Exploring the Stability of Behavioral Biometrics in Virtual Reality in a Remote Field Study

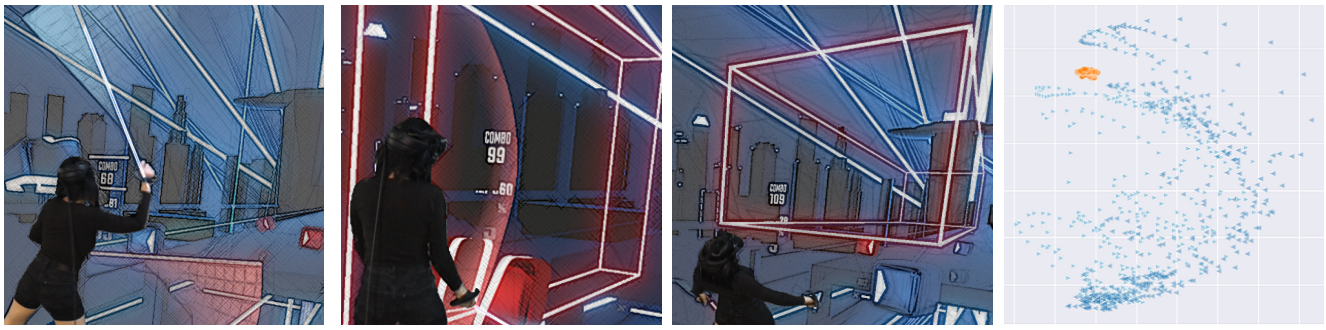
Towards Implicit and Continuous User Identification through Body Movements

Jonathan Liebers
University of Duisburg-Essen
Essen, Germany
jonathan.liebers@uni-due.de

Christian Burschik
University of Duisburg-Essen
Essen, Germany
christian.burschik@stud.uni-due.de

Uwe Gruenefeld
University of Duisburg-Essen
Essen, Germany
uwe.gruenefeld@uni-due.de

Stefan Schneegass
University of Duisburg-Essen
Essen, Germany
stefan.schneegass@uni-due.de



(a) Slashing blocks. (b) Sidestepping cuboids. (c) Ducking from cuboids. (d) Elicited spatiotemporal data.

Figure 1: We conduct a field study on behavioral biometrics in VR. Participants played *Beat Saber* where they sliced colored blocks with light sabers attached to their controllers (a), had to side-step cuboids (b), and were required to duck during the game (c). We elicit their spatiotemporal motion data and explore continuous user identification over a time span of eight weeks (d).

ABSTRACT

Behavioral biometrics has recently become a viable alternative method for user identification in Virtual Reality (VR). Its ability to identify users based solely on their implicit interaction allows for high usability and removes the burden commonly associated with security mechanisms. However, little is known about the temporal stability of behavior (i.e., how behavior changes over time), as most previous works were evaluated in highly controlled lab environments over short periods. In this work, we present findings obtained from a remote field study (N = 15) that elicited data over a period of eight weeks from a popular VR game. We found that there are changes in people's behavior over time, but that two-session identification still is possible with a mean F1-score of up to 71%, while an initial training yields 86%. However, we also see that performance can drop by up to over 50 percentage points when testing with later sessions, compared to the first session, particularly for smaller groups. Thus, our findings indicate that the use of behavioral biometrics in VR is convenient for the user and practical with regard to changing behavior and also reliable regarding behavioral variation.

CCS CONCEPTS

• **Human-centered computing** → *Field studies*; • **Security and privacy** → *Usability in security and privacy*; *Biometrics.*

KEYWORDS

Implicit User Identification; Virtual Reality; Field Study, Continuous Identification.

ACM Reference Format:
Jonathan Liebers, Christian Burschik, Uwe Gruenefeld, and Stefan Schneegass. 2023. Exploring the Stability of Behavioral Biometrics in Virtual Reality in a Remote Field Study: Towards Implicit and Continuous User Identification through Body Movements. In *29th ACM Symposium on Virtual Reality Software and Technology (VRST 2023), October 9–11, 2023, Christchurch, New Zealand.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3611659.3615696

VRST 2023, October 9–11, 2023, Christchurch, New Zealand

ACM ISBN 979-8-4007-0328-7/23/10…$15.00
https://doi.org/10.1145/3611659.3615696

1 INTRODUCTION

In recent years, the number of Virtual Reality (VR) users has been rising continuously. Today, VR empowers users to experience a variety of applications, ranging from productive use cases (e.g., meeting coworkers in a virtual space [9]) to entertaining ones (e.g., playing immersive games [8]). In many applications, identifying the user can be beneficial since it enables personalized experiences and assures the protection of personal user data. Today's user identification is mostly explicitly designed, which means that one's primary action (e.g., playing a game) is interrupted to allow identity verification by, say, interacting with a password window. Behavioral biometrics is a promising approach that tackles this problem. It utilizes distinct patterns in human behavior to identify users and distinguish between them, potentially enabling implicit and continuous user identification [2, 51].

Researchers have proposed various means for behavioral biometrics in VR [21, 28, 37, 42, 50]. In most cases, researchers conduct user studies in their labs to explore behavioral biometrics, which are often conducted only within a single session (i.e., a single day of each participant participating in the study) [29, 49] or, at maximum, taking place across two sessions [27, 36, 42]. However, human behavior tends to change over time. For example, gait can be impacted by long-term time effects, such as aging [11], short and medium-term time effects, such as shifts in mood [7] or context [14]. Thus, the extent to which this change impacts identification systems remains unclear. The change in behavior over time is specific to behavioral biometrics and is called "stability" or "permanence" [13]. While such changes are present in physiological biometrics, too (e.g., the changes of a face due to aging), it is yet not clearly understood how this affects human behavior (e.g., during VR activities).

Therefore, this work investigates the mid-term stability of behavioral biometrics for user identification in VR. We study user behavior over eight weeks in a field experiment, where we distributed *Meta Quest 2* headsets to 16 participants. We asked them to play *Beat Saber*[1], a popular VR game, at least twice a week. In *Beat Saber*, users must slice floating cubes to the rhythm of a song. We elicited the users' spatiotemporal motion data and used it for implicit identification and to study how time affects the identification rates. We selected this game because it elicits various (upper) body movements similar to those previously studied in other VR games [23, 27].

The findings of this work help in creating systems that identify users by their actions in *Beat Saber*. Thereby, overarching systems can establish trust in a user's identity after they played the game; thus, the need for re-identification at a later point might vanish.

Contribution Statement. The contribution of our work is threefold. First, we introduce a method of conducting remote field studies in VR, enabling researchers to collect data outside controlled laboratory settings. Second, we report insights into the stability of behavior from a remote field study that lasted over eight weeks and analyze its effects on user identification. At last, we publish our dataset to enable replicability in addition to the source code of our background application.

2 RELATED WORK

Our work takes place at the crossroads of implicit identification and behavioral biometrics in VR. Combining both allows for seamless user identification, where users are relieved of having to remember dozens of complex passwords [2]. In addition, they also save considerable time by not having to explicitly enter them anymore [2]. Passwords, the current most prevalent form of determining the identity of a user through a knowledge-based component, are associated with a number of problems. First, with the increasing number of passwords [24], users are increasingly overwhelmed. This leads to passwords being frequently reused [45], being guessable [15] and predictable [38]. Furthermore, the failure rate of entering passwords is associated with approximately 10% [5, 47], hence they are considered to be imperfect [4]. At last, the requirement of users having to memorize passwords leads to a phenomenon known as "the great authentication fatigue" [2, 47].

However, alternatives to passwords do exist. One is the employment of security tokens that users need to carry with them and which can be lost [39], which are again subject to usability issues [22]. The other is the employment of biometrics which are, in comparison to tokens, always with the user and which, in comparison to passwords, cannot be forgotten [19]. Biometrics are often distinguished into physiological biometrics, using primarily physiological attributes such as finger prints [16] or finger veins [10] recognition and behavioral biometrics, using behavioral features such as gait, eye gaze [12, 21, 29] or full body movement [40, 42]. However, physiological biometrics in many cases need to be provided explicitly to the device's sensor (e.g., the index finger needs to be actively moved to the sensor), hence they still interrupt users in many cases.

