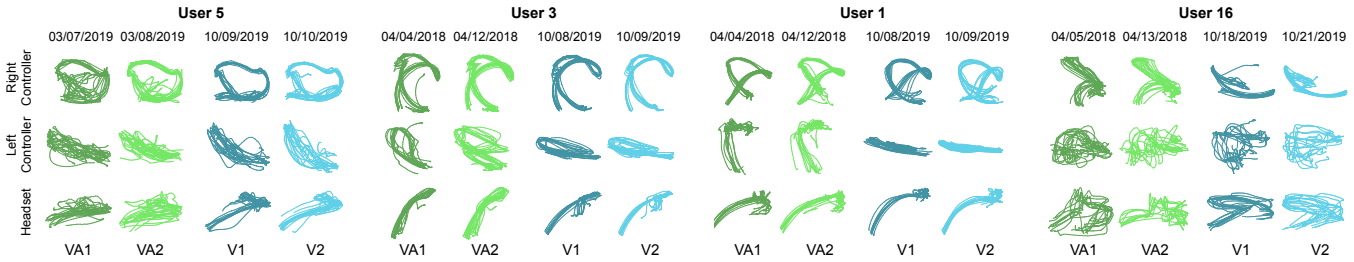

Figure 3: Controller and headset trajectories for VA1, VA2, V1, and V2 sessions for Users 5, 3, 1, and 16 from Ajit et al. and Miller et al. dataset.

statistically significant differences we use Cliff's Delta.

Network 1. We train Network 1 using VA1 as enrollment and VA2 as input for the training users. Network 1 is trained to recognize differences between users who provided enrollment and input data over a short period of temporal separation of on average 4.13±2.80 days. We conduct four tests to assess the effectiveness of Network 1 in successfully identifying and authenticating input data from the left out user by comparing the data to enrollment data from all 16 users in common with Ajit et al. and Miller et al.

- *Test 1.1* compares input data from VA2 against enrollment data from VA1, and assesses the accuracy of Network 1 at classifying data given at the same time period. Table 3 shows that as expected, we receive an identification accuracy of 98.75% with a single misclassified input trajectory for 2 users and minimum authentication EER of 1.12%.
- *Test 1.2 and Test 1.3* compare input data from V1 and V2 respectively against enrollment data from VA1. V1 and V2 are separated by 7 to 18 months from VA1. As shown in Table 3, our identification accuracy and authentication EER rate show a high drop in performance since V1 and V2 were given by users much later in time. Most users exhibit divergence in behavior that cannot be captured by a network trained on users who provide input data close to the enrollment data.
- *Test 1.4* compares input data from V2 against enrollment data from V1. While the data from V1 and V2 comes from a different time period than VA1 and VA2, the separation between V1 and V2 at 2.38±2.19 days is comparable to that between VA1 and VA2. While the accuracy 78.75% is higher than for trajectories that are separated further out, the accuracy is considerably lower than for data provided as part of the Ajit et al. dataset, and similarly EER is considerably higher. As shown by the confusion matrix for Test 1.4 in Figure 4, the most misclassifications are observed for users 12, 16, and 36. While the behavior over short timescales for all these users is consistent suggesting that accuracy should be high similar to Test 1.1, Test 1.4 uses enrollment data for multiple users with altered behavior patterns. The weights of Network 1 are tuned to the behavior patterns of the users from VA1. When altered behavior is provided on the enrollment limb of Network 1, features generated by applying the network weights may be incapable of representing the altered behavior, introducing a drop in identification accuracy. It should be noted that while for some users, e.g., users 5, 11, and 12 intra-capture separation is on the order of 7 months, while for other users, e.g., users 1, 3, and 16 intra-capture separation is nearly 18 months, large differences in inter-user capture separation do not appear to contribute to change in identification or authentication success.

We observe that networks trained solely with short temporal separation in behavior data are susceptible to being overly tuned to fine-scale differences between users, leading to low success when applied to assessment of identification or authentication when behavior data is separated by long timescales as shown by Test 1.2 and Test 1.3. They are also susceptible to being finely tuned to systematic high-level similarities in user behavior over short timescales. When multiple users exhibit non-systematic behavior change with novel enrollment data provided for all users as used in Test 1.4, identification and authentication for input behavior data provided over short timescales with respect to that enrollment data is hampered. Our post-hoc test reveals that the differences between Test 1.1 and all other tests for Network 1 are significant with p-value $< 2 \times 10^{-16}$, $< 2 \times 10^{-16}$, and 1.2×10^{-5} and effect size 0.859375, 0.859375, and 0.6328125 for Test 1.2, Test 1.3, and Test 1.4 respectively.

