ClickID: Multi-modal VR User Authentication based on Click-derived Hand Biometrics

ABSTRACT

While virtual Reality (VR) has attracted increasing attention in recent years, efficiently identifying a VR device user remains challenging. Current solutions require users to input passwords using handheld controllers or in-air hand gestures, which are slow, difficult, and prone to input errors. Moreover, the hand movements can be observed by others in proximity, posing security concerns. In this paper, we propose ClickID, a user-friendly VR user authentication system based on unique hand biometrics that are naturally born with user-controller interactions. To authenticate, a user simply holds the controller and clicks the trigger (an atomic interaction in VR, analogous to a finger tap on a smartphone screen). Two signal modalities, i.e., vibration and sound, are generated from click action and shaped by the user's hand, which is unique in geometry, palm size, and characteristic impedance. We thus extract hand biometric information from the controller inertial sensor and headset microphone for VR user authentication. In particular, to eliminate the influence of variable click behaviors, we design a multi-objective feature disentanglement network. Three optimization objectives are carefully designed, ensuring the disentangled features are precise, meaningful, and independent. We also propose a bifocal correlation learning framework, which harnesses the power of attention mechanisms for synergizing multi-modal information. Our experiments, conducted with two commercial VR devices, achieve a commendable balanced accuracy of over 95.87% for both hands, proving ClickID's efficacy and resilience against potential threats.

CCS CONCEPTS

• **Security and privacy** → *Authentication.*

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

KEYWORDS

Hand Biometrics, User Authentication, Virtual Reality

ACM Reference Format:
. 2026. ClickID: Multi-modal VR User Authentication based on Click-derived Hand Biometrics. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 18 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Virtual reality (VR) technique has attracted increasing attention in recent years. Shipments of VR devices are expected to reach 24.19 million by 2025, with an annual growth rate of 17.36%[1]. With the capability to deliver interactive and immersive experiences, VR has been increasingly adopted in education, entertainment, healthcare, and e-commerce [45], enabling a plethora of personalized applications. These applications are often connected with user privacy, including private app content, browsing habits, and financial records. Therefore, it is crucial to verify a VR device user before granting personal access. However, current authentication methods usually require users to input passwords using handheld controllers or perform gestures in the air. These methods can be slow, difficult, and prone to errors [43]. Additionally, although the user interacts in a virtual space, their hand movements are visible in the physical world, making it possible for attackers to guess sensitive credentials [11, 43].

To address these challenges, researchers have explored alternatives for VR user authentication. Some methods require users to perform specific tasks in the virtual space, such as watching spherical videos [42], following/throwing balls [30, 32], or manipulating a 3D Rubik's cube [28]. However, these methods identify users based on similarity of behavioral patterns, which are observable in proximity and open the door to imitation by adversaries. Recent studies have also explored physiological traits, such as visual- [34] or auditory-pupillary responses [57] to authenticate VR users. But these methods rely on eye trackers, which are unavailable on many low-end VR headsets, and collecting pupillary responses can be time-consuming due to the need for multiple stimuli. Some widely deployed biometrics, such as iris scanning [6], face recognition [20], and fingerprints, are naturally considered as candidates for VR authentication. However, they are less integrated into commodity VR devices, as their implementation necessitates either sophisticated measuring devices that increase deployment costs, or

context-breaking interactions that may undermine the user's immersive experiences. Collectively, the above-mentioned issues, including the requirement for complex tasks, time-consuming data collection, and the inevitable disruption of the immersive experience, all fundamentally compromise the crucial authentication attribute of user-friendliness.

By examining the VR usage scenarios, we find that the user-controller interaction process exhibits several favorable user-friendly attributes. Firstly, this interaction is rapid in most cases, with a simple trigger click often lasting less than one second. Secondly, this interaction is inherently part of the immersive experience, serving as a convenient and precise means for users to execute selections or confirmations in the virtual environment. In addition, handheld controllers have been widely deployed as the default configuration for most VR devices, as demonstrated in Appendix A. These attributes position the user-controller interaction as an ideal basis for establishing novel authentication mechanisms.

