Journal of Information Security and Applications 66 (2022) 103130

Contents lists available at ScienceDirect

Journal of Information Security and Applications

journal homepage: www.elsevier.com/locate/jisa

WiCapose: Multi-modal fusion based transparent authentication in mobile environments☆

Zhuo Chang [a,b], Yan Meng [c], Wenyuan Liu [a], Haojin Zhu [c], Lin Wang [a,*]

[a] School of Information Science and Engineering, Yanshan University, Qinhuangdao, 066004, China
[b] School of Cyber Security and Computer, Hebei University, Baoding, 071000, China
[c] Department of Computer Science Engineering, Shanghai Jiao Tong University, Shanghai, 200240, China

ARTICLE INFO

Keywords:
Channel state information
Two-factor authentication
Dual side-channel
Rear camera
Mobile sense

ABSTRACT

The prosperity of mobile sensing technology makes smartphone-based authentication more prevalent in mobile environment. The present single-modality security authentication based on human biometrics is vulnerable to counterfeit, and the procedure of basic two-factor authentication (2FA) is cumbersome. In this work, we propose a 2FA method without additional devices and procedures, namely *WiCapose*. The essential technique is to use the unique correlation of the two side-channel to complete multi-modal feature extraction and fusion authentication. *WiCapose* can simultaneously extract the radio frequency (RF) gain feature that directly characterizes the finger movement and the inter-frame spatio-temporal difference from the rear camera that indirectly portrays the tapping behaviors, to blend the micro-scale behavior pattern feature of the user's finger and implicit biological features. Then we design and train a deep neural network to fuse both factors for mitigating the limitation of the RF factor in environmental generalization and the uniqueness of the image factor on authentication performance. Experiments involving ten participants demonstrate that our method can achieve a 98% average accuracy on authentication and effectively resist shoulder-surfing attacks and mimic attacks.

1. Introduction

With the popularization of mobile environment, its security has attracted widespread attention. Many biometric authentication schemes have emerged, relying on the human body's exterior unique physiological or behavioral features, such as fingerprint, iris, face, voiceprint, and habitual behaviors.

However, these authentications based on exterior biometric information are highly vulnerable to imitation attacks, such as attacks on face recognition by printed eyeglass frames [1], invasion of facial privacy by the front camera [2,3], and mimic voiceprint by the pre-recorded voice [4], counterfeit fingerprints by silica gel [5]. Nevertheless, the present Two-Factor Authentication (2FA) can prevent these attacks but raises authentication complexity, e.g. Yubikey [6], requiring users to carry additional hardware devices, or increase authentication procedures which reduces user experience and system availability. So is it feasible to take a transparent 2FA?

In the mobile scenarios, the widely deployed Wi-Fi devices provide the Channel State Information (CSI), a non-intrusive and location-dependent side-channel information, which are harnessed to perceive human behavior [7,8], even breathing [9], tracking [10], sleep detection [11], fall detection [12], gesture recognition [13], ReID [14], and respiration detection [15]. Furthermore, the evolved two-factor authentication [16,17], combined with these advantages of wireless perception and the unique biological features of fingers, is transparent to users who only need a single rapid Personal Identification Number (PIN) input process to simultaneously obtain dual side-channel data to complete the entire 2FA mobile enhanced security authentication.

This paper proposes a mobile security authentication system based on an intelligent fusion of interior multi-modal side-channel sensing information, *WiCapose*, as shown in Fig. 1. It is well-known that every ordinary person has his living habits and unique behavior features [18,19]. When the mobile system authenticates the user, biometric features of the hands and fingers [20–22] can be collected by the CSI through

☆ The work supported in part by the National Science Foundation of China under Grant 61772453, Grant 61972453; and in part by the Natural Science Foundation of Hebei Province under Grant F2020203074.

* Corresponding author.
E-mail addresses: changzhuo@stumail.ysu.edu.cn (Z. Chang), yan_meng@sjtu.edu.cn (Y. Meng), wyliu@ysu.edu.cn (W. Liu), zhu-hj@cs.sjtu.edu.cn (H. Zhu), wlin@ysu.edu.cn (L. Wang).

https://doi.org/10.1016/j.jisa.2022.103130

Available online 29 March 2022

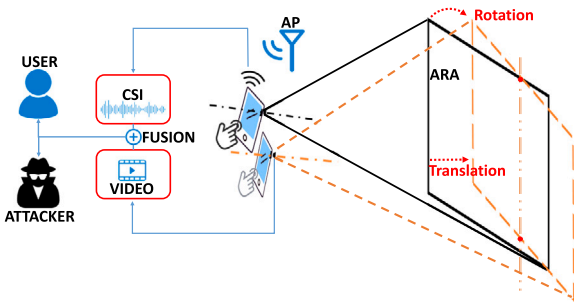


Fig. 1. *WiCapose*, a novel dual side-channel mobile perception authentication system.

ubiquitous WiFi transceiver devices. Moreover, the posture of the hand and the habitual force of tapping the screen during input cause the phone to tilt a certain angle, which is expressed by pose information, and people with different habits lead to different tilt angles when hitting keys. From there, the smartphone's pose could be extracted from a video shot by the rear camera face to compose a unique feature of personal behaviors, which has a strong correlation with CSI. The rear camera faces an unobstructed Authentication Reflection Area (ARA) covering an area of arbitrary plane in mobile computing environments, such as floor and desktop. *WiCapose* adopts the interior, non-imitation, and non-copying of the correlation between the two side-channel biometric modes of micro-scale finger behaviors without additional devices and procedures. It can effectively and conveniently guarantee the mobile system's security and robustness by preventing both mimic attacks and shoulder-surfing attacks.