Network 2. We train Network 2 as a baseline network to evaluate accuracy of comparing input data from V2 against enrollment data from V1 using Test 2.1 when the same pairs V1/V2 are used for training. Network 2 provides a high accuracy of 100.00% and low EER of 0.06%. As expected, the difference between Test 1.1 and Test 2.1 is not significant.

Network 3. We train Network 3 using enrollment data from VA1 and input data from V1. Network 3 is trained to recognize similarities between users when input and enrollment data is separated by longer periods, i.e., between 7 to 18 months. Within this period, some users' devices exhibit coarse motion changes, while others exhibit fine-grained changes, enabling the network to acquire enrollment/input pairs for learning coarse and fine differences. We conduct three tests by using Network 3.

- *Test 3.1* uses Network 3 to compare input data from V1 to enrollment data from VA1. The accuracy is low at 84.38% with EER of 9.48%. However, the accuracy is higher than when closely spaced data is used in training Network 1 to obtain results for Test 1.2 that also compares V1 input to VA1 enrollment. We observe a reduced identification and authentication success for user 16, who confounds with user 12. The V1 right-hand throw trajectories for user 16 shown in Figure 3 resemble the underhand throw of user 12 shown in Figure 1. When comparing Test 3.1 to Test 1.2 in Network 3 the difference is significant with a p-value of 1.9×10^{-9} and effect size 0.6601562.
- *Test 3.2* compares input data from V2 to enrollment data from VA1. The separation between V2 and VA1 is similar to that between V1 and VA1, i.e., around 7 to 18 months. Similar to Test 3.1, inclusion of long temporal separation improves accuracy over Test 1.3 that also compares V2 input against VA1 enrollment, however, at 77.50% with EER of 13.02% the improvement is lesser than with Test 3.1. With Test 3.1, we observe reduced success for users 12 and 1 in addition to user 16, unlike in Test 3.1 where the identification and authentication for users 12 and 1 shows high success with the cyclic rate. While for users 12 and 1, V1 and V2 trajectories are more consistent than their V1 and VA1 trajectories, slight variations are observable in the shape of the right controller trajectory in V2 which may induce the behavior to match closer with an incorrect user who shows similar behavior. In the case of user 1, confounding largely happens with user 15, both of whom have similar overall right controller trajectory shape and head motion. The difference between Test 3.2 and Test 1.3 is significant with a p-value of 1.6×10^{-6} and effect size 0.5859375.

Network E/I	**Network 1 VA1/VA2**				**Network 2 V1/V2**	**Network 3 VA1/V1**			**Network 4 VA1/VA2, VA1/V1 VA2/V1**		
Test (E/I)	Test 1.1 (VA1/VA2)	Test 1.2 (VA1/V1)	Test 1.3 (VA1/V2)	Test 1.4 (V1/V2)	Test 2.1 (V1/V2)	Test 3.1 (VA1/V1)	Test 3.2 (VA1/V2)	Test 3.3 (V1/V2)	Test 4.1 (VA1/V1)	Test 4.2 (VA1/V2)	Test 4.3 (V1/V2)
Acc.	98.75	31.25	32.50	78.75	100.00	84.38	77.50	97.50	90.62	90.62	100.00
EER	1.12	29.25	30.10	17.60	0.06	9.48	13.02	1.31	2.85	2.38	0.29

Table 3: Accuracy (Acc.) and EER as percentages for biometrics using behavior separated over long timescales with the 16 subjects common to the data sets of Ajit et al. [1] and Miller et al. [33, 34]. VA* = HTC Vive data from the Ajit et al. [1] collection. V* = HTC Vive data from the Miller et al. [33, 34] collection, '1' and '2' refer to the session number within each collection, with session 2 occurring later in time than session 1. Temporal separations are 4.13±2.80 days for VA1/VA2, 2.38±2.19 for V1/V2, 344.56±166.30 for VA1/V1, and 346.94±165.52 for VA1/V2. The separation range for V1 and V2 from VA1 is 7 to 18 months.

