Wave-for-Safe: Multisensor-based Mutual Authentication for Unmanned Delivery Vehicle Services

Huanqi Yang[1,2,†], Mingda Han[1,2,†], Shuyao Shi[3], Zhenyu Yan[3], Guoliang Xing[3], Jianping Wang[2], Weitao Xu[1,2*]

[1]City University of Hong Kong Shenzhen Research Institute,
[2]City University of Hong Kong, [3]The Chinese University of Hong Kong

ABSTRACT

In recent years, the deployment of unmanned vehicle delivery services has increased unprecedentedly, leading to a need for enhanced security due to the risk of leaving high-value packages to an unauthorized third party during pickup or delivery. Existing authentication methods such as QR code and one-time password are inadequate, as they are susceptible to attacks and provide only one-way authentication. This paper, for the first time to our best knowledge, proposes Wave-for-Safe (W4S) — a novel mutual authentication system that utilizes multi-modal sensors on both the user's smartphone and the unmanned vehicle. W4S uses random hand-waving by the legitimate user to achieve robust authentication by obtaining highly correlated sensory data measured by the Inertial Measurement Unit (IMU) in the smartphone and sensors in the unmanned vehicle (e.g., mmWave radar and camera). We propose several novel methods to overcome challenges such as heterogeneous data processing, asynchronization, and imitating attacks. The prototype is implemented on an unmanned vehicle and various smartphones, and evaluation in different real-world scenarios shows that W4S achieves an equal error rate below 0.013 against various attacks.

CCS CONCEPTS

• **Security and privacy** → **Authentication**; • **Networks** → **Mobile and wireless security**.

KEYWORDS

Authentication, unmanned vehicle delivery, sensor fusion

ACM Reference Format:
Huanqi Yang[1,2,†], Mingda Han[1,2,†], Shuyao Shi[3], Zhenyu Yan[3], Guoliang Xing[3], Jianping Wang[2], Weitao Xu[1,2*]. 2023. Wave-for-Safe: Multisensor-based Mutual Authentication for Unmanned Delivery Vehicle Services . In *International Symposium on Theory, Algorithmic Foundations, and Protocol Design for Mobile Networks and Mobile Computing (MobiHoc '23), October 23–26, 2023, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3565287.3610253

Table 1: Comparison with existing approaches.

Approaches	T1	T2	T3	T4
Biometric [26, 46]	✔	✖	✖	✖
QR code [18]	✔	✖	✖	✔
Password [8, 25]	✔	✖	✖	✔
Distance-bounding [28]	✖	✔	✖	✔
W4S	✔	✔	✔	✔

1 INTRODUCTION

Unmanned vehicle delivery services are believed to potentially revolutionize last-mile delivery in a more sustainable and cost-effective way. The market for unmanned delivery vehicle services is estimated to exceed $90.21 billion by 2030 [27]. In this context, many giant courier service companies (e.g., FedEx [12] and DHL [16]), large retailers (e.g., Amazon [32] and Alibaba [7]), and startup companies (e.g., Udelv [36] and Nuro [25]) have deployed mature unmanned vehicles for package delivery in the community.

The exponential growth of package deliveries using unmanned vehicles [24, 37] has raised serious concerns regarding package security. Impersonation attacks, where an attacker impersonates a legitimate user, pose the primary threat to these delivery services [23]. Four different types of impersonation attacks have been identified, depending on the stage of the delivery process, including: 1) malicious consignee during the delivery stage, where the attacker impersonates a legitimate consignee to steal the package; 2) malicious vehicle during the delivery stage, where a malicious unmanned vehicle controlled by the attacker impersonates a legitimate unmanned vehicle to deliver fake packages; 3) malicious vehicle during the pickup stage, where a malicious unmanned vehicle controlled by the attacker impersonates a legitimate unmanned vehicle to steal packages; and 4) malicious consignor during the pickup stage, where the attacker impersonates a legitimate consignor to load fake packages into the unmanned vehicle. Given the above security concerns, it is imperative to provide a secure authentication method for unmanned vehicle delivery services.

Existing authentication methods mainly adopt one-way authentication in the following aspects: 1) Biometric-based authentication methods use distinctive biometric information, such as fingerprint and face, for authentication [26, 46]. However, there are many known attacks against this kind of authentication [19, 45]. Moreover, it uploads private user biometric information to the server, which can lead to user privacy leakage. In addition, it needs to profile before use, which impairs usability. 2) QR code has also been used for the authentication of unmanned vehicle delivery such as autonomous vehicle delivery company Zoox [18]. It proposes to authenticate a user by having the vehicle scan a QR code on the user's smartphone. However, it is vulnerable to vision relay attacks. For example, a malicious unmanned vehicle tricks the user

† indicates equal contribution, * indicates corresponding author.

MobiHoc '23, October 23–26, 2023, Washington, DC, USA

ACM ISBN 978-1-4503-9926-5/23/10…$15.00
https://doi.org/10.1145/3565287.3610253

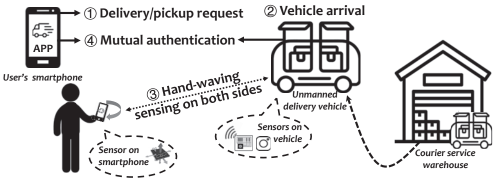

① Delivery/pickup request
④ Mutual authentication
APP
User’s smartphone
② Vehicle arrival
③ Hand-waving sensing on both sides
Unmanned delivery vehicle
Sensors on vehicle
Sensor on smartphone
Courier service warehouse

Figure 1: Application scenario of W4S. The unmanned vehicle and smartphone sense the user’s random hand-waving to obtain similar sensing results for authentication, as the key-protected channel alone is insufficient.

to obtain his QR code and forwards the QR code to the attacker to open the legitimate unmanned delivery vehicle. 3) One-time password-based authentication [8, 25] is a common authentication method, but it is also vulnerable to relay attacks [20]. The above one-way authentication techniques that solely authenticate one party’s identity are vulnerable to impersonation attacks. Furthermore, certain approaches have privacy and usability concerns. To improve security, distance-bounding protocols, a distance-based mutual authentication method, are proposed [28]. It verifies the proximity of authentic users by calculating the round-trip time between the authenticator and authenticatee through a challenge and response mechanism, thereby thwarting relay attacks. However, the effectiveness of the distance bound calculated by this protocol can be easily compromised by even the slightest processing delay, necessitating additional hardware support that is not yet widely available. Furthermore, there have been limited studies on the security issues of this protocol [9], and researchers have proposed new attacks [3] against it, rendering it unreliable. In summary, the current authentication methods have limitations including the need for additional hardware, one-way authentication, vulnerability to common attacks, and privacy leakage risks.

The aforementioned limitations motivate us to develop an authentication scheme for unmanned delivery vehicle services that satisfy four targets as illustrated in Tab. 1: T1) no need for additional hardware, T2) mutual authentication, T3) being resistant to common attacks, and T4) no user privacy leakage risks. To this end, we propose Wave-for-Safe (W4S), a multisensor-based mutual authentication method for unmanned vehicle delivery services. Fig. 1 illustrates an exemplary scenario of W4S. We use the built-in sensors in vehicles and smartphones to perform mutual authentication **without requiring extra hardware**. Specifically, we leverage the lower bound on the similarity of the sensed data **from both sides** as an additional requirement for an insufficient key-protected channel [1], which is based on the observation that the user’s random waving motion data sensed by the vehicle and the smartphone should be correlated. In addition, our system is **resistant to common attacks** due to the fact that it is difficult for imitating attackers (imitating a legitimate user’s hand-waving) as well as eavesdropping attackers to obtain accurate sensing data of a legitimate user’s wave operation. Although an attacker can obtain similar sensing data in a certain dimension, W4S can accurately detect such an attack using 3D motion data and a siamese neural network-based model. Furthermore, W4S does not need to upload personal information, indicating **no user privacy leakage risks**. In summary, the main contributions of this paper are as follows:

Table 2: Mutual authentication for unmanned delivery services.