2.1 Implicit Identification

Implicit identification is a term that is composed of two combined key concepts. The first key concept is identification, a mechanism related to user authentication. User authentication means that a computing system establishes trust in a user's identity through either verification or identification [18]. The second key concept of identification is its associated implicitness which is related to the interaction itself. Identification in the context of biometrics denotes the ability of a system to identify the user solely based on an obtained biometric sample [17]. It is, therefore, similar in concept to verification, which besides the sensor sample, also requires a claim of identity from the user (e.g., a provided user name) [17]. Identification is beneficial as it only requires the sensor sample and no other information from the user.

Hence, implicit identification is defined as the ability of a device to identify its users through "actions they would carry out anyway" [20]. Thereby, implicit identification is based on implicit interactions by the user [48]. As a consequence, the identification process does not demand time from the user, thus being transparent.

An implicit identification scheme thereby allows for the creation of a continuous identification system. Continuous identification denotes the ability of a system to continuously determine the user's

[1]VR Game: Beat Saber. https://beatsaber.com, last retrieved September 26, 2023.

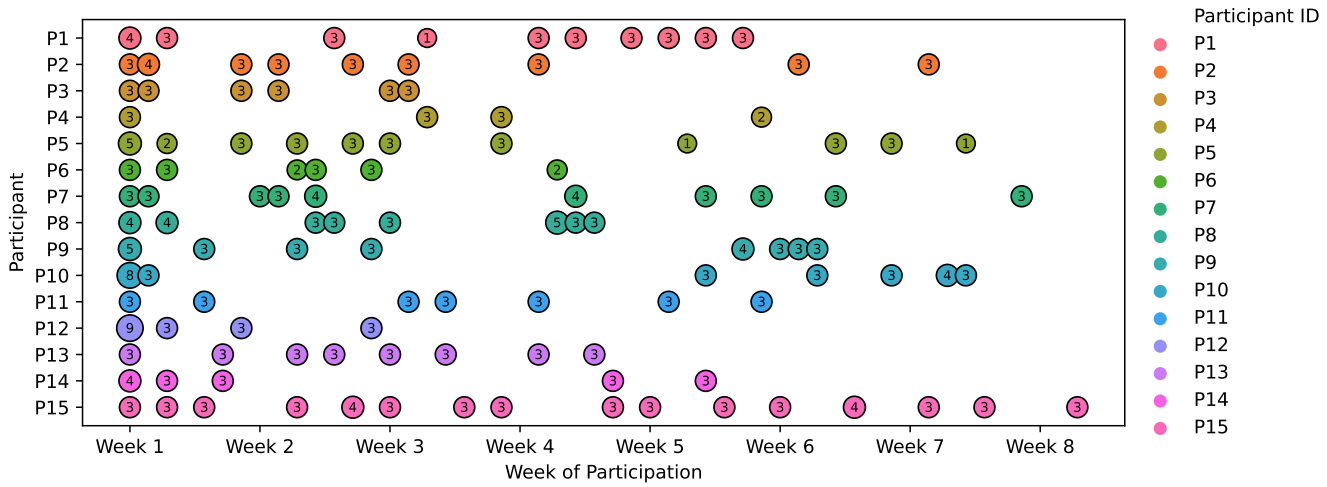


Figure 2: Overview of the days each participant contributed to the data set. Participants are color-coded. The number in every circle equals the number of recordings per day, which is also reflected in each circle's size.

identity, which stands in contrast to most other systems that determine the user's identity only once (e.g., at the beginning of a usage session.) [51, 52].

2.2 Behavioral Biometrics in VR

Behavioral biometrics and VR are a particularly good fit, as the rich interactivity of VR allows the implicit sampling of data elicited by the headset anyway. Additionally, traditional verification mechanisms such as pattern locks were shown to be insecure in VR [6, 53].

One such behavior is head movement. The movements associated with the head are a very distinct trait, as they combine head movement patterns with body height, where especially the latter is also a strong biometric indicator. M. R. Miller et al. for example showed an identification rate of over 90% in a user study with 511 participants, based on the head movements during the observation of a 360-degree VR video [33]. Prior to their work, Mustafa et al. and Sivasamy et al. conducted two works respectively, where Mustafa et al. found mean equal error rates of 7% when authenticating users based on their head and movement body patterns, and Sivasamy et al. found an accuracy of 99% for continuous authentication [30, 37].

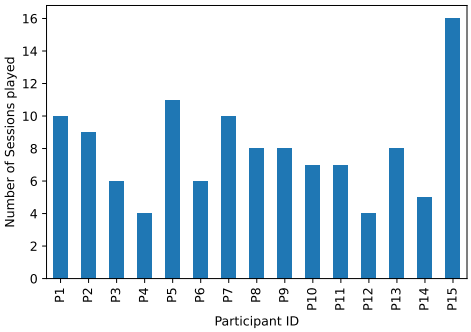
Besides the movement of the head, a number of works have been published that use full body motions in VR for user identification and many works utilize controllers for this reason. For example, Kupin et al. have shown that throwing a ball in VR results in distinct movement patterns using the features of the right controller, coining the term "task-driven biometrics" [23]. Miller et al. and Ajit et al. later also conducted works on the ball-throwing activity, using more features and deep learning, crossing the boundaries of VR systems and also including real-world constraints [1, 34, 35, 43]. Other examples of activities used for user identification in VR are full body kinesiological movements [40], games such as archery or bowling [27], pointing, grabbing, walking and typing [42], and the movements associated with the interaction with a Rubik's cube [32]. In contrast to Head-Mounted Displays (HMDs) that have controller tracking, research also showed that behavioral biometric modalities, such as head movement, finger-tracking, or eye-tracking can be a good fit if controllers are unavailable, such as on devices like Microsoft's Hololens 2 or Apple's Vision Pro [28, 29, 37].

However, little is known so far about the stability of behavioral biometrics and how it impacts identification systems. Although there have been carried out some works for other behavioral biometrics traits such as gait, it remains largely unclear how the change of behavior over time impacts identification systems, particularly for VR [7, 11, 14, 46]. An exception here is a work by Miller et al. who fused data sets to determine effects in motion behavior over a period of 7 to 18 months, finding differences in consistency of the movement, suggesting that short and medium timescales have an impact in altering VR behavior [36].

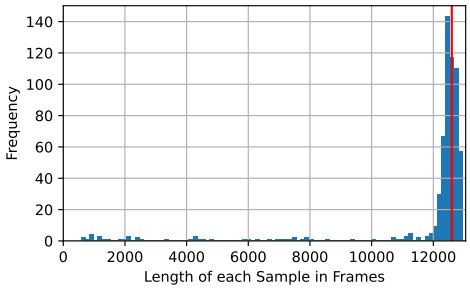
To the best of our knowledge, there is no other work besides the previous work of Miller et al. that explores full-body behavioral biometrics in VR over medium timescales [36]. Additionally, our work utilizes a field study under realistic conditions, whereas other works were primarily lab-based, and we provide a dense sample of recorded activities in VR (cf., Figure 2).

3 REMOTE FIELD STUDY

In previous work, researchers mostly used highly controlled lab environments to explore user identification systems. Moreover, they have been tested only for short periods of time, often within one session [29] or across two sessions at maximum [28, 36, 42]. Hence, the temporal stability of behavioral biometrics and its impact on identification performance remains mostly unaddressed. In this paper, we investigate the temporal stability of behavioral biometrics over a longer time (eight weeks). To do so, we conducted a remote field study for which we handed out *Meta Quest 2* headsets to participants and asked them to play the VR game *Beat Saber* at least twice a week to elicit spatiotemporal user data. Using live monitoring, we were able to follow the remote study and send out reminders when needed.



(a)



(b)

Figure 3: (a) Overview of the days ("sessions") each participant contributed to the data set. Note that participants contributed multiple songs per session. (b) Histogram of the length of each sample in the database, measured in frames. The red line marks the end of the song after 2:21 minutes. Samples that ended much earlier were unsuccessful or canceled attempts by the players, and samples that end a little earlier were subject to unplanned frame drops by the Meta Quest 2 device. Longer samples can occur due to device lag.