As expected, the difference between Test 3.1 and Test 3.2 is not significant as the temporal difference between the V1 and V2 sessions are small and our short and medium timescale studies did not reveal any significant finding.

- *Test 3.3* compares input data from V2 to enrollment data from V1. Test 3.3 assesses the effectiveness of Network 3, trained using data with long temporal separation, in matching data provided with shorter temporal separation of an average of 2.38±2.19 days from the enrollment. We find that the results are improved over the similar Test 1.4, with accuracy of 97.50% and EER of 1.31%. When comparing Test 1.4 and Test 3.3, we find a significant difference with p-value of 5.0×10^{-6} and effect size 0.5859375.

Tests for Network 3 show that by presenting examples of coarse and fine variations, accuracy is improved for behavior data provided over short timescales and for behavior data provided over longer timescales when the pairs are representative of the overall pattern of behavior change between users over long timescales.

Network 4. To boost the quantity of training data for the network to learn separability, we train Network 4 by providing pairs from VA1/VA2, VA1/V1, and VA2/V1. The pairs represent behavior variations over short and long timescales. We repeat the tests Test 3.1 to Test 3.3 for Network 3 as Test 4.1 to Test 4.3.

- *Test 4.1* uses Network 4 to compare input data from V1 to enrollment data from VA1. Test 4.1 demonstrates higher accuracy of 90.62% than Test 3.1 and lower EER of 2.85%. Users 12 and 16 are the only ones with high misclassification. While higher accuracy than Test 3.2 seems surprising, Network 4 is boosted with more examples of long-range behavior variation through the provision of pairs from VA2 and V1.
- *Test 4.2* compares input data from V2 to enrollment data from VA1. Test 4.2 demonstrates similar patterns as Test 4.1, with a higher maximum accuracy of 90.62% with EER of 2.38%. User 16 is the only user with high misclassification.
- *Test 4.3* compares input data from V2 to enrollment data from V1. Test 4.3 demonstrates a 100% accuracy and EER of 0.29%.

Overall, we see higher accuracies for behavior data separated by long and short timescales by augmenting the network with trajectory set VA2 that show short separation from VA1 and long separation from V1, i.e., that include coarse and fine motion differences between enrollment/input pairs. When comparing Test 4.1 to Test 3.1, Test 4.2 to Test 3.2, and Test 4.3 to Test 3.3, we find no significant difference. However, given that the sample size is small, in general, we recommend augmenting the long timescale training samples with some examples of enrollment/input pairs that are closer in time.

5 Discussion

In this paper, we perform the first investigation of temporal changes in VR behavior over short, medium, and long timescales, and their impact on the success of mechanisms that use VR behavior as a biometric for identification and authentication. Our findings suggest that short timescales have minimal impact in altering VR behavior for common repeatable tasks such as ball-throwing, where users may tap into prior real-world experience to rapidly acquire task familiarity. Over medium timescales, the data suggests minimal impact as well. Studies of two datasets collected over a time period separated by 7 to 18 months demonstrate that over long timescales, users demonstrate varying levels of behavior evolution, with differing changes observed on a per-device level. Behavior evolution in long timescales appear to negatively impact the success of security mechanisms that are trained to recognize behavior differences over short timescales in protecting using input separated from enrollment data by longer timescales. While security may be increased by adjusting the threshold to reduce the FAR, the higher EER demonstrates that FAR reduction will come at the risk of increased FRR, reducing usability. Instead, we find that using data with longer time separations may capture multiple granularities of behavior evolution, enabling security mechanisms trained with longer separations of months to years to work on enrollment data distanced by intermediate levels of separation, such as days to weeks. The reduced EER obtained on capturing behavior over multiple timescales indicates that usability may also be improved.