In this work, we present ClickID, a user-friendly VR user authentication system based on extracting hand biometrics that are naturally manifested during user-controller interactions. To authenticate, a user simply grips the controller and clicks the trigger with a finger, typically the index, as shown in Figure 1. This *single click* suffices for authentication, an efficacy underpinned by the unique Dual-modal Biometric Encoding Paradigm (DBEP). Specifically, when holding a VR controller, the user's hand naturally divides into two regions: one in contact with the controller, and the other exposed to air. Two signal modalities, i.e., structure-borne vibrations and air-borne sound, are generated from the click. They interact with the gripping hand at two complementary regions, undergoing absorption, scattering, and reflection before reaching the controller's inertial measurement unit (IMU) and the headset's microphone. As a result, the user's distinct physiological traits (e.g., hand geometry and tissue property) are physically encoded in the signal of both modalities, yielding rich information from complementary perspectives for user authentication.

Despite its promising features, the design of ClickID faces several significant challenges: *(1) Unpredictable Noise in Signal Segmentation:* In real-world scenarios, noise sources are highly unpredictable, ranging from motion artifacts to ambient noise. This renders reliable signal segmentation based on specific modality difficult. *(2) Intrinsic Entanglement of Behavioral Variability and Biometric Signature:* The hand-characterizing signals from user clicks can exhibit significant variability (e.g., duration and density) due to clicking dynamics. This behavioral variability is intrinsically entangled with the user's hand biometric signature, making reliable feature extraction more challenging. *(3) Correlation Learning from Heterogeneous Modalities:* The two signals in ClickID are intrinsically correlated, as they both originate from the same

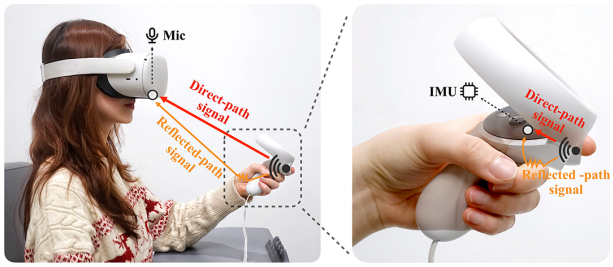


Figure 1: Authentication Scenario for ClickID.

click action and are influenced by the user's gripping hand. While this correlation provides valuable, user-specific information, effectively learning it remains challenging due to the inherent modality heterogeneity.

To mitigate the unpredictable noise effect, we develop a noise-aware event segmentation strategy. Specifically, we first apply low-pass filters to the sensor data, focusing on the frequency band of interest. We then assess the noise levels of both modalities and apply a customized three-stage localization algorithm to the clearer one, which also guides event segmentation in the other. In addition, we design a Multi-objective Feature Disentanglement Network (MFDNet) to derive behavior-irrelevant features. Three optimization objectives are carefully designed, ensuring the disentangled features are precise, meaningful, and independent of each other. After that, the covariates, such as additive noise and behavioral variability, are isolated, yielding identity-discriminative traits for further authentication. Furthermore, we propose a Bifocal Correlation Learning Network (BCLNet) for intrinsic modality dependency learning. Harnessing the expressive power of attention mechanisms, BCLNet not only distills identity-salient biometric signatures within each modality (intra-modal learning branch), but more crucially, elevates discriminative power by capturing the intrinsic correlations between the two (inter-modal learning branch). The representations derived from the two branches are then integrated into user-specific binary classifiers for authentication. In summary, our key contributions include:

- We present ClickID, a user-friendly authentication system that can be seamlessly integrated into commodity VR devices. It is the first work that shows distinct click-derived hand biometrics can be extracted using built-in VR embedding sensors.
- We develop a noise-aware segmentation strategy, which utilizes noise assessment and a three-stage algorithm for event localization. We further design MFDNet, driven by three tailored optimization objectives, enabling effective identity and covariate feature disentanglement.
- We propose a bifocal correlation learning framework for multimodal synergy authentication. It harnesses the strength of attention mechanisms to distill modality-specific cues and capture cross-modality correlations.