To the best of our knowledge, *WiCapose* is the first to leverage wireless side-channel features of dynamic behavior, combined with camera pose change for smart-home user authentication. Combined with the above factors, *WiCapose* has the following three challenges.

(1) ***How to extract the unique features of user behavior through high-resolution CSI segmentation.*** Using smartphones to collect CSI information is a pervasive instrumental technology, but its defect is that the power of the Network Interface Controller (NIC) chip in a smartphone is much weak. So there is much noise in the collected CSI data.

(2) ***How to extract posture transformation features of the handheld devices synchronously.*** When users login, and tap on the smartphone, the procedure is completed according to individual unique habits, such as the angle of the handhold, the strength of the tapping, and length of duration, the continuous space scope of finger operations. All can be utilized to authenticate the user's uniqueness.

(3) ***How to realize multi-modal heterogeneous information fusion on the energy-constrained platform.*** The deep model's training requires extreme computing capability and power to achieve multi-modal heterogeneous information fusion. Moreover, the location-dependent of CSI could be mitigated by weighted deep fusion.

The main contributions of the paper can be summarized as follows.

- We present a transparent 2FA system, *WiCapose*, which segments the CSI to obtain the high-resolution information, performs multi-modal fusion combined with the camera pose transformation features of the smartphone, to extract the side-channel unique features of the user for authentication, prevent a variety of attacks, and enhance the security capabilities of mobile environment and other systems.
- We devise a deep neural network for finger behaviors' features extraction given the diversity of finger movements in complex scenes. It can effectively mitigate location-dependent RF gain features and relieve the system's constraints on user posture and background noise to improve authentication accuracy.
- We design and implement *WiCapose* on Nexus5, which could noninvasively perceive user behaviors to expand the application scenarios of CSI. Abundant experimental results show that the accuracy achieves 98.3% through the unique identification of side-channel features. The result verifies the safety and reliability of system authentication.

The remainder of this paper is organized as follows. In Section 2, we introduce the preliminaries of this work and the attacks that this system can defend. We present the architecture design in Section 3, which is followed by CSI Representation, Camera Pose Calibration, Multi-Modal Fusion, Evaluation, and Discussion in Sections 4, 5, 6, 7, respectively. Finally, we conclude this paper in Section 8.

2. Preliminaries

2.1. Channel state information

CSI represents the influence of multi-path, such as Line Of Sight (LOS), reflection, diffraction, and impact of actions on the channel during signal transmission [23]. When the environment is fixed, the static part is a constant independent of time-variant. During signal transmission, there are phase shifts caused by the devices and noise interference in the environment, so the corresponding transmission signal is derived as the (1):

$$\begin{aligned}\hat{H}(t,f) &= e^{-jn(t)f}H(t,f)+\mu(t,f)\\ &= e^{-jn(t)f}(H_s(f)+H_d(t,f))+\mu(t,f)\\ &= e^{-jn(t)f}H_s(f)+e^{-jn(t)f}H_d(t,f)+\mu(t,f).\end{aligned} \tag{1}$$

In (1), $H_s(f)$ is a static variable in the multi-path transmission, $H_d(t,f)$ represents the dynamic part, $e^{-jn(t)f}$ is the frequency offset due to hardware, and $\mu(t,f)$ represents additive white Gaussian noise. The received CSI signal is affected by many factors, including frequency deviation caused by hardware defects, environmental noise, and signal superimposition caused by the multi-path effect. Especially as limited hardware performance, the signal deviation is large, and the CSI amplitude fluctuates wildly. So, the amplitude and phase information of collected CSI contains a lot of noise, and denoising cannot perform well. In the following work, these factors which impact authentication are removed.