Given the importance of the temporal factor in shaping user behavior in VR environments as elucidated by our work, it is critical to design behavior-based security mechanisms that are cognizant of behavior evolution, especially over long timescales. Future work should investigate types of tasks and portions of task performance that positively or negatively impact identification and authentication in the presence of behavior evolution, as well as impact on continuous usage and security. The interplay of biomechanics of movement in VR, cognitive understanding and retention of the VR environment, and muscle memory for task completion is complex, and deserves attention. Prior higher-level knowledge of the VR environment may reduce variability over successive short timescales, however, more work needs to be performed to clearly elucidate the role of prior knowledge, both overall and task-related, and behavior adaptation on multiple timescales. Additionally, the role of prior real-world experience in influencing VR behavior and its contribution toward security success must also be investigated. The task investigated in this work is a simple ball-throwing action. Future work in VR should explore complex actions where users may need to simultaneously perform physical and cognitive tasks, such as navigating an unknown territory, where evolution of cognitive capabilities symbiotically influences the physical behavior such as, e.g., a user learning that a specific pathway requires the least amount of physical effort and employing the optimum movement in a future trial.

An important aspect to consider for behavior over short timescales is the impact of transient events such as injuries. If a user suffers an injury immediately after they provide enrollment data, behavior alteration due to injury may negatively impact identification or authentication, and may cause an intended user to be locked out of the application. Future work should investigate the role played by transient changes in impacting security, for instance, by isolating body parts on which injury occurs, and investigating the contribution

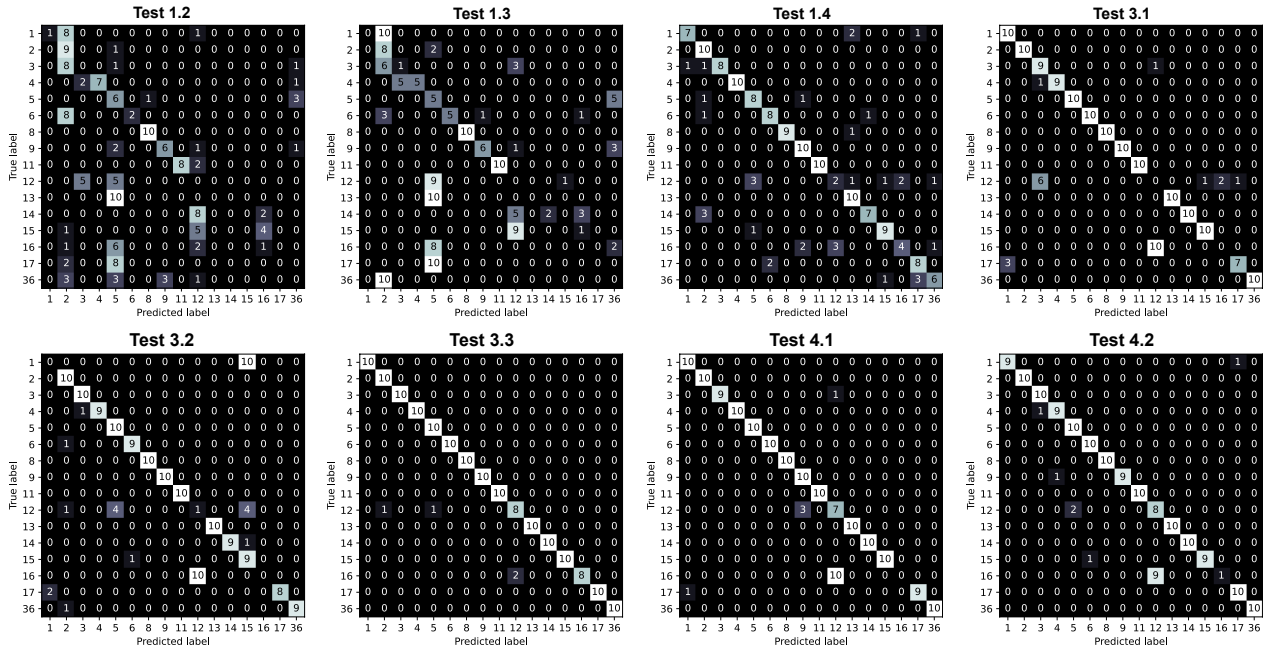

Figure 4: Confusion matrices for 16 users in common with Ajit et al. and Miller et al. tested using 8 of the 11 tests discussed in Table 3.

of that body part and its degrees of freedom toward authentication. For instance, security mechanisms may use a voting scheme to eliminate devices handled by injured body parts from contributing toward identification or authentication. Another aspect to investigate is how user behavior in VR is influenced over short timescales by events that may occur in the real world, e.g., if a user is distracted from their VR task by a phone call and returns to the task after attending the call. Future work should investigate how VR behavior is influenced by the cognitive load in handling real-world transient experiences with varying degrees of positive or negative impact on the person at the moment of distraction. To track impact of real-world events on cognitive load over short timescales, future work may use techniques such as eye-tracking [42] to assess pupil dilation or wearables for heart-rate monitoring [16]. Emotional events may be simulated by having subjects watch positive or negative video, and increased cognitive loads may be simulated by having subjects perform puzzle-solving or form-filling.