- We validate ClickID by conducting extensive experiments using two commercial VR devices. The results show that it achieves a commendable balanced accuracy of over 95.87% for both hands, proving its efficacy and resilience against potential threats.

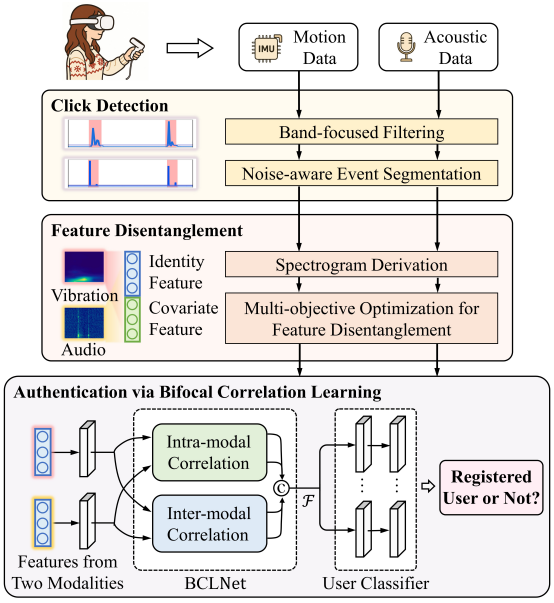


Figure 4: Workflow of ClickID.

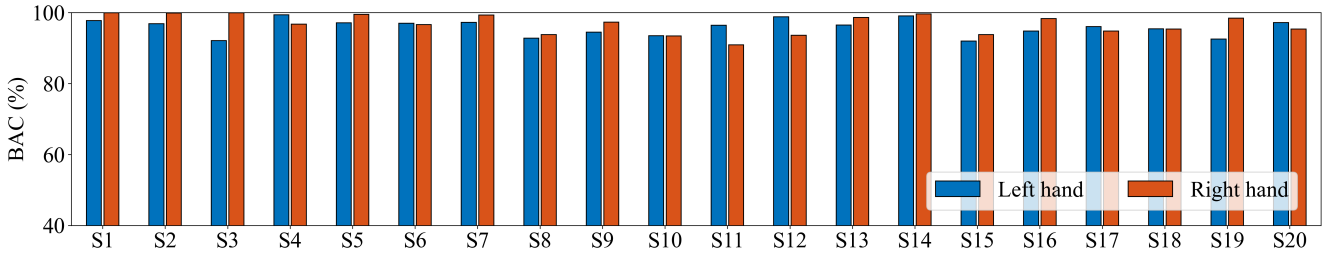
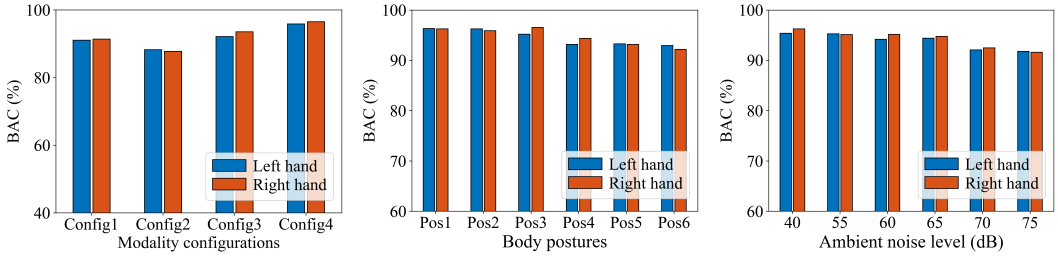


Figure 9: System performance across different subjects.



(a) (b) (c)

Figure 10: Impact of different modality configurations (a), body postures (b) and ambient noises (c).