3.1 Research Questions

Since related work mainly focuses on the feasibility of behavioral biometrics in virtual reality or on session independence by using two dedicated sessions, it remains an open research question how such biometrics behave over more sessions and a longer period of time. Thus, the overarching research question of this work is:

(RQ) How stable are behavioral biometrics in VR over time?

3.2 Apparatus

The apparatus consists of an *Meta Quest 2*, the game *Beat Saber* and its modification, as well as a study dashboard to monitor the progress of the study.

Virtual Reality Headset. To achieve comparable results, we opted for using the same VR headset (i.e., a *Meta Quest 2*) for all participants. The *Meta Quest 2* consists of the head-mounted display and two hand-held controllers that use a six-degrees-of-freedom inside-out tracking system supporting positional and rotational tracking. The display runs at a variable refresh rate of 60 to 120 Hz and offers a field of view of 97°. It can operate without being connected to a computer, thus, being a wireless, consumer-grade device using wifi for its connection to the internet.

Beat Saber. For the VR game, we selected *Beat Saber*, a game in which users must slice floating cubes to the rhythm of a song (cf., Figure 1). We selected this virtual reality game for three main reasons: 1) games are the most popular application category and *Beat Saber* is currently the most popular VR application, 2) it requires continuous interaction as the distance between cubes that require slicing is short, and 3) it requires body movement over a larger space (boxes spawn in a different direction: up/down, left/right). The game is mostly stationary, i.e., players do not need to walk in the game but may have to turn themselves (depending on the game mode) or duck or sidestep when obstacles appear. The slicing of the blocks happens through light sabers that are attached to the player's controllers at a forward angle, and slicing the block in the rhythm of the game and in an optimal angle leads to an increase in player score. The player's score can further be increased by not making any mistakes for a sequence of blocks, and on the contrary, when one fails to slice the blocks correctly, it might lead to failing the song. *Beat Saber* offers multiple different songs that all have specific boxes appearing at pre-defined locations and with pre-defined rhythms. We selected the song *Commercial Pumping* since it contains a wide variety of different patterns and movements. The difficulty level influences the speed at which boxes appear. We opted for *Standard & Normal* settings since it would be feasible across participants. For comparability, all participants were asked to play the same song at the same difficulty level.

Beat Saber Modification. To elicit the participants' spatiotemporal motion data from *Beat Saber*, we create a background application[2] for the *Meta Quest 2*, which is our target VR device. Our application consists solely of a logger that elicits the user's coordinates for rotation and position of their head-mounted display and the left and right controller in Euler and Quaternion angles. It detects when the user launches a song in *Beat Saber* and terminates upon successful completion or fail and tracks their movement. Per each rendered frame of the game, our application fills a buffer and transmits this buffer to our server over an encrypted connection every two seconds through an HTTP POST request. Additionally, it transmits the name of the played song, its difficulty and modifier, the acquired user's score per frame, and a timestamp, as well as a user-specific ID token. We store this pseudonymized data in a relational database on our server (MariaDB 10.3 on Ubuntu 20.04).

Study Dashboard. To monitor participants and make sure that the data is properly sent to our server, we developed a web-based dashboard. The dashboard shows how often a specific player played the game and provided data to us (e.g., the number of songs played in the last two and seven days and the most played song) and it additionally provides descriptive statistics of the collected data set (e.g., recorded number of frames or songs). We used the data elicited from the dashboard in combination with a number of SQL-based

[2]BSMG Wiki. Making Mods. https://bsmg.wiki/modding, last retrieved on September 26, 2023.

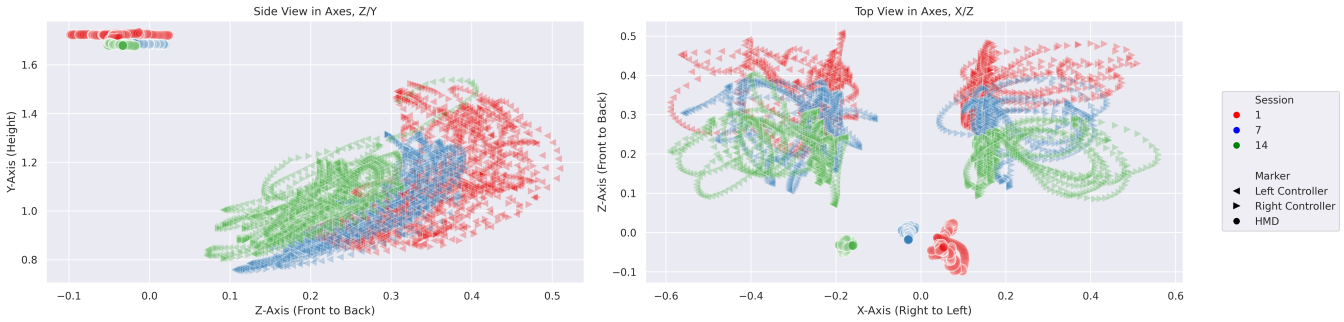


Figure 4: Movement data of participant P12 with 3 songs sampled at the beginning, mid, and end of the study. The expressiveness of the movement – particularly the head front-to-back movements on the z-axis – is reduced by time.

views to inform participants via e-mail of their progress and to remind them of their study participation.

3.3 Participants

We recruited 16 participants (8 female, 8 male) through University mailing lists aged between 25 and 33 years ($M = 27.77$, $SD = 2.78$). Unfortunately, one female participant broke her hand during a sports activity unrelated to our study. Hence, we excluded her data and conducted further analysis on the remaining data ($N = 15$).

3.4 Procedure

At the beginning of the study, we explained the procedure to each participant and asked each for their written and informed consent. We then fully answered any questions from the participants and pointed out that it is possible to cancel their participation in the study at any time without any detriments.

Next, we asked the participants to fill in a brief demographic questionnaire. We recorded two five-point Likert items at the beginning of the study. The first item asks participants to rate the statement "prior to participating in this study, I used VR regularly" on a scale from 1 (completely disagree) to 5 (completely agree). Participants' median response was 2 (IQR: 2.5). Furthermore, we asked to rate a second statement, "prior to participating in this study, I often played *Beat Saber*", on the same scale. Here, participants reported a median response of 2 (IQR: 3.0). We additionally asked participants for their dominant hand, and all were right-handed.

Since only four participants owned a *Meta Quest 2* device, we equipped the other twelve participants with a *Meta Quest 2* headset for the duration of the study. These devices were exclusively used for participation in the study. We instructed all four participants with their own private devices on how to install our background application and verified that the procedure was executed correctly. For the other twelve participants, we provided them with the head-mounted display with the application preinstalled by us. Participants could participate in the study at their preferred location; however, we instructed them to use the headset only in a safe space (i.e., indoors without any objects nearby and by enabling the guardian system). Furthermore, we explicitly asked participants not to give the headset to any other person, as we otherwise would not be able to distinguish the participants' data from any other data.

Overall, the field study took eight weeks. We asked participants to play at least twice a week on two distinct days three levels of *Beat Saber*, particularly the song "*Commercial Pumping*" in the game mode "Solo" on difficulty "*Standard & Normal*". Additionally, participants were able to play other songs in *Beat Saber* or other VR games on the provided HMD; we did not restrict their potential usage of the device. We chose "*Commercial Pumping*" as it has an average number of beats per minute across all songs of *Beat Saber*, a length of 2:21 minutes, and it includes dynamic obstacles that players need to sidestep and duck under. Figure 1 shows the three movements included in playing the song. Additionally, we chose this song to acquire comparable data, as the potential number of combinations for the song, its difficulty, and associated modifiers in *Beat Saber* is very high. Following this, participants invested less than 15 minutes per week in playing the game. The elicited data was collected implicitly in the background.

3.5 Ethics

We received ethical clearance from our local institutional review board at the University of Duisburg-Essen, Faculty of Business Administration and Economics, for conducting our user study. The findings in this work present insights into designing continuous identification systems. The authors want to remark that it is ethically required that the user's consent is acquired when employing the presented findings in a real-world deployed system.