2.2. Feature component

The smartphone's motion is approximately regarded as a rigid body movement, denoted by the rotation matrix V_R and the translation matrix V_T, extracted from the video taken by the smartphone's rear camera. In high-security application scenarios, both rotation and translation caused by tapping the screen during the authentication process, represents the user's behavior features, regarded as a tuple: $\langle V_R, V_T\rangle$. Combining with the previously extracted CSI features of finger behavior compose unique user behavior features: $\langle CSI,(V_R,V_T)\rangle$. The user's input process is a sequential continuous process $\{\langle t_1,t_2\rangle\langle t_3,t_4\rangle\cdots\langle t_i,t_j\rangle\cdots\}$, where $t_{i/j}$ denotes the time index of each independent action, and the entire authentication process can be represented as:

$$\begin{aligned}Sequence_{input} =\{&\langle CSI,(V_R,V_T)\rangle_{\langle t_1,t_2\rangle},\\ &\langle CSI,(V_R,V_T)\rangle_{\langle t_3,t_4\rangle},\dots,\\ &\langle CSI,(V_R,V_T)\rangle_{\langle t_i,t_j\rangle}\dots\},\end{aligned} \tag{2}$$

where $\langle CSI,(V_R,V_T)\rangle_{\langle t_i,t_j\rangle}$ represents the feature matrix of an behavior which starting from t_i to t_j. The deep neural network fuses all timing-related unique features generated during the entire procedure to express each different user, which is formula as:

$$Auth_{\mathrm{y/n}} = f_{DNN}(Feature_{sequence_i}). \tag{3}$$

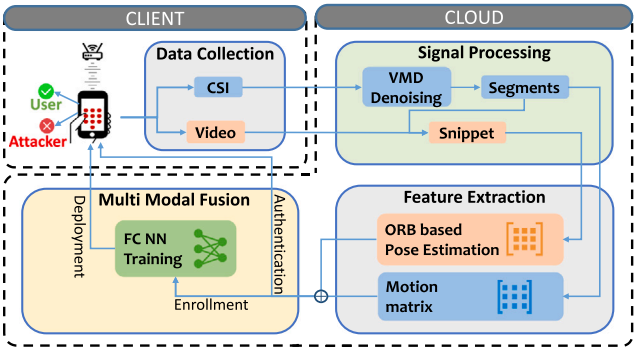


Fig. 2. The authentication framework of *WiCapose* consists of four components: data collection, signal processing, feature extraction, and multi-modal fusion.

2.3. Threat model

The access to devices and resources of mobile environment requires high secure and reliable authentication to protect user privacy and security. However, current authentication methods are vulnerable to mimicry attacks, shoulder-surfing attacks, so enhancing the user authentication security of mobile system is a very challenging issue. First, we assume that the attackers are not resourceful in terms of cloud servers, and the attacks on cloud servers are beyond the scope of this paper.

Shoulder-surfing Attack: There is always someone from behind who can spy on or monitor your input process to record your input pattern or concrete password content [24]. Its weakness lies in that the input required in the system authentication process is only the password string or a graphical unlock pattern which can be optionally recorded by anybody. Without additional security authentication information, it brings challenges to the security of the system.

Mimicry Attack: In certain attack scenarios, an experienced adversary can shoulder surfing the victim's authentication mode, such as key string, input behavior pattern and other information. When the victim is not present, imitate the victim's behavior to invade the mobile payment system [25].

3. System design

As different people have different behavior unique patterns, *WiCapose* is mainly base on the finger's unique features. When the user enters the smartphone's critical authentication information, it can get the user's behavior pattern data to extract the user's unique features and then authenticate the user safely on the smartphone. We use the WiFi signal to collect CSI information, and the rear camera collects posture transformation data caused by finger tapping, integrates both into multi-modal behavior features to perform two-factor authentication. Fig. 2 shows the overview of the system design, which consists of four components, data collection, signal processing, feature extraction, and multi-modal fusion.

Data acquisition section: An Android application collects the CSI data of behavior and camera pose information during the user's input stage on the mobile device. The user's original data, including CSI data packets and video sequences, are transmitted to the cloud through the wireless network and processed efficiently.

The signal processing part mainly denoises and segments the CSI, extracts corresponding amplitude fluctuation data of the behavior, and then uses the start and end time points of the CSI segmentation to extract the corresponding period's snippets from the video frame sequence. Precisely, in the denoising part, the Variational Mode Decomposition (VMD) algorithm and smooth filter remove signal noise. For more effectively highlighting each activity period's fluctuation level, the Short-Time Fourier Transform (STFT) transform is used for time–frequency analysis. On this basis, accumulated signal energy on each frequency, and the average value of energy is set as the threshold to find the user's activity's start and end time points.

The third part, feature extraction, uses the Oriented FAST and Rotated BRIEF (ORB) eight-point algorithm to calculate the camera pose transformation matrix of every two frames in the segmented video. Including rotation matrix V_R, translation matrix T, and CSI amplitude matrix are fused to form the user's final feature matrix. Finally, in the fourth part, feed the combined feature matrix to a Fully Connected (FC) neural network for training. The original matrix input to the neural network contains three parts of feature matrices with different dimensions. We do extraction and prediction through the FC network's nonlinear ability with high efficiency and low energy consumption and then deploy the trained network model to authenticate users.

4. CSI representation

4.1. Denoising

First of all, we use smart mobile devices based on nexmon firmware [26] to directly collect CSI, which provides vital portability, mobility, effortless deployment, and feature enhancement. According to the Fresnel Zone model [27], the aggregated signal is precisely affected by these transmitters' location and angle, so if the device's distance cannot be controlled, it is tough to portray the uniqueness of CSI-based user behavior features. However the CSI information collected by smartphone is not as stable as the data collected by traditional devices. Influenced by the environment, deviations of software and hardware, the composite signal fluctuations have been seriously interfered by noise, and the core signal segmentation of behaviors is confused. For extracting useful CSI segment indexes and features of user behaviors, it is necessary to denoise the signal in advance.