Scheme	Scenarios	Sensors	Sensing Target
G2Auth [41]	Drone	IMU, Camera	1D hand-waving
Smile2Auth [33]	Drone	Camera	Facial expression
H2Auth [42]	Drone	Microphone	Drone noises
W4S	Unmanned vehicle	mmWave radar, IMU, Camera	3D hand-waving

- We propose W4S, the first mutual authentication system for unmanned vehicle delivery services, which requires no additional hardware, is mutual authentication, resistant to relay and imitation attacks, and has no user privacy leakage risks.
- We develop a series of advanced signal processing methods to enable accurate 3D acceleration extraction of the user’s hand-waving movements from the multi-modal sensors (i.e., IMU, mmWave radar, and camera) to facilitate robust authentication.
- We propose a spatial-temporal synchronization approach to fuse multisensor data from the unmanned vehicle and the smartphone, and design a novel two-stage siamese neural network to discriminate between legitimate users and malicious attackers.
- We implement the W4S prototype on an unmanned vehicle and various types of smartphones. Real-world experiments show that W4S can authenticate a user with an average equal error rate below 0.013. Security analysis is also conducted to show that W4S is resistant to common attacks.

2 RELATED WORK

Mutual authentication for unmanned delivery services. Tab. 2 categorizes the authentication methods for unmanned delivery services. G2Auth [41] is a mutual authentication system for drone delivery services, which utilizes 1D acceleration obtained from the drone’s camera and the user’s smartphone IMU for authentication. Smile2Auth [33] is another drone-user authentication method that captures video of the user’s expression via the drone’s camera and the front camera of the user’s smartphone, respectively. Meanwhile, H2Auth [42] performs mutual authentication between the user and the drone using unique drone noise characteristics without requiring any sound fingerprint. Though there are many schemes for drone-user authentication, there is no solution that satisfies the four targets mentioned in Sec. 1 for unmanned vehicle delivery services. This paper, for the first time, presents a mutual authentication scheme explicitly designed for unmanned vehicle delivery services. W4S builds upon the concept of the above authentication methods for drone delivery [33, 41, 42], utilizing randomness derived from different sensing targets to establish authenticity. However, W4S further enhances security by leveraging multi-model sensors on unmanned vehicles to extract 3D hand-waving. Specifically, the commonly equipped mmWave radar on unmanned vehicles is utilized in our study as a side channel for mutual authentication. The mmWave side channel provides depth information regarding the user’s hand movement, which can lead to enhanced mutual authentication performance when compared to previous studies that relied solely on camera [41] as will be introduced in Sec. 8.2.

Similarity measurement-based authentication. W4S can be categorized as similarity measurement-based authentication, which uses the similarity measurement of different devices to the same target (e.g., human’s movement pattern) for authentication. P2Auth [22]

[1] A key-protected communication channel is studied in prior existing authentication methods for unmanned vehicle delivery [2]. Due to radio relay attacks [14, 40], a key-protected channel alone is insufficient for authentication.

uses similar timestamps measured by clocks on different IoT devices for the same event for authentication. EchoKey [17] uses the fine-grained similar spatial context of the receiving device carried by the ambient sound signal for authentication. Different from these works, W4S, the first multisensor-based mutual authentication scheme, is proposed for unmanned vehicle delivery services.

Biometric-based authentication. Many authentication schemes are based on users' physiological information (e.g., face [46], fingerprint [26]) or behavioral characteristics (e.g., gait [13]), which raise privacy concerns and require the users' bio-features to be registered into the database in advance. Moreover, users' behavior characteristics may change over time. Unlike these solutions, W4S eliminates the requirement of registration and uploading private data by only using the user's real-time hand-waving information.

3 FEASIBILITY STUDY AND CHALLENGES

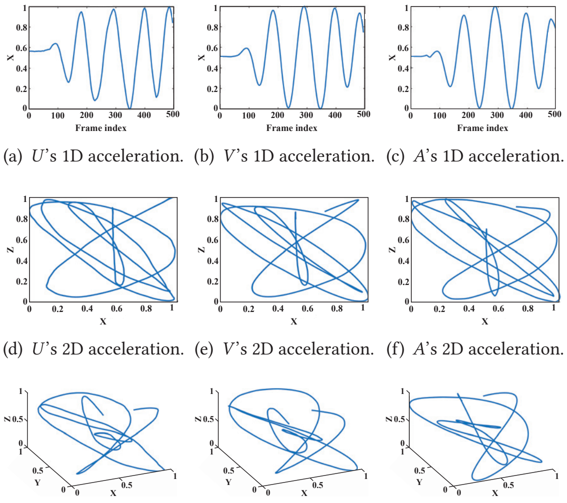
3.1 Design Choices

Sensor choice. Our goal is to achieve mutual authentication without introducing additional hardware costs as described in Sec. 1. Although the existing unmanned vehicles are equipped with various sensors, we find that most commercial unmanned delivery vehicles are equipped with millimeter-wave (mmWave) radars and cameras for cost reasons [8, 12, 25, 35, 36]. Inertial Measurement Unit (IMU) is a common sensor in modern smartphones, which can be used to obtain acceleration information when the user waves the smartphone. Therefore, we choose the camera, mmWave radar of the vehicle, and IMU of the smartphone for authentication.

Metric choice. Since the information obtained from these three sensors needs to be fused and compared, we have to choose a unified metric for them. However, fine-grained trajectory and velocity inferences based on IMU sensory data from smartphones remain an unsolved problem [34] due to noisy acceleration data derived from imperfect hardware. Moreover, the integration of such noisy acceleration leads to noise accumulation, thus making the inferred velocity and trajectory information inaccurate. Therefore, we choose acceleration as the unified metric for authentication.

3.2 Feasibility Analysis

Based on the above analysis, we can utilize the mmWave radar and camera on the unmanned vehicle and the IMU on the smartphone to sense the acceleration of the user's hand-waving for mutual authentication. A recent work shares a similar idea, but it uses the on-drone camera as well as the IMU in the user's smartphone to acquire 1D acceleration of the user's hand-waving for human-drone authentication [41]. However, based on our preliminary study, the generated 1D acceleration sequences have low complexity and can be easily imitated by attackers. In our experiments, the legal user (U) waves the smartphone (S) towards the unmanned vehicle (V), the camera and mmWave radar on the unmanned vehicle side and the smartphone on the user side recorded acceleration information separately. Additionally, another skilled imitating attacker (A) with a smartphone (S') mimics the hand-waving of the user and attempts to obtain the same acceleration sequence to fool the system.



(a) U's 1D acceleration. (b) V's 1D acceleration. (c) A's 1D acceleration.

(d) U's 2D acceleration. (e) V's 2D acceleration. (f) A's 2D acceleration.

(g) U's 3D acceleration. (h) V's 3D acceleration. (i) A's 3D acceleration.

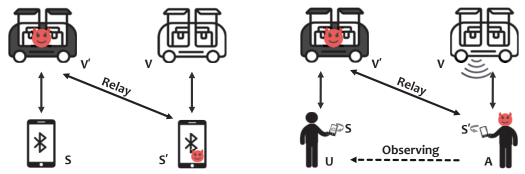
Figure 2: 1/2D acceleration vs. 3D acceleration. When using 1D acceleration, the correlation coefficients between V-U and A-U are 0.9881 vs. 0.9770. When using 2D acceleration, the correlation coefficients between V-U and A-U are 0.9826 vs. 0.9756. When using 3D acceleration, the correlation coefficients between V-U and A-U are 0.9804 vs. 0.8310. The conclusions are: 1) the correlation of A-U is always lower than that of V-U, and 2) using 3D acceleration is more secure than using 1/2D acceleration.

Correlation Analysis. The normalized acceleration we obtained is shown in Fig. 2. Although A can easily access almost the same 1D and 2D acceleration sequence (the 1D and 2D correlation coefficients between A and U are 0.9770 and 0.9756, respectively), the complete 3D acceleration sequence is difficult to access (the 3D correlation coefficient between A and U decreases to 0.8310), which is determined by the A's point of view, because no matter where A is standing, he always fails to get the complete and accurate 3D acceleration information due to the limitations of observation angle. However, V can obtain complete and accurate 3D acceleration information (the correlation coefficient between V and U is 0.9804 when using 3D acceleration) by combining multiple sensors.