4 RESULTS

In the following, we first describe the generated data set, the machine learning approach, and the evaluation results.

4.1 Data Set

Overall, participants played the designated "Commercial Pumping" song in Beat Saber 375 times during 119 sessions over the course of eight weeks (see Figure 2), which corresponds to 4,520,828 sampled frames and ca. 2.4 GiB of data[3]. One session corresponds to the number of plays during a calendar day per participant. We recorded the position and rotation of the headset and both controllers at a sampling frequency of 90 Hz. Naturally, we obtained an imbalanced

[3]Our data set and background application are publicly available online at http://research.hcigroup.de.

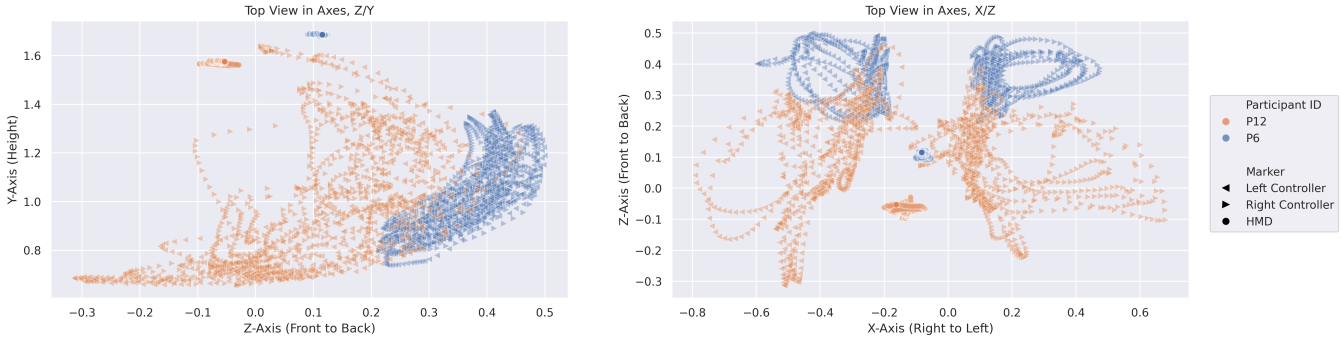
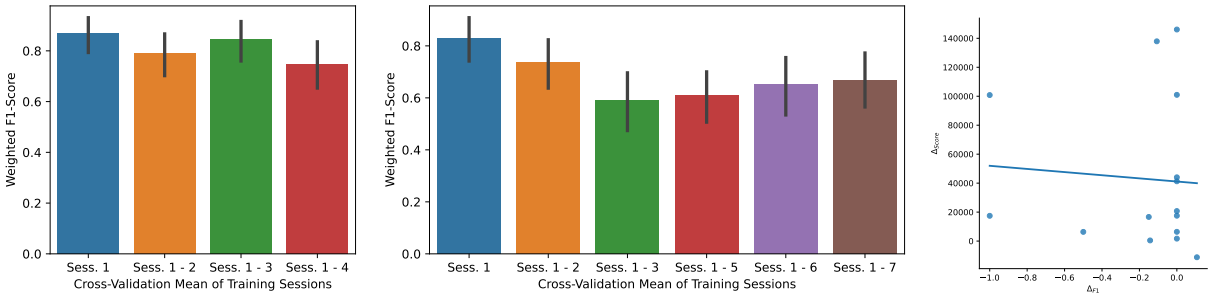


Figure 5: Movement data of participants P12 and P6 in comparison to each other from a side view (left) and top view (right). It is visible that P6's movements have more precision and are more targeted compared to P12. The Y-axis corresponds to "up".



(a) $N = 15$ (b) $N = 8$ (c) Δ_{Score} vs. $\Delta_{F1-Metric}$

Figure 6: Training bar plots in (a) and (b) show for $N = 15$ and $N = 8$ the cross-validated training performance of our model as denoted by F1-score, obtained from a k-cross-validation with $k = 5$. The model then is subject to testing as depicted in Figures 7(a) and 7(b). Error bars show standard errors. (c) Visualization of the correlated values of F1 score and acquired game score.

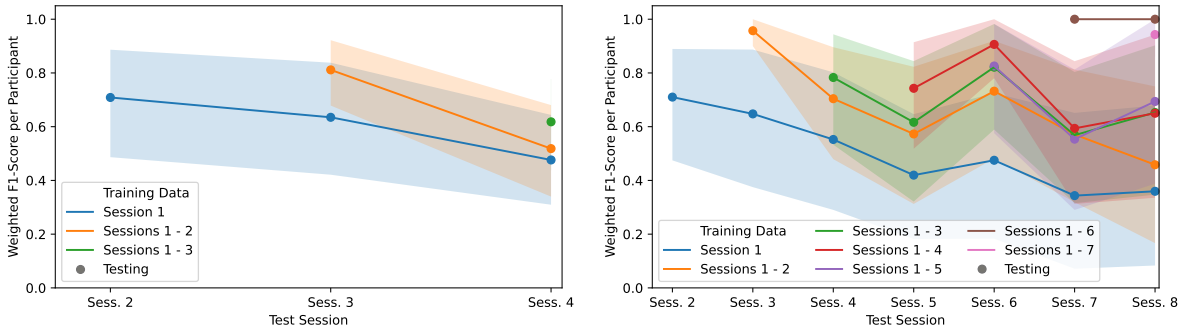
data set since every participant provided a different amount of data (cf., Figure 3(a)). Thus, participants contributed between 11 and 50 songs ($M = 25.53$, $SD = 9.30$).

Next, we plotted and inspected the elicited raw data for outliers to verify if the elicitation worked correctly. We found a few songs that ended too early and verified the cause together with the participant, which yielded that the song ended since the participant involuntarily lost the game (e.g., by failing to evade the obstacles when they appeared for the first time – cf. Figures 1(b) and 1(c)). We furthermore found samples in the data set that ended right after the start, and it turned out that participants started the song mistakenly and canceled it immediately. Therefore, we set up a filter that removes any song from the data set that has a shorter duration than 10 seconds.

In this data set, each sample corresponds to one song. However, the songs are grouped by a second metric, which is the "session" day, denoting on which day since the start of the study the participants played the song (e.g., the third day of participation corresponds to the third session). Here, songs are grouped by calendar day, as multiple songs were played per session. Thereby, participants contributed between one and nine songs for a single day. Since not all participants contributed the same number of sessions, we use two subsets of the data for further analysis. First, we evaluate our approach with all participants ($N = 15$) and with four sessions. Second, we used another subset of $N = 8$ and evaluated their performance with eight sessions.

Figure 3(a) shows the differences in the number of provided samples per participant in the data set. We see that eight participants provided eight complete sessions played (P1, P2, P5, P7, P8, P9, P13, and P15). For all 15 participants, we see that only four sessions were absolved across all, as P4 imposes a lower limit.

We, therefore, opt for a split. We can create a balanced subset of our data set for $N = 15$ that is suitable for cross-validation and it can be tested with sessions number two to four. Additionally, we also train a model for $N = 8$, using sessions two to eight for testing. Consequently, we can create two subsets of our data set, one for testing with three sessions and one for testing with six sessions, where both subsets are balanced in terms of class distribution after applying the sampling, which is particularly important for the training of the Random Forest model.



(a) $N = 15$ (b) $N = 8$

Figure 7: Average identification results over time, taken from the test set for a continuous identification system. Previous data recordings are used as training data and subsequent data of each week for testing. The circular markers show the point in time when the system was tested, and the point in time of training is not depicted.