Unlike keystroke or gesture-based behavioral biometrics where data can be captured at scale using a web-based key logger or the user's personal device, the relative novelty of VR makes large-scale collection challenging. Until VR devices become as ubiquitous as smartphones and laptops, longitudinal data collection for VR spanning days, months, and years will require users to visit a capture site which can be infeasible as a large pool of subjects are less likely to visit for multiple captures. The largest capture to date for VR biometrics is the work of Miller et al. [31] with 511 subjects captured in a technology museum and a university setting, however the data is captured in a single session making analysis over medium and long timescales impossible. The work of Miller et al. [33, 34] has 41 subjects providing data across multiple sessions spanning days and weeks. However, once subjects are binned into VR experience, real-world task experience, and temporal difference between session the number of subjects in each bin is not sufficient to draw statistically significant conclusions. The 16 common users across Ajit et al. [1] and Miller et al. [33, 34] enable long-timescale analysis, however, a larger sample size is needed for concrete conclusions on temporal effects over long timescales. Existing VR biometrics datasets lack longitudinal data on cognitive or physical changes induced by a subject aging, growing in height, or gaining or losing weight. A large-scale dataset encompassing physical and cognitive changes at varying timescales will help comprehensive analysis of impact of temporal changes in VR biometrics.

Recognizing the challenge of collecting longitudinal data at scale, we advocate incentivizing research teams to work with local middle schools, high schools, and colleges to perform data collection over 4-10 year spans where within- and cross-subject physical and cognitive changes may be studied in real-world and VR environments. Efforts should include reaching average users by incorporating task-based data collection deployed in consumer VR applications such as games or puzzle solving that may be attempted by multiple users. Given the potential of VR to facilitate physical therapy and engagement, incentivization should also be provided to reach out to senior citizen communities so as to provide applications that involve older adults in VR interaction that is secure and requires minimal learning time. Support is critical from the VR community for elucidating ethical concerns, facilitating seamless yet informed collection from average consumers, and providing mechanisms for secure storage of data. Successful at-scale collection has the potential to transform VR security, as it enables performing behavior-based authentication by leveraging the benefits of VR in performing controlled mimicking of the real world. Modern smartphones provide multiple front facing cameras and higher-end processors which may be leveraged to create Cardboard style VR systems that perform camera-based hand tracking similar to the Oculus Quest and HTC Vive Cosmos. Such systems may allow VR to become more accessible leading to data collection at scale. Behavior biometrics in VR are likely to remain the de facto approach of authentication until VR system manufacturers embed fingerprint scanners in controllers or iris cameras in headsets with high-assurance behavior-independent continuous authentication. By providing high-success-rate behavioral biometrics, VR applications have the future potential to bypass traditional credentials and provide uninterrupted user experiences in secure VR environments.

REFERENCES

[1] A. Ajit, N. K. Banerjee, and S. Banerjee. Combining pairwise feature matches from device trajectories for biometric authentication in virtual reality environments. In *Proc. AIVR*. IEEE, New York, USA, 2019.

[2] F. A. Alsulaiman and A. El Saddik. A novel 3d graphical password schema. In *Proc. VECIMS*. IEEE, New York, USA, 2006.

[3] F. A. Alsulaiman and A. El Saddik. Three-dimensional password for more secure authentication. *IEEE Transactions on Instrumentation and Measurement*, 57(9):1929–1938, Sep 2008.

[4] F. Battistone and A. Petrosino. Tglstm: A time based graph deep learning approach to gait recognition. *Pattern Recognition Letters*, 126(9):132–138, Sep 2019.

[5] M.-S. Bracq, E. Michinov, and P. Jannin. Virtual reality simulation in nontechnical skills training for healthcare professionals: A systematic review. *Simulation in Healthcare*, 14(3):188–194, Jun 2019.