At first, we use ordinary least squares to quadratic fitting the data trend and subtract the current fitting curve from the CSI so that the mean value of the detrended data is zero. After that, *WiCapose* focuses on analyzing the data's variation, highlighting the finger gesture's signal fluctuation, and it is easier to complete the subsequent segmentation and extraction of the behavior features. Next, *WiCapose* eliminates outliers. Because the wireless NIC chip in the smartphone is unstable, the finger gesture's amplitude information is concealed in the outlier noise. Hampel filter is used to filter out most outliers without affecting the overall signal's fluctuation scale and integrity.

Secondly, we apply VMD [28] to the signal. The signal trend fluctuation is the most important, and CSI is a kind of random signal, so the VMD adaptive decomposition is used to separate the effective signal fluctuation trend and remove the noise of other frequencies. One of the VMD algorithm's benefits is that the number of Intrinsic Mode Function (IMF) components can be set manually according to device characteristics and experience. Moreover, as the number of decomposition components increases, high-frequency noise produces intermittent phenomena, which does not affect the low-frequency region's pattern trend where the finger moves. We have collected some data in the meeting room for verification. According to experience, we set the number of IMFs to 3 to achieve a better decomposition effect, as shown in Fig. 3. The top of the figure is the original signal of several subcarriers. The lower part of the figure is the result after VMD denoising, which can show the signal segments.

4.2. Subcarrier selection

Subcarriers of different frequencies give further feedback to operations on the wireless communication link, and some have more noticeable effects on behavior features expression. When extracting features in action's CSI segments, they can effectively characterize actions and serve as neural networks' input to extract unique features. On the

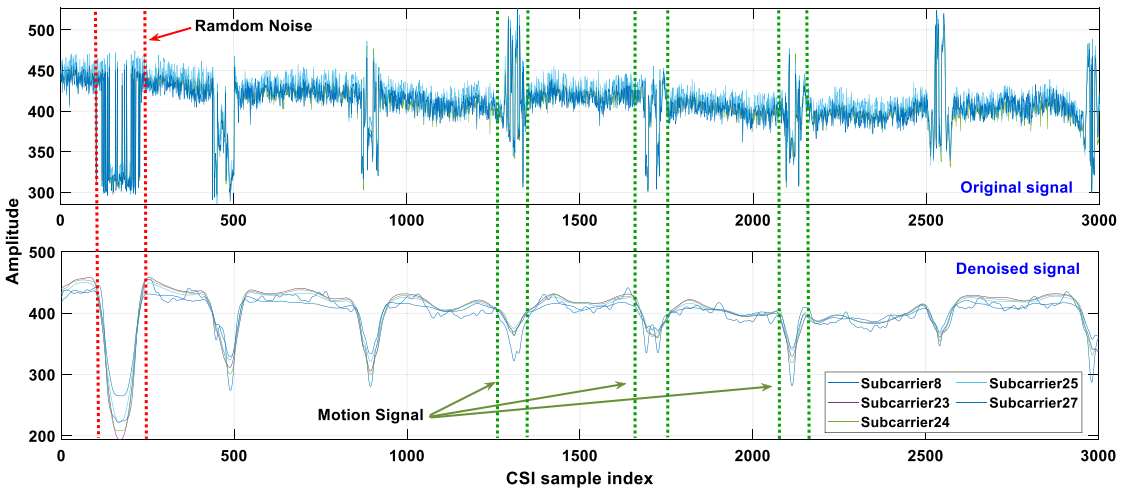


Fig. 3. Illustration of amplitude of CSI signal and motion section after denosing.

contrary, if the subcarrier signal fluctuates weak or with no regularity, it is not easy to segment it, which cannot fully reveal the behavior's feature information and form the feature expression corresponding to the amplitude-phase of the behaviors' signal. So it is a critical step to efficiently and automatically make dynamic subcarrier selection that appropriately responds to the behavior mode. The Auto Correlation coefficient (ACF) technique [29] evaluates the effectiveness of different subcarriers in CSI and efficiently extracts subcarrier signals that can reflect stable behavior features. The definition of ACF for static signals as (4) below:

$$\rho(\tau) = \frac{Cov[x(t), x(t-\tau)]}{\sqrt{Var[x(t)]}\sqrt{Var[x(t-\tau)]}}, \tag{4}$$

where $\rho(\cdot)$ is autocorrelation coefficient, $x(t)$ is time signal, τ denotes the lags, $Cov[\cdot]$ is covariance, and $Var[\cdot]$ represents variance. $Var[x(t)]$ and $Var[x(t-\tau)]$ are equal, so we define

$$\hat{\gamma}(\tau, f) = \frac{1}{n}\sum_{t=1}^{n-\tau}[x(t+\tau, f) - \bar{x}][x(t, f) - \bar{x}], \tag{5}$$