4.2 Sampling and Preprocessing

To oppose the prevalent class imbalance within our data set, we meet two precautions. First, we randomly sample a balanced number of samples from our data set every time we train or validate our model. We apply a random undersampling for overrepresented classes and a random oversampling to underrepresented classes. Thereby, although the size per class differs (cf., Figure 3(a)), we obtain a balanced number of samples per class. We set three as the number of samples to sample to, as this number corresponds to the number of repetitions that the song should be played per session in our study. Our strategy first tries to sample without any redraws, but if it cannot, it might draw samples with replacement (i.e., upsample). This approach, of course, comes with the drawback that we cannot fully utilize our data set and that the over-sampling may sample the same data twice. Second, as for a metric, we focus on the weighted F1 score that would take any prevalent class imbalance into account as it calculates the F1 score, which is the harmonic mean of precision and sensitivity, weighted by the given support per class.

In the next step, we unify the data shape of our elicited data. Our data consists of three positional columns ("pos.x", "pos.y", and "pos.z"), where "pos.x" corresponds to right and left, "pos.y" to the height of the player and "pos.z" to the front and back. Here, "pos.x" is positive to the right and "pos.z" positive to the front. Additionally, we take the four Quaternion-based rotation coordinates into account ("quat.x", "quat.y", "quat.z", and "quat.w").

As previous works suggest, the head is a strong biometric feature [3, 28, 30, 31, 37]. One particularly strong biometric property that can be inferred from the head is the user's height, corresponding to the "pos.y"-feature of the head-mounted display. Also, another side effect that can occur is the initial positioning of the user within their tracking space which might be dependent on other objects in the room, such as furniture. To resolve both issues, we transform the head-mounted display's and both controllers' positional coordinates by calculating the offset vectors from the headset to the respective controller by subtracting their absolute positional coordinates. The resulting vector describes where the hands are positioned in relationship to the head, but it does not bear any absolute height information anymore, nor does it express information about the user's positioning within the tracking space. However, we acknowledge that those vectors still might be loosely correlated with a user's height to a certain extent, as the arm-length correlates with height, and this might affect the vector's maximum length (cf., da Vinci's principle of the Vitruvian man). We, therefore, obtain in total 18 features (two times "pos.x", "pos.y" and "pos.z" for each offset vector and three times "quat.x", "quat.y", "quat.z" and "quat.w" for the rotation of the headset and both controllers).

As for the length of each sample, we find that the mean song length is 11631 frames long (SD: 2788.58). Since this standard deviation is high, we explore the data and find that it originates from two factors: first, a number of songs have been exited early, either voluntarily (i.e., the player stopping the game) or involuntarily (i.e., the player failing the song). Additionally, we find that the length of each song also varies a little, most likely due to occurring frame drops. Figure 3(b) depicts a histogram. To unify the length of data, we choose to calculate the minimum, maximum, mean, and standard deviation for each of the 18 different features. Hence, we obtain a feature vector that always corresponds to a length of 72 values, encoding each feature column by four floating point values.

We furthermore inspect and visualize the elicited data during this step. Figure 5 depicts the spatial movements of two players plotted in a comparison against each other. Here, the differences in style can be seen as the movements of P6 are more focused and precise in comparison to P12. Figure 4 also shows the learning of P12, where it is notable that P12 performs fewer movements after sessions 7 and 14 in comparison to one, as for example, the head is moving less in the later participation within our study. Hence, the movements are becoming more focused and precise again.

4.3 Machine Learning Model for Identification and Training Process

To identify our participants, we utilize a closed-set Random Forest multiclass-classifier from scikit-learn [41]. Random Forests were used by previous works as means for user identification, and we choose our procedure to be similar to previous work to enable comparability [28, 42]. For this reason, we also intentionally leave the model's default parameters in place (e.g., $n\_estimators = 100$). We also considered deep learning models for this task, as these models are frequently used in behavioral biometrics [27, 29, 36, 44]. However, we refrained from using deep learning since Random Forests allow for a feature analysis using their mean decrease in impurity, and we wanted to obtain insights into the model's learned features. Finally, we train the Random Forest model with the feature vectors obtained from our preprocessing and participants' labels.

Due to the large amount of data we elicited in our study over time, we opt for a Training-Validation-Testing-process for evaluating our machine learning model. The training set is used to train our model, and the validation set is used to determine its performance with regard to the given training set. In general practice, the validation set can be used to determine model parameters and model performance during training. However, a test set must only be used to test the model's performance once and must not be used to tune the overall process. Therbey, we stand in contrast to previous works that elicited data mostly during a single study day [29, 33] or two days [27, 28, 36, 42] and split their data into training and validation, often not having access to a true distinct test set.

For the training of our Random Forest, we group the data by each participant's study session day, which is counted relative to the beginning of their participation in the study. Here, for example, the second study session is the second day when they participated in the study, independent of the time that passed between the first and second day. As each sample in the data set corresponds to a song, we thereby group the samples by day. We use a 5-fold cross-validation to evaluate our model within-session, choosing an 80% to 20% split for training and validation, respectively. We randomly sample the set from our data set and then apply the cross-validation to this set. The ratio of samples belonging to each class is balanced with regard to the classes, as otherwise, the model could be biased by overrepresented classes. For evaluation between sessions, we first train the Random Forest model with all balanced, sampled data from the previous day or days in chronological order.

4.4 Single Training Session Approach

Initially, we create a traditional identification system by using the data from the very first session to train and cross-validate our model. We used all songs that participants provided except for one song that was used for validation and performed k-fold cross-validation with $k = 5$. We use the two subsets of our data set for $N = 15$ and $N = 8$ for training our model, respectively. The training results are depicted in Figure 6, and the testing results are shown in Figure 6. The F1 Score drops from a cross-validated F1 score of 86% at the beginning for $N = 15$ to 48% after testing with session 4, a difference of -38 percentage points (cf., Table 2). For $N = 8$, we find a mean F1 score of 83% at the beginning and 27% at testing with session 8, a decrease of -56 percentage points (cf., Table 1).

4.5 Multi-Session Training Approach

Next, we create an identification system that is trained with multiple sessions, meaning we perform frequent re-training with new data that our system is able to acquire. We, therefore, test with session n and train with session $[1, ..., (n-1)]$ (e.g., to evaluate the data recorded in week three, we train our classifier with the data of the first two weeks). Figure 7 depicts the performance of the different classifiers over time. Note that the blue lines in Figure 7 correspond to the blue lines in Figure 8. Overall, we trained 3 classifiers for $N = 15$ and 7 classifiers for $N = 8$, evaluating their performance for the subsequent weeks. Table 2 contains the results for four sessions and $N = 15$ and Table 1 for $N = 8$ and eight sessions. Our results show that a classifier trained more recently in the majority of cases outperforms a classifier that has been trained earlier on fewer data. We found a significantly strong negative correlation (Pearson's coefficient) between the number of past days (i.e., the days between training to test) and the F1-score ($r(13) = -0.828$, $p < 0.001$).

4.6 Performance and Learning

Besides testing our data set for identification performance, we also question whether participants got better at playing the game and if their behavior would reflect this. We choose the logged game scores of the game as a metric and determine the mean game score per participant at the beginning of the study and at the end of the study, i.e., during their first and last session. We subtract these values and determine $\Delta_{Score} = Mean(Score_{End}) - Mean(Score_{Beginning})$. Next, we determine $\Delta_{F1} = Mean(F1_{End}) - Mean(F1_{Beginning})$ alike the score, by subtracting the F1 scores per participant.

We then correlate the Δ_{Score} values with Δ_{F1} that we obtained from testing with the last session for $N = 15$. We find that the change in-game performance and the identification metric have a close to zero correlation ($r(13) = -0.07$), indicating no relationship. We could not find a significant linear correlation ($p = .792$). Figure 6(c) visualizes the correlation.