Complexity Analysis. Afterward, we calculate the complexity for different dimensions of acceleration information in our dataset as will be introduced in Sec. 7, which contains 1,200 3D acceleration sequences. The complexity is measured by sample entropy, which reflects the complexity of time-series sequences. The average sample entropy of 1D, 2D, and 3D time-acceleration sequences are 0.6225, 0.6694, and 1.4509, respectively, indicating that the 3D information is more complex to ensure security. These preliminary experiments illustrate both the limitations of using only 1D or 2D acceleration and the advantages of using 3D acceleration. Therefore, we choose to use 3D hand-waving acceleration for authentication.

However, we need to address several non-trivial challenges.

- *Challenge 1:* W4S relies on the accurate sensing of the user's hand-waving movements by three different sensors (i.e., mmWave radar, camera, and IMU). The raw data captured by different sensors contains a lot of interference (e.g., dynamic/static environmental noise and movement noise generated by the user's body). Each sensor exhibits diverse resilience to different types of interference, and how to design the corresponding signal processing algorithms to get accurate 3D acceleration according to the characteristics of different sensors is the first challenge.
- *Challenge 2:* Since the unmanned delivery vehicle and the user's smartphone do not share the same clock signal, how to design an accurate time synchronization scheme is a prerequisite for accurate sensory data matching. In addition, the 3D acceleration



(a) Relay attack. (b) Imitating attack.

Figure 3: Threat model. (a) In a relay attack, V' and S' fool V and S by relaying the Bluetooth signal. (b) In an imitating attack, V' induces U to start hand-waving, while A imitates U's movements in front of V to make V obtain the same acceleration as U. All communications between V and S are relayed through V' and S'.

obtained from the vehicle side and the user's smartphone side are not under a unified spatial coordinate system. Therefore, a temporal-spatial synchronization method is required in W4S.

- *Challenge 3:* As aforementioned, even if we use the 3D acceleration for authentication, the correlation coefficient between the attacker (A) and legitimate user (U) is still higher than 0.8, indicating that we cannot use the linear correlation coefficient as a single discriminant for authentication. Therefore, a fine-grained discrimination method is needed to obtain promising accuracy.

4 SYSTEM MODEL

We consider a mutual authentication system involving two entities, the user's smartphone (S) and the unmanned vehicle (V). In this scheme, both S and V serve as the authenticator and authenticatee. Both parties capture the user's hand-waving, process the 3D acceleration information locally, and exchange it via a pre-established key-protected channel. Authentication is deemed successful only when both S and V have independently passed the authentication.

4.1 Threat Model

Relay attacks. The relay attack can easily break the key-protected channel-based authentication system on unmanned vehicles [14, 40], which is also applicable to unmanned delivery scenarios. For example, as shown in Fig. 3(a), given a key-protected Bluetooth channel, without knowing the key, the attacker's smartphone (S') and malicious unmanned vehicle (V') can simply relay the Bluetooth signals between the user's smartphone (S) and the unmanned vehicle (V), such that both V and S can be fooled to believe the proximity and conduct the authentication even if V and S are far away from each other. Our threat model assumes attackers have the ability to launch relay attacks, such that an attacker can use a malicious unmanned vehicle to fool a victim user to start the authentication procedure and relay the encrypted traffic.

Imitating attacks. As shown in Fig. 3(b), after launching relay attacks, an adaptive attacker (A) familiar with W4S can replicate a user's hand-waving movements. By observing the user's hand-waving trajectory, the attacker can mimic similar hand-waving movements in front of V, while V' prompts the legitimate user (U) to hand-wave. The attacker's goal is to induce V to obtain the same acceleration as U. It should be noted that all encrypted traffic between V and S is relayed by V' and S'. W4S can handle such attacks, as detailed in Sec. 8.2.

Attacks out of scope. The attacker may employ computer vision-based methods along with a robotic arm to imitate the user's hand-waving, which is known as the robotic arm-based imitation attack. However, such attacks are hard to implement due to the significant cost [29] and vulnerability to time delays [6], thereby making them an impractical avenue for our current study.

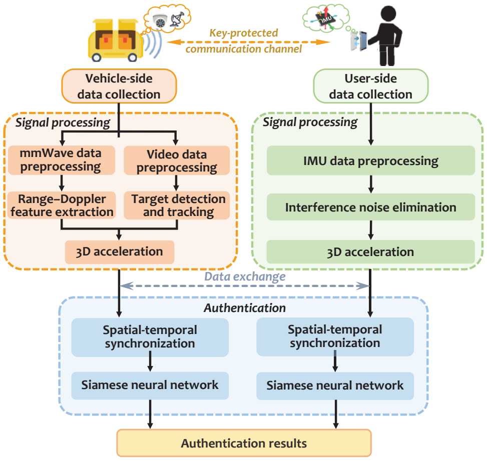


Figure 4: Approach overview.

4.2 System Overview

Assumptions. We assume a pre-established key-protected communication channel (e.g., Bluetooth and Wi-Fi) between V and S for data transmission. To enable authentication, V is equipped with at least a pair of mmWave radar and camera, which are synchronized in the time domain and transformed to the same coordinate system using known calibration parameters. Calibration can be achieved using existing methods in the literature [5]. We assume the data collected by built-in sensors are trustworthy [15] and processed in the local trusted execution environment [30]. However, the current communication method is vulnerable to relay attacks, as previously discussed. Assuming that other people or objects do not obstruct the line-of-sight between the user and vehicle during authentication.

Approach overview. As illustrated in Fig. 4, W4S is composed of three primary components: vehicle-side signal processing, user-side signal processing, and authentication. The unmanned vehicle and the user's smartphone utilize their respective sensors, including mmWave radar and camera in the unmanned vehicle, and an IMU in the smartphone, to acquire acceleration information when the user waves the smartphone. Through the spatial-temporal synchronization and similarity detection of obtained 3D acceleration information, mutual authentication between the user and the vehicle is achieved. The procedure of W4S is as follows:

1. U places a delivery/pickup order using the delivery app on S. After V arrives at the designated location, S establishes a key-protected communication channel with V. V sends a start notification to S and opens built-in mmWave radar and camera.
2. Upon receiving the start notification, S generates a vibration to notify U to start hand-waving movements. After S collects data of time duration t (t is studied as a parameter) through the built-in IMU, it generates another vibration to notify U to stop waving and sends a stop notification to V.
3. The hand-waving data sensed by S and V are exchanged after local processing, and the authentication decision is made independently. If the authentication is successful on both sides, the package delivery is executed; otherwise, it goes back to step 2 until the maximum number of attempts is reached.

5 SIGNAL PROCESSING

5.1 Vehicle-side Data Processing

5.1.1 ***mmWave Data Processing.*** The mmWave radar mounted on the unmanned delivery vehicle is leveraged to obtain the 1D radial acceleration information when the user waves the smartphone.

- *Range-Doppler Maps generation.* The mmWave radar transmits the frequency modulated continuous wave (FMCW) signal, which is called chirp. The frequency of the chirp signal increases linearly with time and can be expressed as $f = f_0 + St$, where f_0 is the start frequency and S is the frequency modulation slope. Suppose the amplitude of the transmitted signal at time t is A_T, then the transmitted sinusoidal FMCW signal $T(t)$ can be expressed as

$$T(t) = A_T \cos\left[2\pi\left(f_0 t + St^2/2\right)\right]. \tag{1}$$

When the transmit signal encounters an obstacle (e.g., the legitimate user) at distance d, the radar will receive a delayed version of the transmitted signal $R(t)$, which can be expressed as

$$R(t) = A_R \cos\left[2\pi\left(f_0\left(t-\tau\right) + S(t-\tau)^2/2\right)\right], \tag{2}$$

where A_R is the amplitude of the received signal, $\tau = 2d/c$ is the time delay, and c is the speed of light. Finally, the transmitted signal $T(t)$ is mixed with the received signal $R(t)$, and a low-pass filter is used to filter out the sum frequency components to obtain the intermediate frequency (IF) signal:

$$Y(t) = LPF\{T(t)\cdot R(t)\} = A_{IF}\cos\left(2\pi f_{IF} t + \phi_{IF}\right), \tag{3}$$

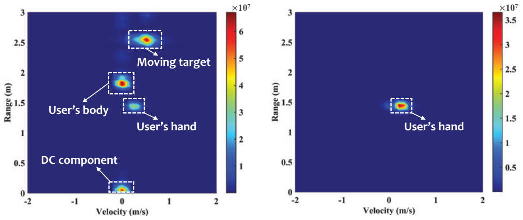
where A_{IF} is the amplitude of the IF signal, $f_{IF} = 2Sd/c$ is known as the beat frequency, and ϕ_{IF} is the phase. The IF signal is initially subject to range dimension (fast time) Fast Fourier Transform (FFT) to obtain target range information. Subsequently, a second FFT is conducted in the Doppler dimension (slow time) to obtain velocity information. Using the combined range and velocity information, we generate the Range-Doppler Map (RDM) for each frame, as illustrated in Fig. 5(a).