$\bar{x}$ denotes mean of x, then

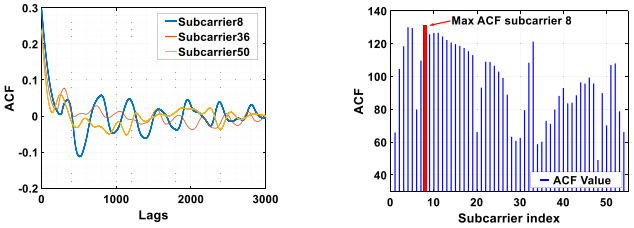
$$\rho_{\hat{H}(t,f)}(\tau, f) = \frac{\hat{\gamma}(\tau, f)}{\hat{\gamma}(0, f)}. \tag{6}$$

Generally speaking, the ACF calculation of each subcarrier can be obtained by (6), where $\hat{\gamma}(0, f)$ represents variance, $\hat{\gamma}(\tau, f)$ is covariance of $x(t, f)$ User behaviors influence CSI subcarriers to fluctuate. Especially, verification behaviors are periodic and comply with established habits and unique patterns. Therefore, we need only to perform autocorrelation calculation on each subcarrier internal lag to get the current fluctuating curve and select the most periodic subcarrier for subsequent data segmentation and feature extraction. As shown in Fig. 4(a), all subcarriers' ACF is calculated, and the value of subcarrier 8 is maximum. Fig. 4(b) indicates that the ACF trend of subcarrier 8 is the most obvious, and the periodic behavioral feature is also the most explicit and reasonable.

4.3. Time–frequency analysis

The CSI subcarrier signal is affected in many ways. One of them is additive white Gaussian noise. After the signal is denoised, the selected subcarrier's signal shows distinct fluctuation of the action sequence. Next, we need to extract the section related to the behavior, that is, segment the time series CSI signal.

After CSI filtering and denoising, we need to isolate the motion-related CSI segments for feature extraction. However, some ineffective signal fluctuations caused by hardware defects are apparent. So we



(a) ACF Curves of Subcarriers (b) ACF Value of 54 Subcarriers

Fig. 4. Subcarriers selection by ACF.

perform the STFT and convert the CSI signal to the frequency domain for spectrum analysis according to (7):

$$\begin{aligned} G(t, f) &\triangleq k\left|\hat{H}(t, f)\right|^2, k = \frac{2}{F_s\sum_{n=1}^{N}|w(n)|^2} \\ &= k(|H(t, f)|^2 + 2\,\mathrm{Re}\{n^*(t, f)e^{-jn(t)f}\,|H(t, f)|\} \\ &+ |\mu(t, f)|^2) \\ &\triangleq k|H(t, f)|^2 + k\mu(t, f), \end{aligned} \tag{7}$$

where $w(n)$ is hamming window, F_s is the sampling rate, and k represents the coefficient of signal energy. The movement of the Hamming window can reflect the signal fluctuations generated by user actions at various moments. Next, we accumulate all Power Spectral Density (PSD) along the frequency dimension in the spectrum according to (8).

$$\begin{aligned} P_a(t) &= \sum_{f=f_1}^{f_2} G(t, f) = \sum_{f=f_1}^{f_2} k\left|\hat{H}(t, f)\right|^2 \\ &= \sum_{f=f_1}^{f_2}\left[k|H(t, f)|^2 + k\mu(t, f)\right]. \end{aligned} \tag{8}$$

As we can see from Fig. 5, the difference in signal strength caused by behaviors is evident in spectrum and has a robust periodic effect. The red areas in this figure mean the tapping behaviors. But on the left in the spectrum analysis graph, there is a signal fluctuation due to hardware defects. The cumulative value of the power spectral density of different frequencies cannot clearly distinguish the boundary of the segment and even miss some segment information. To facilitate the better distinction of the action signal boundary, we proceed to calculate the variance of accumulation PSD, as shown in (9), where $E(\cdot)$ denotes expectation, P_a is PSD accumulation. The result of $Variance_G$ is more

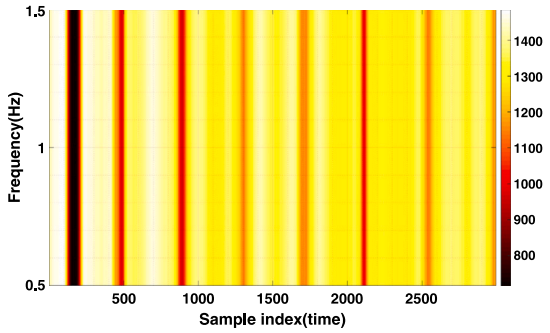


Fig. 5. The spectrogram of time–frequency analysis. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