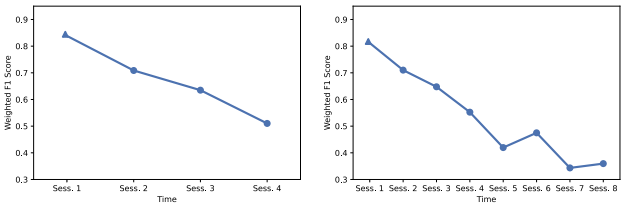
We furthermore consider the influence of participants' regular VR usage and whether they played *Beat Saber* often before on participants' performance and learning. To do so, we linearly correlate their self-reported VR and *Beat Saber* experience (five-point Likert items) with the maximum in-game score they acquired in *Beat Saber* during their very first day of participation in the study given the questions "prior to participating in this study I used VR regularly" and "prior to participating in this study, I often played *Beat Saber*". We find a strong positive linear correlation between their previous VR usage ($\rho(13) = 0.7333, p = 0.0019$) and their acquired score on the first day and also a strong linear correlation between their previous frequency of playing *Beat Saber* before the study and their acquired score ($\rho(13) = 0.8490, p < 0.0001$). We repeat this procedure by correlating participants' maximum in-game scores of their last participation day of the study for the same respective Likert items and find again two strong positive linear correlations (VR usage: $\rho(13) = 0.5193, p = 0.0473$ and frequency of playing *Beat Saber*: $\rho(13) = 0.6973, p = 0.0039$). Participants' scores median increase is 10815 points (IQR: 31880) between their first and last participation in the study.

Table 1: Mean (Standard Deviation) of the identification results (F1-score) for $N = 8$ participants for a continuous identification system. The training column shows the origin of the training data.

	Testing						
Training	**Session 2**	**Session 3**	**Session 4**	**Session 5**	**Session 6**	**Session 7**	**Session 8**
Session 1	0.71 (0.33)	0.65 (0.41)	0.55 (0.41)	0.42 (0.37)	0.48 (0.41)	0.34 (0.44)	0.27 (0.39)
Session 1 to 2		0.96 (0.08)	0.70 (0.34)	0.57 (0.39)	0.73 (0.34)	0.57 (0.40)	0.34 (0.40)
Session 1 to 3			0.78 (0.33)	0.62 (0.39)	0.82 (0.34)	0.57 (0.38)	0.49 (0.44)
Session 1 to 4				0.74 (0.33)	0.91 (0.19)	0.59 (0.42)	0.57 (0.48)
Session 1 to 5					0.83 (0.35)	0.55 (0.42)	0.61 (0.51)
Session 1 to 6						1.00 (0.00)	0.75 (0.46)
Session 1 to 7							0.71 (0.44)

Table 2: Mean (Standard Deviation) of the identification results (F1-score) for $N = 15$ participants for a continuous identification system. The training column shows the origin of the training data.

	Testing		
Training	**Session 2**	**Session 3**	**Session 4**
Session 1	0.71 (0.42)	0.63 (0.42)	0.48 (0.34)
Session 1-2		0.81 (0.26)	0.48 (0.36)
Session 1-3			0.62 (0.36)



(a) $N = 15$ (b) $N = 8$

Figure 8: Average identification results for training with only the first session (Sess.) denoted by the F1 score. The very first session is used as training data, and the subsequent data of each further study participation session is for testing the classifier. The triangle marker depicts the point in time with which the model was trained and cross-validated (only data from the first session went into training and testing). The circular marker indicates the session with which a test of the classifier was performed (the classifier was trained with the data of the first day and tested on the subsequent ones.

4.7 Feature Analysis

We conducted a feature analysis to understand what movements contribute most to our classifiers. We analyzed the resulting five most influential features for the $N = 8$ classifiers by determining the mean decrease in impurity (MDI) of the Random Forest, which ranks features that the Random Forest was trained on by their importance [26]. We found that for the first classifier (i.e., the one trained with the data from the first session), the head rotation is three times among the five most influential features. For the classifiers that include data from other sessions, the head rotation was never part of the five most influential features. Additionally, we found that the rotation of the right controller is present among each of the five most influential features for each classifier.

5 DISCUSSION

In the following, we discuss our findings concerning the temporal stability of behavioral biometrics. First, we reflect on classification performance and thereafter discuss different training strategies. We conclude with limitations and research directions for future work.

5.1 Classification Performance

We overall achieved an identification performance, denoted by F1-score, of up to 86% for a cross-validated single-session and 71% for a two-session evaluation. Additionally, we find that it largely degrades over time, as seen in Figure 8. This sometimes happens monotonously falling (cf., Figure 8(a)), but there can also be outliers in the overall trend (cf., Session 6 in Figure 8(b)). The reason for this bump remains, unfortunately, unclear: it might be a random effect that the rules for the Random Forest that were created during the very first training session suddenly apply slightly better to Session 6 compared to Session 5. By checking the data, we found that a single participant's performance did create the increase in Session 6. Although we excluded data with obvious flaws (e.g., participant 12 changed the height from 1.5m in every other song to 2.0m in one specific song), the type of study might add further noise to the data (e.g., participants are disturbed while playing) which is limited in lab-based user studies.

We also wondered whether participants' learning of *Beat Saber* would influence the identification performance. While we could find strong statistically significant correlations given participants' prior knowledge of VR and *Beat Saber* with regards to their performance as denoted by the in-game score at the beginning of the study, we also saw that this correlation became less strong towards the end of the study. Thus, it turned out that the participation led to an approximation of participants' performance over time, yet the correlation between participants' Δ_{Score} vs. $\Delta_{F1-Metric}$ did not turn out to be statistically significant. Thereby, we could not find evidence for an influence of participants' in-game performance denoted by the score on the identification rate.

5.2 Upfront Training vs. Multi-Session Training

Our results show that classifiers trained with more recent data generally perform better than classifiers trained with the initial data. Thus, continuous training improved the stability of behavioral biometric systems. Our findings also suggest that the features influencing the classifier most change after the first session. While most research conducted studies with at max two sessions, it remains unclear if their results change in a similar way and if that influences identification performance. In a practical scenario this would mean that such a system requires constant re-training. This has ethical implications, such as users being aware of the system and how their data is processed. It is important that this processing occurs with the consent of the individuals involved, which is a key requirement for all biometric identification systems for personal identification.

5.3 Limitations

We acknowledge the following limitation to our study. Studies outside the lab after highly influenced by the context in which the study is conducted. Since we conducted a remote field study, we had little to no control over the context in which participants played *Beat Saber*. While the VR HMD shielded participants from direct influences of their environment, other factors still came into play. Participants, for example, did not equally contribute to our data set. Reasons included being on vacation during the study or forgetting to play. This also influenced the length of the time period between the two sessions. At last, as all our participants were right-handed, we did not evaluate our approach on a mixture of left-handed and right-handed people, which might impact the results. Another inherent methodologic drawback of conducting a remote field study is the limited control of participants' usage of the devices. We did not ask our participants to use the VR device only for the study; they potentially could play other games in VR or other levels in *Beat Saber*. We acknowledge that this could influence our results. However, we see this as an essential part of the chosen methodology that yields data obtained under realistic conditions.

5.4 Future Work

In this paper, we explored spatiotemporal user data elicited with the VR game Beatsaber over the course of eight weeks. We will continue to collect user data to investigate the long-term stability of behavioral biometrics and continue to update our data set. Moreover, we plan to share our modified BeatSaber version with a wider audience, deploying it in the wild. Thereby, we aim to not only investigate the effect of time but also consider a larger sample population. Finally, in this paper, we focused on explainable machine learning algorithms that require manual feature engineering. In the future, we will investigate deep learning with a sliding window approach to examine the upper limit for user identification performance [25]. The data elicited in our study was obtained using an HMD with controller tracking and the *Beat Saber* application. Another research opportunity would be to see whether comparable results could be obtained from other application contexts or from devices that make no use of controllers. Moreover, the release of our data set allows for further experimentation; for example, an analysis of the hysteresis component in the data could be conducted to understand the potential lag of identification performance with relation to participants' in-game performance.

6 CONCLUSION

In this work, we report on a remote field study that identifies $N = 15$ participants while playing *Beat Saber*. Our work contributes to the understanding of behavioral biometrics for virtual reality by providing insights into a mid-term study which we evaluated with up to 8 sessions per participant. We found that the identification performance decreases over time and that continuous training with recently recorded data can improve the tracking performance. Our results shed light on the influence of multiple sessions on identification performance and how continuous retraining of classifiers can help to improve identification systems.

ACKNOWLEDGMENTS

We thank Emmanuel Anastas Mbawala for his support in the creation phase of the background application. The presented work was funded by the German Research Foundation (DFG) under project no. 426052422.