- *Noise reduction.* Besides hand-waving information, raw RDM contains static noises (e.g., user's body and DC component) and dynamic noises (e.g., nearby moving objects and pedestrians), which are shown in Fig. 5(a). Therefore, we need to denoise the raw RDMs and keep the user's hand-waving information only. We use the mean value of each RDM along the slow time dimension to remove the static noises. During the authentication process, there may also be dynamic noises caused by pedestrians or moving objects in the RDM, which can be eliminated by the following method. As described in Sec. 4.2, since we assume that no other person or objects pass between the user and the unmanned vehicle during the authentication process, the user is the closest object to the vehicle. Therefore, for each RDM frame, we only keep the closest object, which reflects the range and velocity information of the user's hand-waving movements. The denoised RDM is shown in Fig. 5(b).

- *Acceleration acquisition.* To obtain radial movement information of the smartphone, we use the following equation to transform the RDM of all frames into a 2D time-velocity feature map:

$$V_{(n,i)} = \frac{\sum_{j=1}^{N_R}\left[RDM_{(n,i,j)}\cdot R_j\right]}{N_R}, \quad i\in[1,N_D], j\in[1,N_R], \tag{4}$$

where N_R is the number of Range FFT, N_D is the number of Doppler FFT, R_j is the range bin index, and $RDM_{(n,i,j)}$ represents the value corresponding to Doppler bin i and range bin j in the nth RDM frame. We show the correspondence between the waving smartphone process (phase *a*-*e*) and the time-velocity feature map in Fig. 6. Due to the static noise elimination algorithm described above, the obtained time-velocity feature map contains missing values near the zero velocity component. Meanwhile, there are other anomalies in the obtained feature map. A K nearest neighbor (KNN) smoothing filter is used to filter out the anomalies, and the image dilation operation is used to connect discontinuous points. Finally, the 1D radial acceleration of the smartphone is obtained by taking the derivative of the velocity.



(a) Raw RDM. (b) Denoised RDM.

Figure 5: Noise reduction.

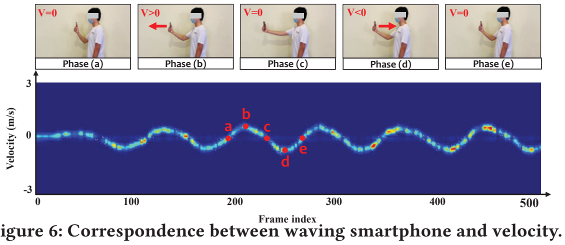
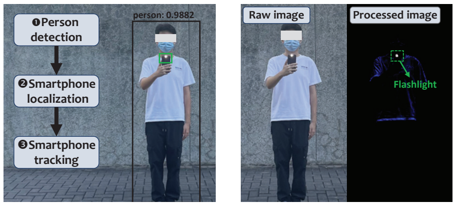


Figure 6: Correspondence between waving smartphone and velocity.

5.1.2 ***Video Data Processing.*** In order to obtain the 2D acceleration information when the user waves the smartphone, we first locate the smartphone and track its trajectory. However, directly locating and tracking the user's smartphone is difficult due to varying light conditions and other environmental distractions. Therefore, we design a three-step smartphone acceleration acquisition scheme from the video as shown in Fig. 7(a).

- *Person detection.* In the first step, we use YOLO-FastestV2 [10], which is a lightweight, fast, and easy-to-deploy objection detection method, to detect the user's body. Although there may be multiple people in the camera frame, we assume that the user is dominant (i.e., the user with the largest bounding box in the frame).

- *Smartphone localization.* Direct target detection of the user's smartphone is difficult because smartphones have different sizes and colors, and the area covered by the user's hand is also varying. We locate the flashlight on the user's smartphone to get the smartphone location even at night [41]. Since the brightness of the flashlight part is significantly different from the other parts in the image, we adjust the contrast limits of the video frame to highlight the flashlight part as shown in Fig. 7(b). Meanwhile, due to the interference from other static light sources such as street lights in the surrounding environment, we use a simple frame difference method to eliminate interference. Then, a contour detection method [1] is used to search the flashlight point in the first frame, and this point is marked as the initial tracking position. Person detection is necessary during the day due to reflective surfaces, using the obtained bounding box to reduce the search range. While at night,



(a) Workflow. (b) Flashlight localization.

Figure 7: Video data processing.

the flashlight's high brightness distinguishes the smartphone from the surroundings, eliminating the need for person detection.

- *Smartphone tracking.* Once we get the flashlight (smartphone) location, we use SiameRPN++ [21] to track the flashlight. The output of target tracking is the center point coordinates of the bounding box. Then, a Savitzky-Golay filter is used, which can smooth the raw trajectory data while retaining the variation information more effectively. Finally, we obtain the final 2D acceleration by taking the second derivative of the smoothed trajectory.

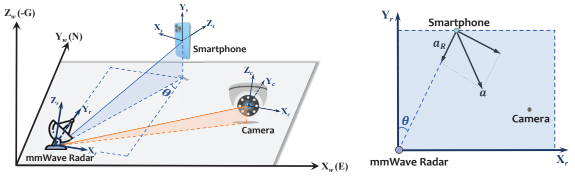
5.2 User-side Data Processing

The 3D acceleration on the user side is obtained directly from the smartphone's IMU. First, we preprocess the raw acceleration data to remove the effect of gravity [38]. In addition, there are other interfering noises such as slight hand tremors when the user waves the smartphone, which affect the raw acceleration information sensed by the IMU. To solve this problem, we use independent component analysis (ICA), a blind source separation method, to separate the waving component of the acceleration from other interference components. Assume that the acceleration measured by the IMU is $A(t)$, which is a mixture of hand-waving and hand trembling. Then, the ICA model of our problem can be expressed as $A(t) = A \cdot S(t)$, where A is the mixing matrix and $S(t)$ represents the independent sources. The FastICA algorithm is leveraged to obtain the unmixing matrix $W = A^{-1}$ and the source signals can be estimated by $S(t) = W \cdot A(t)$. Finally, the denoised 3D acceleration can be obtained by selecting the independent component with the lowest dominant frequency [43].

6 AUTHENTICATION

6.1 Spatial and Temporal Synchronization

W4S relies on the similarity of the 3D acceleration sequences obtained from the user's smartphone-side and the unmanned vehicle-side. However, the unmanned vehicle and smartphone do not share the same clock and coordinate system, which leads to a temporal and spatial asynchronization of the obtained 3D acceleration sequences. This asynchronization can seriously affect the similarity of the sequences and thus the authentication results. To solve this problem, we design spatial and temporal synchronization methods to synchronize the raw 3D acceleration sequences.



(a) Different coordinate systems in W4S. (b) Top view in $\mathcal{R}$.

Figure 8: Spatial synchronization.

6.1.1 *Spatial Synchronization*

Since the 3D acceleration sequences obtained from the smartphone side and the unmanned vehicle side are not in the same coordinate system, we need to spatially synchronize them. As shown in Fig. 8(a), the world coordinate system $\mathcal{W}$ is defined by east, north, and the reverse gravity direction (X_w, Y_w, Z_w), the smartphone coordinate system $\mathcal{S}$ is (X_s, Y_s, Z_s), the camera coordinate system $\mathcal{C}$ is (X_c, Y_c, Z_c), and the mmWave radar coordinate system $\mathcal{R}$ is (X_r, Y_r, Z_r). The vehicle coordinate system $\mathcal{V}$ is (X_v, Y_v, Z_v), which is not shown in the figure. We define $\mathcal{W}$ as the reference coordinate system and explain how to transform the 3D acceleration sequences from the vehicle side and the smartphone side to the reference coordinate system.