[6] A. G. Campbell, T. Holz, J. Cosgrove, M. Harlick, and T. O'Sullivan. Uses of virtual reality for communication in financial services: A case study on comparing different telepresence interfaces: Virtual reality compared to video conferencing. In *Proc. FCC*. Springer, Berlin, Germany, 2019.

[7] B. J. Concannon, S. Esmail, and M. Roduta Roberts. Head-mounted display virtual reality in post-secondary education and skill training: A systematic review. In *Proc. FIE*. Frontiers, Switzerland, 2019.

[8] E. Czerniak, A. Caspi, M. Litvin, R. Amiaz, Y. Bahat, H. Baransi, H. Sharon, S. Noy, and M. Plotnik. A novel treatment of fear of flying using a large virtual reality system. *Aerospace medicine and human performance*, 87(4):411–416, Apr 2016.

[9] H. Feng, C. Li, J. Liu, L. Wang, J. Ma, G. Li, L. Gan, X. Shang, and Z. Wu. Virtual reality rehabilitation versus conventional physical therapy for improving balance and gait in parkinson's disease patients: A randomized controlled trial. *Medical science monitor: international medical journal of experimental and clinical research*, 25(5):4186–4192, Jun 2019.

[10] M. Funk, K. Marky, I. Mizutani, M. Kritzler, S. Mayer, and F. Michahelles. Lookunlock: Using spatial-targets for user-authentication on hmds. In *Proc. CHI Extended Abstracts*. ACM, New York, USA, 2019.

[11] C. George, D. Buschek, A. Ngao, and M. Khamis. Gazeroomlock: Using gaze and head-pose to improve the usability and observation resistance of 3d passwords in virtual reality. In *Proc. AVR*. Springer, Berlin, Germany, 2020.

[12] 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 *Proc. VR*. IEEE, New York, USA, 2019.

[13] 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 *Proc. NDSS*, 2017.

[14] J. Gurary, Y. Zhu, and H. Fu. Leveraging 3d benefits for authentication. *International Journal of Communications, Network and System Sciences*, 10(08):324–338, Aug 2017.

[15] L. Jensen and F. Konradsen. A review of the use of virtual reality head-mounted displays in education and training. *Education and Information Technologies*, 23(4):1515–1529, Jul 2018.

[16] P. Jerčić, C. Sennersten, and C. Lindley. Modeling cognitive load and physiological arousal through pupil diameter and heart rate. *Multimedia Tools and Applications*, 79(5):3145–3159, Jan 2020.

[17] A. Kupin, B. Moeller, Y. Jiang, N. K. Banerjee, and S. Banerjee. Task-Driven Biometric Authentication of Users in Virtual Reality (VR) Environments. In *Proc. MMM*. Springer, Berlin, Germany, 2019.

[18] B. M. Kyaw, N. Saxena, P. Posadzki, J. Vseteckova, C. K. Nikolaou, P. P. George, U. Divakar, I. Masiello, A. A. Kononowicz, N. Zary, et al. Virtual reality for health professions education: systematic review and meta-analysis by the digital health education collaboration. *Journal of medical Internet research*, 21(1), Jan 2019.

[19] M. Lager and E. A. Topp. Remote supervision of an autonomous surface vehicle using virtual reality. *IFAC-PapersOnLine*, 52(8):387–392, Jul 2019.

[20] S. Li, A. Ashok, Y. Zhang, C. Xu, J. Lindqvist, and M. Gruteser. Whose move is it anyway? authenticating smart wearable devices using unique head movement patterns. In *Proc. PerCom*. IEEE, New York, USA, 2016.

[21] X. Li, Y. Makihara, C. Xu, Y. Yagi, and M. Ren. Joint intensity transformer network for gait recognition robust against clothing and carrying status. *IEEE Transactions on Information Forensics and Security*, 14(12):3102–3115, Dec 2019.

[22] J. Liebers, M. Abdelaziz, L. Mecke, A. Saad, J. Auda, U. Grünefeld, F. Alt, and S. Schneegass. Understanding user identification in virtual reality through behavioral biometrics and the effect of body normalization. In *Proc. CHI*. ACM, New York, USA, 2021.

[23] J. Liebers and S. Schneegass. Gaze-based authentication in virtual reality. In *Proc. ETRA*. ACM, New York, USA, 2020.