REFERENCES

[1] A. Ajit, N. Banerjee, and S. Banerjee. 2019. Combining Pairwise Feature Matches from Device Trajectories for Biometric Authentication in Virtual Reality Environments. In *2019 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*. IEEE Computer Society, Los Alamitos, CA, USA, 9–97. https://doi.org/10.1109/AIVR46125.2019.00012

[2] Florian Alt, Stefan Schneegass, Jens Grossklags, Heather Richter Lipford, and Jessica Staddon. 2022. Beyond Passwords—Challenges and Opportunities of Future Authentication. *IEEE Security & Privacy* 20, 1 (2022), 82–86. https://doi.org/10.1109/MSEC.2021.3127459

[3] Arman Bhalla, Ivo Sluganovic, Klaudia Krawiecka, and Ivan Martinovic. 2021. MoveAR: Continuous Biometric Authentication for Augmented Reality Headsets. In *Proceedings of the 7th ACM on Cyber-Physical System Security Workshop*. Association for Computing Machinery, New York, NY, USA, 41–52. https://doi.org/10.1145/3457339.3457983

[4] Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Frank Stajano. 2015. Passwords and the evolution of imperfect authentication. *Communications of the ACM* 58, 7 (2015), 78–87. https://doi.org/10.1145/2699390

[5] Sascha Brostoff and M. Angela Sasse. 2003. "Ten strikes and you're out": Increasing the number of login attempts can improve password usability. In *Workshop on Human-Computer Interaction and Security Systems at CHI 2003*. ACM, Fort Lauderdale, Florida, USA, 4 pages. https://discovery.ucl.ac.uk/id/eprint/19826/

[6] Ceenu George, Mohamed Khamis, Emanuel von Zezschwitz, Henri Schmidt, Marinus Burger, Florian Alt, and Heinrich Hu\textsectionmann. 2017. Seamless and Secure VR: Adapting and Evaluating Established Authentication Systems for Virtual Reality. In *Proceedings 2017 Workshop on Usable Security*. Internet Society, San Diego, CA, USA. https://doi.org/10.14722/usec.2017.23028

[7] Peter P. K. Chan, Chao-Ying Chen, Hussein Ayache, Lobo Louie, Alan Lok, Nathan Cheung, and Roy T. H. Cheung. 2021. Gait difference between children aged 9 to 12 with and without potential depressive mood. *Gait & posture* 91 (2021), 126–130. https://doi.org/10.1016/j.gaitpost.2021.10.012

[8] David Checa and Andres Bustillo. 2020. A review of immersive virtual reality serious games to enhance learning and training. *Multimedia Tools and Applications* 79, 9-10 (2020), 5501–5527. https://doi.org/10.1007/s11042-019-08348-9

[9] Vuthea Chheang, Patrick Saalfeld, Fabian Joeres, Christian Boedecker, Tobias Huber, Florentine Huettl, Hauke Lang, Bernhard Preim, and Christian Hansen. 2021. A collaborative virtual reality environment for liver surgery planning. *Computers & Graphics* 99 (2021), 234–246. https://doi.org/10.1016/j.cag.2021.07.009

[10] Sarah Faltaous, Jonathan Liebers, Yomna Abdelrahman, Florian Alt, and Stefan Schneegass. 2019. VPID: Towards Vein Pattern Identification Using Thermal Imaging. *i-com* 18, 3 (2019), 259–270. https://doi.org/10.1515/icom-2019-0009

[11] A. Gabell and U.S.L. Nayak. 1984. The Effect of Age on Variability in Gait. *Journal of Gerontology* 39, 6 (1984), 662–666. https://doi.org/10.1093/geronj/39.6.662

[12] Ceenu George, Daniel Buschek, Andrea Ngao, and Mohamed Khamis. 2020. GazeRoomLock: Using Gaze and Head-Pose to Improve the Usability and Observation Resistance of 3D Passwords in Virtual Reality. In *Augmented Reality, Virtual Reality, and Computer Graphics*, Lucio Tommaso de Paolis and Patrick Bourdot (Eds.). Springer International Publishing, Cham, 61–81.

[13] John Harvey, John Campbell, Stephen Elliott, Michael Brockly, and Andy Adler. 2017. Biometric Permanence: Definition and Robust Calculation. In *2017 Annual*

IEEE International Systems Conference (SysCon). IEEE, 1–7. https://doi.org/10.1109/SYSCON.2017.7934760

[14] John H. Hollman, Robert H. Brey, Richard A. Robb, Tami J. Bang, and Kenton R. Kaufman. 2006. Spatiotemporal gait deviations in a virtual reality environment. *Gait & posture* 23, 4 (2006), 441–444. https://doi.org/10.1016/j.gaitpost.2005.05.005

[15] J. Bonneau. 2012. The Science of Guessing: Analyzing an Anonymized Corpus of 70 Million Passwords. In *2012 IEEE Symposium on Security and Privacy*. IEEE, San Francisco, CA, USA, 538–552. https://doi.org/10.1109/SP.2012.49

[16] A. Jain, Lin Hong, and R. Bolle. 1997. On-line fingerprint verification. *IEEE transactions on pattern analysis and machine intelligence* 19, 4 (1997), 302–314. https://doi.org/10.1109/34.587996

[17] Anil K. Jain, Patrick Flynn, and Arun A. Ross (Eds.). 2008. *Handbook of biometrics*. Springer, New York, NY. https://doi.org/10.1007/978-0-387-71041-9

[18] Anil K. Jain, Arun A. Ross, and Karthik Nandakumar. 2011. Introduction. In *Introduction to Biometrics*, Anil K. Jain, Arun A. Ross, and Karthik Nandakumar (Eds.). Springer US, Boston, MA, 1–49. https://doi.org/10.1007/978-0-387-77326-1_1

[19] Anil K. Jain, Arun A. Ross, and Karthik Nandakumar (Eds.). 2011. *Introduction to Biometrics*. Springer US, Boston, MA. https://doi.org/10.1007/978-0-387-77326-1

[20] Markus Jakobsson, Elaine Shi, Philippe Golle, and Richard Chow. 2009. Implicit Authentication for Mobile Devices. In *Proceedings of the 4th USENIX Conference on Hot Topics in Security (HotSec'09)*. USENIX Association, USA, 9.

[21] Christina Katsini, Yasmeen Abdrabou, George E. Raptis, Mohamed Khamis, and Florian Alt. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Regina Bernhaupt, Florian 'Floyd' Mueller, David Verweij, Josh Andres, Joanna McGrenere, Andy Cockburn, Ignacio Avellino, Alix Goguey, Pernille Bjørn, Shengdong Zhao, Briane Paul Samson, and Rafal Kocielnik (Eds.). ACM, New York, NY, USA, 1–21. https://doi.org/10.1145/3313831.3376840

[22] Kat Krol, Eleni Philippou, Emiliano De Cristofaro, and M Angela Sasse. 2015. "They brought in the horrible key ring thing!" Analysing the Usability of Two-Factor Authentication in UK Online Banking. *arXiv preprint arXiv:1501.04434* (2015).