- *Vehicle-side spatial synchronization.* Since mmWave radar only obtains radial acceleration a_R, we need to transform a_R to $\mathcal{R}$. As shown in Fig. 8(b), the transformation can be done by using a_R multiplied by $\cos\theta$ and $\sin\theta$, where θ is the azimuth. To obtain the azimuth θ, we design a method based on the MUSIC algorithm [31], which is a high-resolution direction-finding algorithm for multiple antenna systems. The covariance matrix R_x of the received signals X from K antennas of mmWave radar is first calculated as follows:

$$R_x = \frac{1}{K} \sum_{k=1}^{K} X(i) X^H(i), \tag{5}$$

where X^H is the conjugate transpose of X. Then, the eigendecomposition can be represented as

$$R_x = \begin{bmatrix} U_s & U_n \end{bmatrix} \begin{bmatrix} \Lambda_s & \\ & \Lambda_n \end{bmatrix} \begin{bmatrix} U_s \\ U_n \end{bmatrix}, \tag{6}$$

where U_s and U_n are the signal space and noise space, Λ_s and Λ_n are diagonal matrices consisting of eigenvalues in signal space and noise space, respectively. The spatial spectrum can be expressed as

$$P(\theta) = \frac{1}{\mathbf{a}^H(\theta) U_n U_n^H \mathbf{a}(\theta)}, \tag{7}$$

where $\mathbf{a}(\theta)$ is the steering vector. Finally, θ can be obtained by spatial spectrum peak search. Therefore, the coordinates of radial acceleration a_R in $\mathcal{R}$ can be expressed as $(a_R \sin\theta, a_R \cos\theta, 0)$.

Since the relative positions of the mmWave radar and the camera are fixed, the rotation matrices $\mathbf{R}_\mathbf{r}^\mathbf{c}$ from $\mathcal{R}$ to $\mathcal{C}$ and $\mathbf{R}_\mathbf{c}^\mathbf{v}$ from $\mathcal{C}$ to $\mathcal{V}$ are assumed to be known (can be calculated by calibration method [5]). Therefore, the acceleration obtained by the mmWave radar in $\mathcal{C}$ can be expressed as $[a_{xc}^r, a_{yc}^r, a_{zc}^r]^\top = \mathbf{R}_\mathbf{r}^\mathbf{c}[a_R \sin\theta, a_R \cos\theta, 0]^\top$. Combining the 1D acceleration component a_{yc}^r along Y_c with the 2D acceleration (a_{xc}^c, a_{zc}^c) obtained by the camera, we can obtain the 3D acceleration $(a_{xc}^c, a_{yc}^r, a_{zc}^c)$ in $\mathcal{C}$. Finally, the vehicle-side 3D acceleration in the reference coordinate system can be expressed as $[a_{xw}^v, a_{yw}^v, a_{zw}^v]^\top = \mathbf{R}_\mathbf{v}^\mathbf{w}\mathbf{R}_\mathbf{c}^\mathbf{v}[a_{xc}^c, a_{yc}^r, a_{zc}^c]^\top$, where $\mathbf{R}_\mathbf{v}^\mathbf{w}$ is the transformation matrix from $\mathcal{V}$ to $\mathcal{W}$, which can be obtained by built-in GPS on unmanned vehicle [5, 39].

- *Smartphone-side spatial synchronization.* The acceleration of the smartphone's IMU can be directly transformed to the reference coordinate system by multiplying a transformation matrix, which can be expressed as $[a_{xw}^s, a_{yw}^s, a_{zw}^s]^\top = \mathbf{R}_\mathbf{s}^\mathbf{w}[a_{xs}^s, a_{ys}^s, a_{zs}^s]^\top$, where a_{xs}^s, a_{ys}^s, and a_{zs}^s are the three acceleration components obtained by the smartphone's IMU. $\mathbf{R}_\mathbf{s}^\mathbf{w}$ is the transformation matrix from $\mathcal{S}$ to $\mathcal{W}$ and can be obtained by the smartphone API.

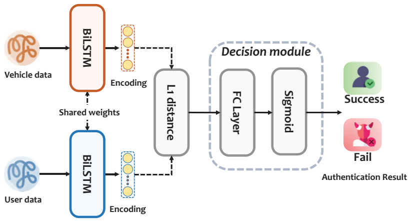


Figure 9: Siamese neural network.

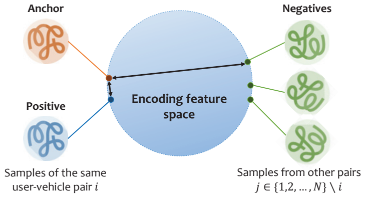

Anchor
Negatives
Positive
Encoding feature space
Samples of the same user-vehicle pair i
Samples from other pairs $j \in \{1,2,\dots,N\} \setminus i$

Figure 10: Triplet training.

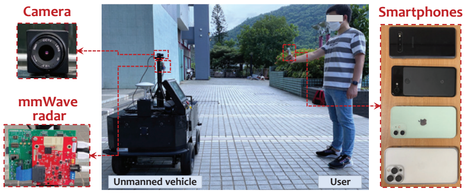


Figure 11: Hardware devices and data collection.

6.1.2 ***Temporal Synchronization***. Since the user's smartphone and the unmanned vehicle do not share the same clock, the data from the two sides need to be temporally synchronized. We first unify the sampling rate of the acceleration obtained from the three sensors by downsampling to 60 Hz. Then a two-step time synchronization method is proposed in W4S as follows.

- *Coarse-grained synchronization.* As introduced in Sec. 4.2, the unmanned vehicle sends the start notification to the user's smartphone, and the smartphone generates a vibration to notify the user to start authentication, but this temporal synchronization can be inaccurate due to the user response time, hardware response time, and transmission delay.
- *Fine-grained synchronization.* Since W4S is based on the acceleration information when the user waves the smartphone, and the obtained acceleration contains multiple extreme points (peaks and valleys), we use extreme points of the axis with the highest variance detected on the user's smartphone side and the vehicle side for fine-grained temporal synchronization. Specifically, we use the time stamps corresponding to the first three extreme points as the key points for synchronization. The origin time point after temporal synchronization can be expressed as

$$T_m = \frac{1}{3}(t_{m,1} + t_{m,2} + t_{m,3}), \quad m \in \{smartphone, vehicle\}, \tag{8}$$

where T_m is the original time point of the acceleration after temporal synchronization at the m-side, $t_{m,i}$ ($i \in \{1,2,3\}$) are the key points for fine-grained synchronization at the m-side.

6.2 Decision Making

The success of authentication is assessed by calculating the similarity between the 3D acceleration data obtained from the user and the unmanned vehicle. However, the correlation coefficient between the user and attacker still remains high even when using 3D acceleration, making traditional correlation threshold-based methods unsuitable. To address this, we propose a novel siamese neural network (SNN) based framework for authentication. SNNs have shown promise in calculating similarity [11, 44], and can learn the input data representation based on the similarity metric used.

As shown in Fig. 9, we design a dual-input SNN framework for mutual authentication in unmanned vehicle delivery. The SNN calculates the similarity between the 3D acceleration sequences obtained from the smartphone and the vehicle. The SNN comprises two identical sub-networks, which process each input and work as encoders. Each sub-network within the model is composed of two layers of Bidirectional Long Short-Term Memory (BiLSTM). BiLSTM is particularly advantageous for processing time-series sequences, as it is capable of capturing both past and future contexts effectively. The first BiLSTM layer has 256 units and returns sequences to allow the second BiLSTM layer to receive sequence input. The second BiLSTM layer has 128 units and does not return sequences, thus its output could serve as the encoding of the input signal. The similarity between the two encodings is obtained using the Euler distance. Specifically, the absolute difference between two encodings is calculated element-wise. This distance vector is then fed into a fully connected layer with a single unit and a Sigmoid activation function, which outputs the final authentication result.