[24] K. R. Lohse, C. G. Hilderman, K. L. Cheung, S. Tatla, and H. M. Van der Loos. Virtual reality therapy for adults post-stroke: a systematic review and meta-analysis exploring virtual environments and commercial games in therapy. *PloS one*, 9(3):e93318, Mar 2014.

[25] M. Maciaś, A. Dabrowski, J. Fraś, M. Karczewski, S. Puchalski, S. Tabaka, and P. Jaroszek. Measuring performance in robotic teleoperation tasks with virtual reality headgear. In *Proc. AUTOMATION*. Springer, Berlin, Germany, 2019.

[26] M. G. Maggio, G. Maresca, R. De Luca, M. C. Stagnitti, B. Porcari, M. C. Ferrera, F. Galletti, C. Casella, A. Manuli, and R. S. Calabrò. The growing use of virtual reality in cognitive rehabilitation: fact, fake or vision? a scoping review. *Journal of the National Medical Association*, 111(4):457–463, Aug 2019.

[27] F. Mathis, H. I. Fawaz, and M. Khamis. Knowledge-driven biometric authentication in virtual reality. In *Proc. CHI Extended Abstracts*. ACM, New York, USA, 2020.

[28] F. Mathis, J. Williamson, K. Vaniea, and M. Khamis. Rubikauth: Fast and secure authentication in virtual reality. In *Proc. CHI Extended Abstracts*. ACM, New York, USA, 2020.

[29] F. Mathis, J. H. Williamson, K. Vaniea, and M. Khamis. Fast and secure authentication in virtual reality using coordinated 3d manipulation and pointing. *ACM Transactions on Computer-Human Interaction*, 28(1):1–44, Jan 2021.

[30] D. S. Matovski, M. S. Nixon, S. Mahmoodi, and J. N. Carter. The effect of time on gait recognition performance. *Transactions on Information Forensics and Security*, 7(2):543–552, Apr 2011.

[31] M. R. Miller, F. Herrera, H. Jun, J. A. Landay, and J. N. Bailenson. Personal identifiability of user tracking data during observation of 360-degree vr video. *Scientific Reports*, 10(1):1–10, Oct 2020.

[32] R. Miller, A. Ajit, N. K. Banerjee, and S. Banerjee. Realtime behavior-based continual authentication of users in virtual reality environments. In *AIVR*. IEEE, New York, USA, 2019.

[33] R. Miller, N. K. Banerjee, and S. Banerjee. Within-system and cross-system behavior-based biometric authentication in virtual reality. In *Proc. VRW*. IEEE, New York, USA, 2020.

[34] R. Miller, N. K. Banerjee, and S. Banerjee. Using siamese neural networks to perform cross-system behavioral authentication in virtual reality. In *Proc. VR*. IEEE, New York, USA, 2021.

[35] T. Mustafa, R. Matovu, A. Serwadda, and N. Muirhead. Unsure how to authenticate on your vr headset? come on, use your head! In *Proc. IWSPA*. ACM, New York, USA, 2018.

[36] S. Neumeier, N. Gay, C. Dannheim, and C. Facchi. On the way to autonomous vehicles teleoperated driving. In *Proc. AmE*. VDE, 2018.

[37] S. Neumeier, P. Wintersberger, A.-K. Frison, A. Becher, C. Facchi, and A. Riener. Teleoperation: The holy grail to solve problems of automated driving? sure, but latency matters. In *Proc. AutomotiveUI*. ACM, New York, USA, 2019.

[38] M. M. North, S. M. North, and J. R. Coble. Virtual reality therapy: an effective treatment for the fear of public speaking. *International Journal of Virtual Reality*, 3(3):1–6, Jan 1998.

[39] I. Olade, C. Fleming, and H.-N. Liang. Biomove: Biometric user identification from human kinesiological movements for virtual reality systems. *Sensors*, 20(10):2944, May 2020.

[40] I. Olade, H.-N. Liang, C. Fleming, and C. Champion. Exploring the vulnerabilities and advantages of swipe or pattern authentication in virtual reality (vr). In *Proc. ICVARS*. ACM, New York, USA, 2020.