[23] Alexander Kupin, Benjamin Moeller, Yijun Jiang, Natasha Kholgade Banerjee, and Sean Banerjee. 2019. Task-Driven Biometric Authentication of Users in Virtual Reality (VR) Environments. In *MultiMedia Modeling*, Ioannis Kompatsiaris, Benoit Huet, Vasileios Mezaris, Cathal Gurrin, Wen-Huang Cheng, and Stefanos Vrochidis (Eds.). Lecture Notes in Computer Science, Vol. 11295. Springer International Publishing, Cham, 55–67. https://doi.org/10.1007/978-3-030-05710-7{_}5

[24] LastPass by LogMeIn. 2019. The 3rd Annual Global Password Security Report. https://lp-cdn.lastpass.com/lporcamedia/document-library/lastpass/pdf/en/LMI0828a-IAM-LastPass-State-of-the-Password-Report.pdf

[25] Arthur Le Guennec, Simon Malinowski, and Romain Tavenard. 2016. Data Augmentation for Time Series Classification using Convolutional Neural Networks. In *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*. HAL, Riva Del Garda, Italy, 9 pages. https://halshs.archives-ouvertes.fr/halshs-01357973

[26] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (2001), 5–32. https://doi.org/10.1023/a:1010933404324

[27] Jonathan Liebers, Mark Abdelaziz, Lukas Mecke, Alia Saad, Jonas Auda, Uwe Gruenefeld, Florian Alt, and Stefan Schneegass. 2021. Understanding User Identification in Virtual Reality Through Behavioral Biometrics and the Effect of Body Normalization. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411764.3445528

[28] Jonathan Liebers, Sascha Brockel, Uwe Gruenefeld, and Stefan Schneegass. 2022. Identifying Users by Their Hand Tracking Data in Augmented and Virtual Reality. *International Journal of Human–Computer Interaction* (2022), 28 pages. https://doi.org/10.1080/10447318.2022.2120845

[29] Jonathan Liebers, Patrick Horn, Christian Burschik, Uwe Gruenefeld, and Stefan Schneegass. 2021. Using Gaze Behavior and Head Orientation for Implicit Identification in Virtual Reality. In *Proceedings of the 27th ACM Symposium on Virtual Reality Software and Technology (VRST '21)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3489849.3489880

[30] M. Sivasamy, V. N. Sastry, and N. P. Gopalan. 2020. VRCAuth: Continuous Authentication of Users in Virtual Reality Environment Using Head-Movement. In *2020 5th International Conference on Communication and Electronics Systems (ICCES)*. IEEE, Coimbatore, India, 518–523. https://doi.org/10.1109/ICCES48766.2020.9137914

[31] Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A. Landay, and Jeremy N. Bailenson. 2020. Personal identifiability of user tracking data during observation of 360-degree VR video. *Scientific Reports* 10, 1 (2020), 10 pages. https://doi.org/10.1038/s41598-020-74486-y

[32] Florian Mathis, John Williamson, Kami Vaniea, and Mohamed Khamis. 2020. RubikAuth: Fast and Secure Authentication in Virtual Reality. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, New York, NY, USA, 1–9. https://doi.org/10.1145/3334480.3382827

[33] Mark Roman Miller, Fernanda Herrera, Hanseul Jun, James A. Landay, and Jeremy N. Bailenson. 2020. Personal identifiability of user tracking data during observation of 360-degree VR video. *Scientific Reports* 10, 1 (2020), 17404. https://doi.org/10.1038/s41598-020-74486-y

[34] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. 2021. Using Siamese Neural Networks to Perform Cross-System Behavioral Authentication in Virtual Reality. In *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*. IEEE, Lisboa, Portugal, 140–149. https://doi.org/10.1109/VR50410.2021.00035

[35] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. 2022. Combining Real-World Constraints on User Behavior with Deep Neural Networks for Virtual Reality (VR) Biometrics. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, Christchurch, New Zealand, 409–418. https://doi.org/10.1109/VR51125.2022.00060

[36] Robert Miller, Natasha Kholgade Banerjee, and Sean Banerjee. 2022. Temporal Effects in Motion Behavior for Virtual Reality (VR) Biometrics. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*. IEEE, 563–572. https://doi.org/10.1109/VR51125.2022.00076

[37] Tahrima Mustafa, Richard Matovu, Abdul Serwadda, and Nicholas Muirhead. 2018. Unsure How to Authenticate on Your VR Headset? Come on, Use Your Head!. In *Proceedings of the Fourth ACM International Workshop on Security and Privacy Analytics (IWSPA '18)*. Association for Computing Machinery, New York, NY, USA, 23–30. https://doi.org/10.1145/3180445.3180450

[38] Alex Nosenko, Yuan Cheng, and Haiquan Chen. 2022. Learning Password Modification Patterns with Recurrent Neural Networks. In *Secure Knowledge Management In The Artificial Intelligence Era*, Ram Krishnan, H. Raghav Rao, Sanjay K. Sahay, Sagar Samtani, and Ziming Zhao (Eds.). Springer International Publishing, Cham, 110–129.

[39] L. O'Gorman. 2003. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE* 91, 12 (2003), 2021–2040. https://doi.org/10.1109/JPROC.2003.819611

[40] Ilesanmi Olade, Charles Fleming, and Hai-Ning Liang. 2020. BioMove: Biometric User Identification from Human Kinesiological Movements for Virtual Reality Systems. *Sensors* 20, 10 (2020), 2944.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[42] Ken Pfeuffer, Matthias J. Geiger, Sarah Prange, Lukas Mecke, Daniel Buschek, and Florian Alt. 2019. Behavioural Biometrics in VR: Identifying People from Body Motion and Relations in Virtual Reality. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300340

[43] R. Miller, N. K. Banerjee, and S. Banerjee. 2020. Within-System and Cross-System Behavior-Based Biometric Authentication in Virtual Reality. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*. IEEE, Atlanta, Georgia, USA, 311–316. https://doi.org/10.1109/VRW50115.2020.00070

[44] Christian Rack, Andreas Hotho, and Marc Erich Latoschik. 2022. Comparison of Data Encodings and Machine Learning Architectures for User Identification on Arbitrary Motion Sequences. , 11-19 pages. https://doi.org/10.1109/AIVR56993.2022.00010

[45] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. 2016. Understanding Password Choices: How Frequently Entered Passwords Are Re-used across Websites. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 175–188. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/wash

[46] Alia Saad, Nick Wittig, Uwe Gruenefeld, and Stefan Schneegass. 2022. A Systematic Analysis of External Factors Affecting Gait Identification. In *IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, Abu Dhabi, United Arab Emirates, 9 pages. https://doi.org/10.1109/IJCB54206.2022.10007994

[47] M. Angela Sasse, Michelle Steves, Kat Krol, and Dana Chisnell. 2014. The Great Authentication Fatigue – And How to Overcome It. In *Cross-cultural design*, David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Doug Tygar, Gerhard Weikum, P. L. Patrick Rau, and P. PatrickL. Rau (Eds.). Lecture notes in computer science Information systems and application, incl. Internet/web and HCI, Vol. 8528. Springer, Cham, 228–239. https://doi.org/10.1007/978-3-319-07308-8{_}23

[48] Albrecht Schmidt. 2000. Implicit human computer interaction through context. *Personal Technologies* 4, 2-3 (2000), 191–199. https://doi.org/10.1007/BF01324126

[49] Maximilian Schrapel, Dennis Grannemann, and Michael Rohs. 2022. Sign H3re: Symbol and X-Mark Writer Identification Using Audio and Motion Data from a Digital Pen. In *Mensch und Computer 2022 - Tagungsband*, Bastian Pfleging, Kathrin Gerling, and Sven Mayer (Eds.). ACM, New York, 209–218. https://doi.org/10.1145/3543758.3543764

[50] Shen Yiran, Wen Hongkai, Luo Chengwen, Xu Weitao, Zhang Tao, Hu Wen, and Rus Daniela. 2019. GaitLock: Protect Virtual and Augmented Reality Headsets Using Gait. *IEEE Transactions on Dependable and Secure Computing* 16, 3 (2019), 484–497. https://doi.org/10.1109/TDSC.2018.2800048

[51] Issa Traore and Ahmed Awad E. Ahmed (Eds.). 2012. *Continuous Authentication Using Biometrics: Data, Models, and Metrics*. IGI Global, Hershey, PA, USA.

[52] Issa Traoré and Ahmed Awad E. Ahmed. 2012. Introduction to Continuous Authentication. In *Continuous Authentication Using Biometrics: Data, Models, and Metrics*, Issa Traore and Ahmed Awad E. Ahmed (Eds.). IGI Global, Hershey, PA, USA, 1–22. https://doi.org/10.4018/978-1-61350-129-0.ch001

[53] Z. Yu, H. Liang, C. Fleming, and K. L. Man. 2016. An exploration of usable authentication mechanisms for virtual reality systems. In *2016 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS)*. IEEE, Jeju, South Korea, 458–460. https://doi.org/10.1109/APCCAS.2016.7804002