Training. The training procedure is performed in two steps: 1) shared layers training and 2) decision module training. To train the shared layers, we use a semi-hard triplet configuration with a triplet loss, which is ideal for training samples with small variability. The loss function can be represented as

$$Loss = \max(d(a,p) - d(a,n) + \text{margin}, 0), \tag{9}$$

where $d(\cdot)$ is ℓ_1 distance, margin is the hyperparameter in semi-hard triple loss, a, p, and n represent anchor, positive and negative, respectively. As shown in Fig. 10, under semi-hard triplet configuration, we have $d(a,p) < d(a,n) < d(a,p) + \text{margin}$, meaning that the distance between the negative samples and the baseline samples (anchor) is greater than the distance between the positive samples and the baseline samples (anchor), but the triplet loss value has not yet reached zero when the network can continuously reduce the loss value through proper learning. Once the shared layers in SNN are trained, their weights are frozen and the fully-connected decision layer is attached to the shared layers. Finally, we train the decision layers using binary cross-entropy as the loss function.

7 DATA COLLECTION

Devices. Fig. 11 shows the devices used in our experiments, including an Apollo unmanned vehicle (3.6 GHz Intel Core i9-9900K CPU and the Ubuntu 16.04 OS) and four smartphones. As shown in the figure, we deploy a camera and a mmWave radar (TI AWR1642) for Apollo unmanned vehicle to capture users' hand-waving movements. The default resolution of the camera is set to 720P at 30 FPS. The default frame rate of the mmWave radar in the vehicle is set to 50 FPS with 255 chirp loops in each frame. We use four types of smartphones (S1-S4) in the data collection: iPhone 13 Pro Max (3.23 GHz CPU, 6 GB RAM, iOS 16), iPhone 12 (3.1 GHz CPU, 4 GB RAM, iOS 15), Samsung S10 (2.84 GHz CPU, 8 GB RAM, Android 11), and Nexus 6P (1.95 GHz CPU, 3 GB RAM, Android 9).

Dataset. The dataset used to evaluate the performance of W4S consists of 40 subjects (24 males and 16 females) [2]. The dataset was collected from participants randomly selected in various public locations, such as college entrances, apartment entrances, and mall entrances. Participants were instructed to hold a smartphone and press the "start" button on the screen to begin waving the device until it vibrated. Each hand-waving motion lasted at least 5 s, and each

[2] Ethical approval has been granted by the corresponding organization.

participant repeated the motion 30 times. There were no specific requirements for how participants should wave the smartphone, allowing for a diverse range of natural waving patterns. This makes the hand-waving dataset highly realistic, reflecting a variety of people and environments. As we use the triplet strategy for training, below we discuss how to generate the anchor, positive, and negative samples in the training procedure. We collect acceleration data from the smartphone (anchor sample) and the unmanned vehicle (positive sample) when a participant performs hand-waving in front of the vehicle, resulting in 40 subjects × 30 repetitions = 1, 200 anchor-positive pairs. For each anchor-positive pair, we randomly select one positive from another subject as the negative sample. Finally, we have 1,200 triplet sample pairs, each consisting of an anchor sample, a positive sample, and a negative sample.

8 EVALUATION

System Implementation The implementation details of W4S are described as follows. The SNN model is implemented based on TensorFlow Lite and TensorFlow frameworks on the smartphone and the vehicle, respectively. The model is first trained offline on a desktop PC with Intel i7-10700 CPU, 64 GB RAM, and RTX 3080 GPU, then deployed on the smartphone and unmanned vehicle.

Metrics: We adopt True Acceptance Rate (TAR), False Rejection Rate (FRR), and False Acceptance Rate (FAR) as the performance metrics for W4S. A lower FRR implies higher usability for authorized users, while a lower FAR indicates better protection against unauthorized access. Additionally, we report the receiver operating characteristic (ROC) curve's area under the curve (AUC) and the equal error rate (EER). AUC provides an overall performance measure, and EER is the value of FRR or FAR when FRR equals FAR. A lower EER indicates superior system performance.

8.1 Overall Performance

We use the collected dataset to evaluate the performance of W4S. We adopt the leave-one-subject-out cross-validation mechanism to obtain the average performance over all subjects. We iteratively choose one subject for testing and use the data of the 39 subjects left to train the model. We report the average performance of all subjects. Thus, we can check whether our system is user-independent, i.e. whether it works for users that we have never seen during training. Fig. 12 shows the ROC curve. We can see that W4S achieves an average EER of 0.0126 and AUC of 0.9987. The low EER indicates that W4S can distinguish authorized accesses from unauthorized ones with high accuracy (i.e., 1-EER) of 0.9874.

8.2 Security Analysis

Against imitating attacks. As discussed in Sec. 4, based on relay attacks, the attacker who knows how W4S works is able to observe the behavior of the legitimate user. Then the attacker tries to imitate the user's waving trajectory to fool the legitimate unmanned vehicle into obtaining similar sensor measurements with the user, as introduced in Sec. 4.1. In this experiment, we recruit 10 participants as victims and another 10 as imitating attackers. Each pair of attacker and victim performs the authentication operations on each smartphone for 15 times. We consider two types of imitating attackers, untrained and trained imitating attackers. Among them, the untrained attackers are told to imitate the victim without previous knowledge, and the trained attackers are provided with a video of the victim's authentication process for multiple practice sessions to improve the imitating ability. Finally, a total of 600 trials are conducted by untrained and trained attackers. As shown in Fig. 13, W4S can successfully defend against untrained and trained imitating attackers with 99.67% and 99.17% detection rates. However, when 2D or 1D information is used by our system, the detection rate drops significantly. Specifically, when using 2D information, the detection rate of trained and untrained attackers decreased by 6.22% and 8.63%, respectively; when using 1D information, the detection rate of trained and untrained attackers decreased by 16.79% and 27.32%, respectively. This is because the 3D information is hard for the attacker to observe as discussed in Sec. 3 and the human reaction time is constantly varying [4], both of which impair the 3D acceleration values observed by the attacker in the temporal and spatial domain. Although the attacker can obtain partially correlated 3D acceleration, our designed decision-making SNN with fine-grained feature extraction capability can still distinguish the attacker due to the robust decision-making model.

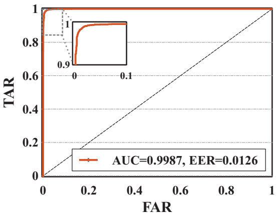


Figure 12: Overall performance.

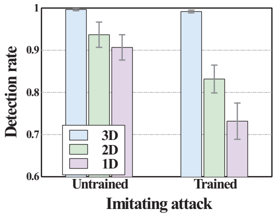


Figure 13: Security analysis.

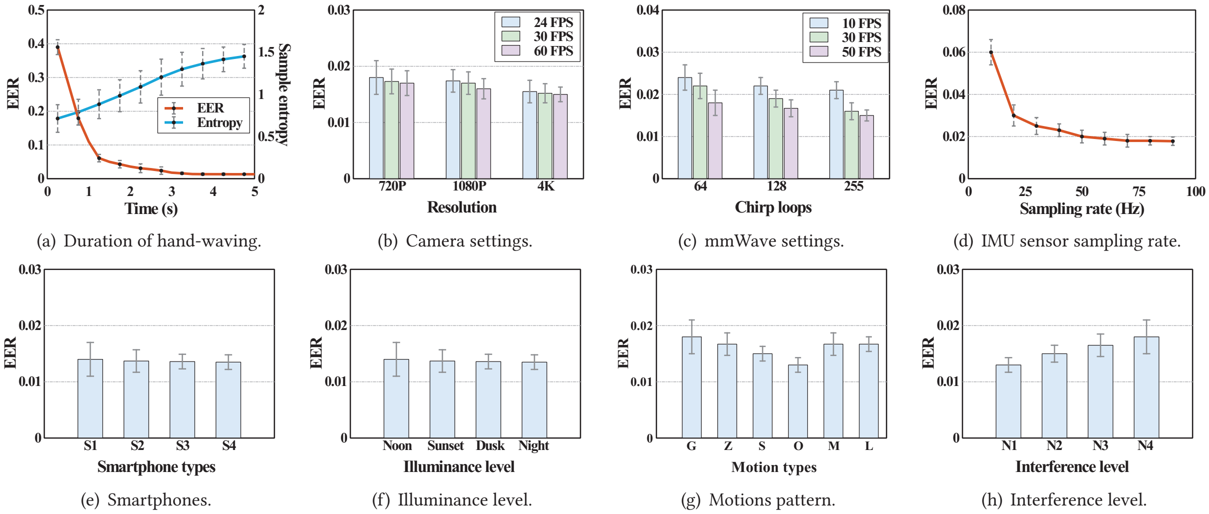
Against relay attacks. As outlined in Sec. 4, The effectiveness of a relay attack depends on the attacker's ability to accurately replicate the user's hand-waving. Our analysis has demonstrated that W4S is able to withstand imitating attacks, thus rendering relay attacks ineffective against the system. Hence, we conclude that W4S is capable of effectively mitigating relay attacks.