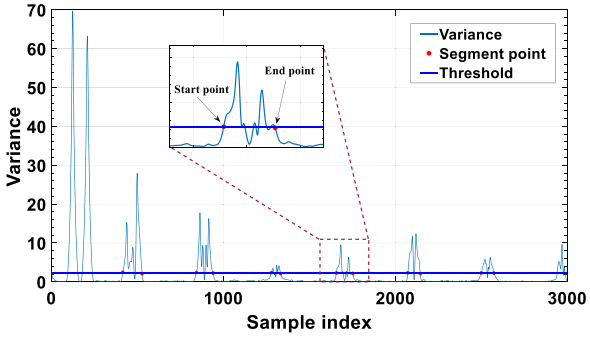


Fig. 6. Energy accumulation variance and boundary points of segmentation.

suitable to extract behavior segments.

$$Variance_G = \frac{1}{N}\sum_{t=1}^{N}(P_a(t) - E(P_a))^2. \tag{9}$$

Then the smooth operation is performed on the spectral variance result, which makes the entire curve easy to process the boundary judgment, as shown in Fig. 6. At this time, the average threshold is used to identify the CSI segments of the finger movement accordingly. To avoid the influence of several interference segments caused by the hardware defects, *WiCapose* designs a segment search algorithm, presented in Algorithm 1, to perform perfect segment boundary adaptive extraction and obtain the CSI sampling point indexes value of the beginning and end of a segment.

Finally, we extract the CSI signal segment by the methods mentioned in the above subsections. The segmented CSI that expresses the fingers' tapping biometric features can be used to fuse with the following video side-channel information for user identity authentication.

5. Camera pose construction

In a video sequence, we can use Visual Odometry to obtain the camera pose. A finger tap on the smartphone causes the smartphone's rear camera to vibrate, which can be regarded as the fundamental transformation of a rigid body in which geometric objects rotate and translate in 3-D space. We estimate the camera's motion state by calculating the motion relationship between adjacent frames in the video and obtaining the rotation and translation matrices to characterize the essential behavioral features during user authentication. The rotation matrix V_R and the translation matrix V_T represent the rigid body's primary motion, that is, the Euclidean Transform of the two coordinate systems to denote the camera pose change. Such as the (10), where a' is the transformed coordinate.

$$a' = V_R a + V_T, \tag{10}$$

Algorithm 1: CSI Motion Segmentation.

Input: N consecutive CSI spectrum $Variance_{\Sigma G}$,
Output: Array of CSI segment index $seg\_index[i]$

initialization boundary threshold=mean$\left(Variance_{\Sigma G}\right)$;

%Find these segment indexes.
while *not end of* $Variance_{\Sigma G}$ **do**
  if $Variance_{\Sigma G}[j] < threshold$ **then**
    i=findnext($Variance_{\Sigma G}[k] > threshold$);
  end
end
Assign index to $seg\_index[i]$;

%Combine the very close adjacent segments
while *not end of* $seg\_index[i]$ **do**
  if $Distanceof\,(i, i+1) < threshold\_distance \&\& Sumlength\,(i, i+1) < threshold\_length$ **then**
    Combine$(i, i+1)$;
  end
end

%remove the very short segments
while *not start of* $seg\_index[i]$ **do**
  if $Seglength\,(i) < threshold\_shortLen$ **then**
    Remove(i);
  end
end

We use homogeneous coordinates and transformation matrices to maintain the linear relationship of multiple transformations include rotation and translation, as (11):

$$\begin{bmatrix} a' \\ 1 \end{bmatrix} = \begin{bmatrix} V_R & V_T \\ 0^T & 1 \end{bmatrix} \begin{bmatrix} a \\ 1 \end{bmatrix} \triangleq T' \begin{bmatrix} a \\ 1 \end{bmatrix}. \tag{11}$$

Using the monocular camera in smartphone to collect video for pose estimation, the action amplitude caused by finger behaviors to the camera is small, which satisfies the hypothesis that there is no excessive motion between two adjacent frames. Given the computing power and memory space constraints, we adopt the highly efficient ORB [30,31] algorithm to perform rotation V_R and translation V_T estimation. The extracted binary feature descriptors must match each other and fulfill data associations of frames. If matching the descriptors (x_t^m, x_{t+1}^n), we can precisely estimate the camera pose, where x_t^k denotes k feature points in frame t. We use hamming distance to measure the distance of feature points so that the closest descriptors are regarded as the matching feature points.

Through the above theoretical analysis, we can prepare to extract the jitter side-channel information in the video. Nevertheless, before that, we still need to do some preparatory work. First, the segments with jitter information in the video sequence must be separated.

5.1. Alignment and segmentation

Before the pose estimation, we should extract video snippets recorded while the finger taps the phone screen. Fortunately, we use the same device clock to collect CSI data and video, so the time signal is synchronized. The sampling rate of CSI is 90 pks/s, and the video's frame rate is 30 fps. A single video frame corresponds to three CSI sampling points. The CSI segment indexes are obtained in the above Section 4.3 maps to the video frame indexes and takes out the corresponding video frames to be snippets. The corresponding camera pose matrices are calculated between every two adjacent frames of them.