[41] N. Palaskar, Z. Syed, S. Banerjee, and C. Tang. Empirical techniques to detect and mitigate the effects of irrevocably evolving user profiles

in touch-based authentication systems. In *Proc. HASE*. IEEE, New York, USA, 2016.

[42] O. Palinko, A. L. Kun, A. Shyrokov, and P. Heeman. Estimating cognitive load using remote eye tracking in a driving simulator. In *Proc. ETRA*. ACM, New York, USA, 2010.

[43] K. Pfeuffer, M. J. Geiger, S. Prange, L. Mecke, D. Buschek, and F. Alt. Behavioural biometrics in vr: Identifying people from body motion and relations in virtual reality. In *Proc. CHI*. ACM, New York, USA, 2019.

[44] G. Pizzi, D. Scarpi, M. Pichierri, and V. Vannucci. Virtual reality, real reactions?: Comparing consumers' perceptions and shopping orientation across physical and virtual-reality retail stores. *Computers in Human Behavior*, 96(0):1–12, Jul 2019.

[45] J. Radianti, T. A. Majchrzak, J. Fromm, and I. Wohlgenannt. A systematic review of immersive virtual reality applications for higher education: Design elements, lessons learned, and research agenda. *Computers & Education*, 147(0), Apr 2020.

[46] C. E. Rogers, A. W. Witt, A. D. Solomon, and K. K. Venkatasubramanian. An approach for user identification for head-mounted displays. In *Proc. ISWC*. ACM, New York, USA, 2015.

[47] X. Shen, Z. J. Chong, S. Pendleton, G. M. J. Fu, B. Qin, E. Frazzoli, and M. H. Ang. Teleoperation of on-road vehicles via immersive telepresence using off-the-shelf components. In *Intelligent Autonomous Systems*, pp. 1419–1433. Springer, Berlin, Germany, 2016.

[48] L. N. Smith. Cyclical learning rates for training neural networks. In *Proc. WACV*. IEEE, New York, USA, 2017.

[49] A. J. Snoswell and C. L. Snoswell. Immersive virtual reality in health care: Systematic review of technology and disease states. *JMIR Biomedical Engineering*, 4(1), Jan-Dec 2019.

[50] Z. Syed, S. Banerjee, Q. Cheng, and B. Cukic. Effects of user habituation in keystroke dynamics on password security policy. In *Proc. HASE*. IEEE, New York, USA, 2011.

[51] Z. Syed, S. Banerjee, and B. Cukic. Normalizing variations in feature vector structure in keystroke dynamics authentication systems. *Software Quality Journal*, 24(1):137–157, Mar 2016.

[52] Z. Syed, J. Helmick, S. Banerjee, and B. Cukic. Touch gesture-based authentication on mobile devices: The effects of user posture, device size, configuration, and inter-session variability. *Journal of Systems and Software*, 149(3):158–173, Mar 2019.

[53] R. Tilhou, V. Taylor, and H. Crompton. 3d virtual reality in k-12 education: A thematic systematic review. In *Emerging Technologies and Pedagogies in the Curriculum*, pp. 169–184. Springer, Berlin, Germany, 2020.

[54] P. Wang, P. Wu, J. Wang, H.-L. Chi, and X. Wang. A critical review of the use of virtual reality in construction engineering education and training. *International journal of environmental research and public health*, 15(6):1204, Jun 2018.

[55] S. Weise and A. Mshar. Virtual reality and the banking experience. *Journal of Digital Banking*, 1(2):146–152, Sep 2016.

[56] L. Xue, C. J. Parker, and H. McCormick. A virtual reality and retailing literature review: Current focus, underlying themes and future directions. In *Augmented Reality and Virtual Reality*, pp. 27–41. Springer, Berlin, Germany, 2019.

[57] S. Yi, Z. Qin, E. Novak, Y. Yin, and Q. Li. Glassgesture: Exploring head gesture interface of smart glasses. In *Proc. INFOCOM*. IEEE, New York, USA, 2016.

[58] S. Yu, D. Tan, and T. Tan. A framework for evaluating the effect of view angle, clothing and carrying condition on gait recognition. In *Proc. ICPR*. IEEE, New York, USA, 2006.

[59] Z. Yu, H.-N. Liang, C. Fleming, and K. L. Man. An exploration of usable authentication mechanisms for virtual reality systems. In *Proc. APCCAS*. IEEE, New York, USA, 2016.