Privacy analysis. Although the user's face is not directly uploaded to the server, privacy concerns may still arise if the vehicle captures the user's face. To mitigate privacy concerns, the user can cover their face in their own way. Moreover, we can employ privacy-protection methods like data anonymization to shield individuals from being identified in captured videos.

8.3 Parameter Evaluation

Duration of Hand-waving. As mentioned in Sec. 4.2, hand-waving duration t is controlled by the app which represents how long the hand-waving lasts. Longer duration provides better security but also takes longer to authenticate, which reduces usability. As shown in Fig. 14(a), the average EER decreases as t increases. The stability increases as t increases. Specifically, the EER drops by 0.099 when t increases from 1 s to 5 s. We can also see that the sample entropy increases with t, indicating that the randomness of the acceleration is increasing. We set t to 3.25 s to balance the security and usability.

Camera Resolution. We evaluate the impact of the camera resolution on W4S by downsampling the resolution of 4K (3840 × 2160) to 1080P (1920 × 1080) and 720P (1280 × 720). As shown in Fig. 14(b), the average EER of W4S decreases as the resolution increases, indicating that the authentication performance is improved with a higher camera resolution. The results show that W4S can achieve EERs lower than 0.02 with different resolutions.



(a) Duration of hand-waving.
(b) Camera settings.
(c) mmWave settings.
(d) IMU sensor sampling rate.
(e) Smartphones.
(f) Illuminance level.
(g) Motions pattern.
(h) Interference level.

Figure 14: Parameter evaluation.

Camera FPS. We analyze how camera FPS affects W4S by reducing video frame rates from 60 FPS to 30 FPS and 24 FPS. Fig. 14(b) shows that higher FPS leads to lower EER, indicating that videos with higher FPS provide more information for authentication. EER drops by 3.41% and 1.15% when FPS increases from 24 to 30 and from 30 to 60, respectively. We find that EER stabilizes at 30 FPS. Therefore, we select it for subsequent experiments.

mmWave Radar Chirp Loops. We examine the impact of mmWave radar chirp loops on W4S's performance, with Fig. 14(c) showing that increasing loops from 64 to 128 and from 128 to 255 reduces ERR by 2.13% and 1.69%, respectively. This decrease in EER can be explained by the fact that larger chirp loops yield smaller velocity resolution, as expressed by $\Delta v = \lambda/(2NT_c)$, where λ is the wavelength, N is the number of chirp loops, and T_c is the chirp period. When the chirp period is fixed, the larger the chirp loops is, the smaller the velocity resolution is, which means a more accurate measurement of the velocity of the smartphone can be obtained.

mmWave Radar Frame Rate. To evaluate the impact of mmWave radar frame rate, as shown in Fig. 14(c), we set the frame rate of the mmWave radar to 10 FPS, 30 FPS, and 50 FPS, respectively. Since hand-waving is a relatively fast process, when the frame rate is set to 10 FPS, the mmWave cannot obtain the accurate acceleration change, and hence the EER becomes higher. Experimental results show that the difference between 30 FPS and 50 FPS is only 0.35%. Therefore, 30 FPS is sufficient to capture accurate acceleration.

IMU Sensor Sampling Rate. We evaluate W4S's robustness to various IMU sensor sampling rates (10 Hz to 100 Hz) and find that its EER drops significantly as the sampling rate increases from 10 Hz to 25 Hz (Fig. 14(d)). However, performance stabilizes at rates higher than 50 Hz, indicating that 50 Hz is optimal for W4S.

Horizontal Distance. We assess W4S's robustness to varying horizontal distances between the user and vehicle (50 cm to 200 cm) with 15 participants performing 20 authentication operations per distance. Tab. 3 shows that while no significant difference is observed as distance increases, performance drops considerably at 200 cm. Nonetheless, W4S maintains EERs smaller than 0.0225, 0.0251, and 0.0363 at distances of 50 cm, 100 cm, and 200 cm, respectively, demonstrating its robustness to varying horizontal distances.

Angle of View. We examine W4S's performance with varying relative view angles between the user and vehicle, measured in azimuth angle from the point of view of the mmWave radar. With 15 participants performing 20 authentication operations per angle, Tab.3 shows that while the EER slightly increases by 0.0126 and 0.0144 when the angle changes from 0 °to 30 °and 30 °to 60 °, W4S still achieves low EERs across various angles.

Smartphones. We evaluate the impact of different smartphones on W4S by testing its performance with each smartphone type. Fig. 14(e) demonstrates that there is no noteworthy difference in performance among the four smartphones. Thus, the size, weight, and operating system of smartphones have minimal effect on W4S.

Illuminance Level. To evaluate the impact of illuminance on the performance of W4S, we collect data based on different times of the day: 1) noon, 2) sunset, 3) dusk, and 4) night. As illustrated in Fig. 14(f), W4S works slightly better with low illuminance levels, probably because it can better position the tracking flash in this case, but no significant differences are observed, indicating that our system can work with different light levels.

Motion Pattern. We assess the robustness of W4S by testing its performance with different hand-waving patterns, each exhibiting varying complexity levels. Six patterns (G, Z, S, O, M, and L) were selected based on their lines and corners. Our evaluation involved 15 participants performing the authentication procedure 20 times for each pattern. Fig. 14(g) shows that all patterns produced comparable EERs below 0.02, indicating the system's robustness. However, the G and S patterns had higher EERs due to more lines and corners.

Interference Level. To examine the impact of background interference on W4S's performance, we conducted experiments under four types of interference: 1) no interference, 2) static interference, 3) dynamic interference, and 4) mixed interference. The results, presented in Fig.14(h), reveal that W4S achieves slightly lower EER in

Table 3: EER under different positions of user.

Angle Distance	0°	30°	60°	Mean
50 cm	0.0122	0.0204	0.0348	0.0225
100 cm	0.0124	0.0235	0.0394	0.0251
200 cm	0.0195	0.0382	0.0512	0.0363
Mean	0.0147	0.0273	0.0418	**0.0280**

the presence of static interference because it can be easily removed using the method discussed in Sec.5. Notably, the system maintains an EER of 0.0157, 0.0171, and 0.0180 under static, dynamic, and mixed interference conditions, respectively.

9 CONCLUSION

In this paper, we propose W4S, a multisensor-based mutual authentication system for unmanned vehicle delivery services. W4S achieves the goals of no need for additional hardware, mutual authentication, being resistant to attacks, and no user privacy leakage risks. We propose a series of signal processing methods to enable the user and the vehicle to extract the 3D acceleration of the user's hand-waving. Moreover, we propose a cross-sensor spatial synchronization and an event-based temporal synchronization approach. Afterward, we propose a novel SNN to obtain the authentication results. Extensive evaluations show that W4S achieves an average EER below 0.013 against various attacks.

ACKNOWLEDGMENTS

The work described in this paper was substantially sponsored by project 62101471 supported by NSFC and was partially supported by the Shenzhen Research Institute, City University of Hong Kong. The work was also partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CityU 21201420, CityU 11201422, and R1012-21), Shenzhen Science and Technology Funding Fundamental Research Program (Project No. 2021Szvup126), NSF of Shandong Province (No. ZR2021LZH010).

REFERENCES

[1] Pablo Arbelaez, Michael Maire, Charless Fowlkes, and Jitendra Malik. 2010. Contour detection and hierarchical image segmentation. *IEEE TPAMI* (2010).

[2] Ron Y Asmar, David T Proefke, Charles J Bongiorno, and Aaron P Creguer. 2017. Method and system for authenticating vehicle equipped with passive keyless system. US Patent 9,710,983.

[3] Gildas Avoine, Muhammed Ali Bingöl, Ioana Boureanu, Srdjan Čapkun, Gerhard Hancke, Süleyman Kardaş, Chong Hee Kim, Cédric Lauradoux, Benjamin Martin, Jorge Munilla, et al. 2018. Security of distance-bounding: A survey. *ACM CSUR* (2018).