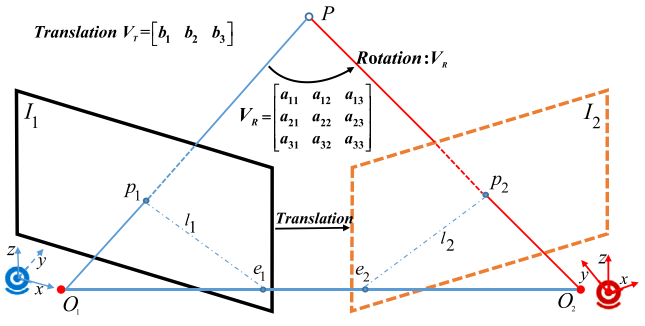


Fig. 7. Visual Odometry schematic diagram.

Considering more details, when the phone shots a video, the storage efficiency is slightly lower, and the signal time lags several milliseconds, so the delay part is deleted from the CSI data. According to experience, we only delete the first 200 sampling points, the CSI and the video frame are aligned accurately.

5.2. Visual odometry pose estimation

The extracted video snippets aim to adjacent frame pose estimation. In the case of several pairs of matching descriptors, we model the camera's rigid motion of two adjacent frames by Epipolar Geometry, as shown in Fig. 7, and derive the corresponding rotation matrix and translation matrix of the Euclidean transformation for reconstructing the movement of the camera lens between the two image planes I_1 and I_2.

O_1, O_2 are centers of the same camera at adjacent times, while p_1 is a feature point in the I_1 plane, and p_2 is a corresponding feature point in I_2. The two points compose a feature point pair, show that a single point project on the imaging plane with a different angle and time. $\overrightarrow{O_1p_1}$ and $\overrightarrow{O_2p_2}$ intersect the object point P in the 3-D space, and then these three points determine an Epipolar plane. The line of $\overrightarrow{O_1O_2}$, and the imaging plane I_1, I_2 intersect at e_1, e_2, two Epipoles, and l_1 and l_2 are the intersection lines between the Epipolar plane and the two image planes, named Epipolar lines. Assume the coordinate of the P in space is $P = [X, Y, Z]^T$, so we know that the relationship between the two image pixels p_1 and p_2 is denoted as (12):

$$Cor_{p2} = K(V_R P + V_T) = V_R K P + K V_T = V_R Cor_{p1} + K V_T, \tag{12}$$

where K is the camera's internal parameter matrix, Cor_{pi} denotes the coordinates of p_i, V_R and V_T are the rotation matrix and the translation matrix, respectively. Thus according to the epipolar constraint, the corresponding pose (13) is derived.

$$p_2^T K^{-T} V_T \bullet V_R K^{-1} p_1 = 0, \tag{13}$$

where $E = V_T \bullet V_R$ is crucial, named Essential Matrix. It is obtained by using the Eight-point-algorithm, and then the Singular Value Decomposition (SVD) is performed on matrix E to get the singular value matrix $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \sigma_3)$, then

$$E = V_R\ diag(\frac{\sigma_1+\sigma_2}{2}, \frac{\sigma_1+\sigma_2}{2}, 0)\ V_T{}^t. \tag{14}$$

At last, the corresponding rotation V_R and translation matrix V_T are obtained by simple decomposition.

We get the video jitter side-channel information from this section and the CSI in Section 4. Fingers' behaviors generate the two kinds of side-channel information, which are highly correlated to be fused for dual side-channel two-factor authentication. Next, we preprocess the data so that neural networks can be used for feature extraction and fusion.

6. Multi-modal feature fusion

6.1. Regularization

Review the above sections, and the CSI data is a $54 \times N$ matrix. The value of CSI amplitude is different from the numerical range of the camera rotation and translation matrix, so we need to normalize the CSI value before splicing. The matrices of the camera pose are all parameters close to value 1. After the CSI is normalized, it is directly spliced with the camera pose matrices.

6.2. Data augmentation

We collected data from multiple volunteers. They set their password string in advance and entered it several times in a scene. Even so, the size of the dataset is still relatively small. For the current deep neural network, data is the primary factor in training, and we need to perform data augmentation on the data collected by each person separately. *WiCapose* applies GaussianBlur, Crop, AverageBlur, Affine and Pad five operations and randomly select one to three operations for data augmentation through the imgaug [32] library. Finally, a dataset containing 39,000 samples is generated. The data quantity is sufficient for a small neural network training like the one mentioned in our paper.