[4] John H Borghi. 1965. Distribution of human reaction time. *Perceptual and motor skills* (1965).

[5] Guowei Cai, Ben M Chen, and Tong Heng Lee. 2011. Coordinate systems and transformations. In *Springer Unmanned rotorcraft systems*.

[6] Gerard Canal, Sergio Escalera, and Cecilio Angulo. 2016. A real-time human-robot interaction system based on gestures for assistive scenarios. *Elsevier CVIU* (2016).

[7] Christine Chou. 2021. Alibaba's Driverless Robots Just Made Their One Millionth E-commerce Delivery. https://www.alizila.com/alibaba-driverless-robots-one-millionth-ecommerce-delivery/.

[8] Clevon. 2022. CLEVON 1. https://clevon.com.

[9] Cas Cremers, Kasper B Rasmussen, Benedikt Schmidt, and Srdjan Capkun. 2012. Distance hijacking attacks on distance bounding protocols. In *IEEE S&P*.

[10] Dog-qiuqiu. 2021. Yolo-FastestV2: V0.2. https://github.com/dog-qiuqiu/Yolo-FastestV2.

[11] Boyu Fan, Xuefeng Liu, Xiang Su, Pan Hui, and Jianwei Niu. 2020. EmgAuth: An EMG-based smartphone unlocking system using siamese network. In *IEEE PerCom*.

[12] FedEx. 2022. Roxo. https://www.fedex.com/en-us/innovation/roxo-delivery-robot.html.

[13] Andrea Ferlini, Dong Ma, Robert Harle, and Cecilia Mascolo. 2021. EarGate: Gait-based user identification with in-ear microphones. In *ACM MobiCom*.

[14] Aurélien Francillon, Boris Danev, and Srdjan Capkun. 2011. Relay attacks on passive keyless entry and start systems in modern cars. In *NDSS*.

[15] Le Guan, Jun Xu, Shuai Wang, Xinyu Xing, Lin Lin, Heqing Huang, Peng Liu, and Wenke Lee. 2016. From physical to cyber: Escalating protection for personalized auto insurance. In *ACM SenSys*.

[16] Anthony James. 2022. DHL teams up with TuSimple on autonomous trucking operations. https://www.autonomousvehicleinternational.com/news/trucks/dhl-teams-with-tusimple-on-autonomous-trucking-operations.html.

[17] Meng Jin and Xinbing Wang. 2022. Pairing IoT Devices with Spatial Keys. In *ACM/IEEE IPSN*.

[18] Timothy David Kentley-Klay, Rachad Youssef Gamara, and Gary Linscott. 2019. Software application to request and control an autonomous vehicle service. US Patent 10,446,037.

[19] Felix Kreuk, Yossi Adi, Moustapha Cisse, and Joseph Keshet. 2018. Fooling end-to-end speaker verification with adversarial examples. In *IEEE ICASSP*.

[20] Zeyu Lei, Yuhong Nan, Yanick Fratantonio, and Antonio Bianchi. 2021. On the insecurity of SMS one-time password messages against local attackers in modern mobile devices. In *NDSS*.

[21] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *CVPR*.

[22] Xiaopeng Li, Fengyao Yan, Fei Zuo, Qiang Zeng, and Lannan Luo. 2019. Touch well before use: Intuitive and secure authentication for iot devices. In *ACM MobiCom*.

[23] Trupil Limbasiya, Ko Zheng Teng, Sudipta Chattopadhyay, and Jianying Zhou. 2022. A Systematic Survey of Attack Detection and Prevention in Connected and Autonomous Vehicles. *arXiv preprint arXiv:2203.14965* (2022).

[24] Marketsandmarkets Research Private Ltd. 2022. Autonomous Last Mile Delivery Market. https://www.marketsandmarkets.com/Market-Reports/autonomous-last-mile-delivery-market-41240862.html.

[25] Nuro, Inc. 2022. Nuro. https://www.nuro.ai/.

[26] Daniel Peralta, Isaac Triguero, Raul Sanchez-Reillo, Francisco Herrera, and José Manuel Benítez. 2014. Fast fingerprint identification for large databases. *Pattern Recognition* (2014).

[27] Sonia Mutreja Prateek Yadav. 2022. Autonomous Last Mile Delivery Market by Application, Solution, Range and Vehicle Type: Global Opportunity Analysis and Industry Forecast, 2021-2030. https://www.alliedmarketresearch.com/autonomous-last-mile-delivery-market.

[28] Kasper Bonne Rasmussen and Srdjan Capkun. 2010. Realization of RF distance bounding. In *USENIX Security*.

[29] RobotoLab. 2022. Price of NAO ROBOT V6. https://www.robotlab.com/store/nao-power-v6-educator-pack.

[30] Mohamed Sabt, Mohammed Achemlal, and Abdelmadjid Bouabdallah. 2015. Trusted execution environment: what it is, and what it is not. In *IEEE Trustcom*.

[31] Ralph Schmidt. 1986. Multiple emitter location and signal parameter estimation. *IEEE TAP* (1986).

[32] Sean Scott. 2019. Meet Scout. https://www.aboutamazon.com/news/transportation/meet-scout.

[33] Jonathan Sharp, Chuxiong Wu, and Qiang Zeng. 2022. Authentication for drone delivery through a novel way of using face biometrics. In *ACM Mobicom*.

[34] Sheng Shen, Mahanth Gowda, and Romit Roy Choudhury. 2018. Closing the gaps in inertial motion tracking. In *ACM MobiCom*.

[35] Starship Technologies. 2022. Starship. https://www.starship.xyz/.

[36] Udelv. 2008. Udelv. https://www.udelv.com/.

[37] Universal Postal Union. 2021. Number of parcels distributed worldwide from 2015 to 2020. https://www.statista.com/statistics/737418/parcel-traffic-worldwide-by-sector/.

[38] Vincent T Van Hees, Lukas Gorzelniak, Emmanuel Carlos Dean León, Martin Eder, Marcelo Pias, Salman Taherian, Ulf Ekelund, Frida Renström, Paul W Franks, Alexander Horsch, et al. 2013. Separating movement and gravity components in an acceleration signal and implications for the assessment of human daily physical activity. *PloS one* (2013).

[39] David Wells, Norman Beck, Alfred Kleusberg, Edward J Krakiwsky, Gerard Lachapelle, Richard B Langley, Klaus-peter Schwarz, James M Tranquilla, Petr Vanicek, and Demitris Delikaraoglou. 1987. Guide to GPS positioning. In *Citeseer Canadian GPS Assoc*.

[40] Wired. 2017. Just a Pair of These $11 Radio Gadgets Can Steal a Car. https://www.wired.com/2017/04/just-pair-11-radio-gadgets-can-steal-car/.

[41] Chuxiong Wu, Xiaopeng Li, Lannan Luo, and Qiang Zeng. 2022. G2Auth: Secure mutual authentication for drone delivery without special user-side hardware. In *ACM Mobisys*.

[42] Chuxiong Wu and Qiang Zeng. 2023. Turning noises to fingerprint-free "credentials": Secure and usable authentication for drone delivery. *arXiv preprint arXiv:2302.09197* (2023).

[43] Weitao Xu, Girish Revadigar, Chengwen Luo, Neil Bergmann, and Wen Hu. 2016. Walkie-Talkie: Motion-assisted automatic key generation for secure on-body device communication. In *ACM/IEEE IPSN*.

[44] Xiangyu Xu, Jiadi Yu, Yingying Chen, Qin Hua, Yanmin Zhu, Yi-Chao Chen, and Minglu Li. 2020. TouchPass: Towards behavior-irrelevant on-touch user authentication on smartphones leveraging vibrations. In *ACM MobiCom*.

[45] Yang Zhang, Peng Xia, Junzhou Luo, Zhen Ling, Benyuan Liu, and Xinwen Fu. 2012. Fingerprint attack against touch-enabled devices. In *ACM SPSM*.

[46] Bing Zhou, Zongxing Xie, Yinuo Zhang, Jay Lohokare, Ruipeng Gao, and Fan Ye. 2021. Robust Human Face Authentication Leveraging Acoustic Sensing on Smartphones. *IEEE TMC* (2021).