6.3. Final decision

Large-scale neural networks are not proper because of smartphones' power and storage capacity. Moreover, the network's input is a spliced fusion matrix, and the features of each sub-matrix have no corresponding spatial features, so they are not regarded as actual images. Furthermore, the ultimate goal is to perform classification certification. We use a 3-layer, fully connected network to achieve the purpose based on the above factors. As shown in Fig. 8, the network input is a 200×202 tensor, and the output is the probability produced by the Sigmoid function. The middle full connect module mainly includes $FC \longrightarrow BN \longrightarrow ReLU \longrightarrow Dropout$. In *WiCapose*, the input matrices have no prominent local spatial characteristics like those between adjacent pixels in an image. So *WiCapose* does not use the classic convolution operation but uses the nonlinear fitting ability of the multi-layer fully connected network to fuse dual side-channel information. The BN layer is used for batch normalization, and the following ReLU is for nonlinear activation. The Dropout layer is combined to prevent the overfitting of the fully connected layers and appropriately improve the training speed. Sigmoid is applied to output prediction probability results at the end of the entire network. We use the Mean Square Error (MSE) as the loss function to calculate the loss. After calculating and fitting the training data in the network training process, try to make the final output close to the probability 1 (legal user) or 0 (attacker) to achieve binary classification. Note that the enrollment and authentication process is the same, but we can apply the trained neural network in smartphones for feature extraction and authentication.

7. Evaluation

7.1. Experimental setup

A regular router is used as the transmitter to send WiFi packets. The nexmon kernel module updates the firmware of the Google Nexus5 smartphone, which is regarded as the receiver to collect side-channel data. We have developed an Android-based prototype Application (APP) on Nexus 5 to collect CSI and video information in real-time. As of the experiment's completion, the nexmon project only supports Nexus phones, so Nexus 5 is currently the only experimental model, shown in Fig. 9(a). However, as the application scenarios increase, the cost of adopting other mobile hardware to collect CSI is not high. The regulation scenarios of APP are generally indoor scenarios. We

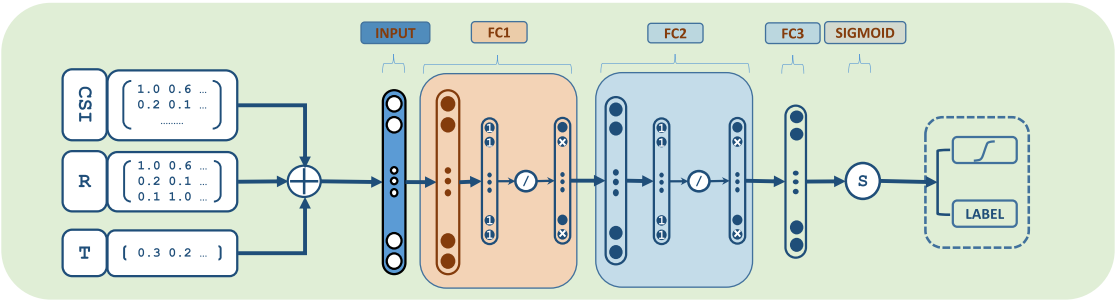
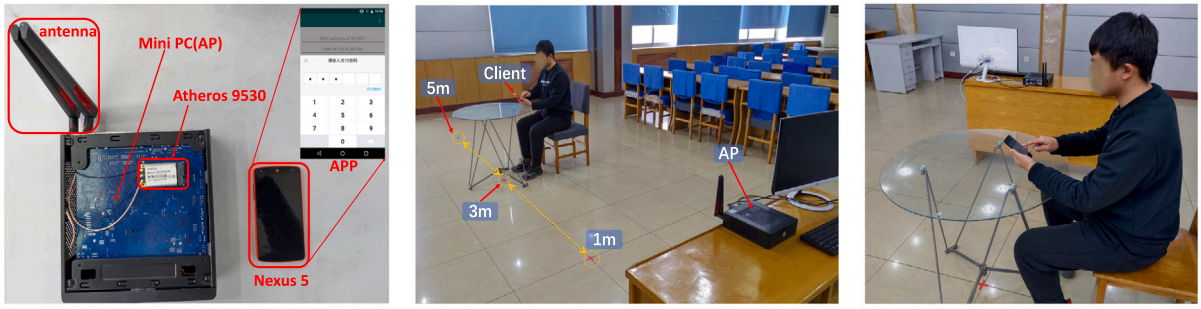


Fig. 8. Architecture of FC neural network.



(a) Illustration of hardware and software (b) Data collection in 3m (c) Typing on screen

Fig. 9. Experimental setup.

experiment with a standard meeting room. In this way, it can reduce the number of influencing factors when collecting CSI, the multipath effects of WiFi signals are all produced by static furniture, and the light is sufficient, which accomplishes the real indoor scenarios.

The D11 core of the nexmon copies the CSI table from the physical layer to the shared memory and pushes it to the collection program in User Datagram Protocol (UDP) packets. The size of shared memory is limited, and system variables for ordinary WiFi networking are also stored. It needs to revise the software to improve signal processing efficiency, especially strengthening the sampling rate. We check the underlying code of nexmon and optimize the micro-code for CSI collection. In addition, we use a mini PC with an Atheros 9530 NIC as the router and send data packets through professional software simulation. Currently, only the beacon packets sent by the router are collected, so the sample rate increase by eight times and reaches 300 pks/s. For more stable CSI collection, the packet sending rate of the router is set at 90 pks/s. Besides, when nexmon collects CSI, a single WiFi antenna supports 80 Mhz bandwidth and gets 256 subcarriers. However, under 20 Mhz bandwidth, many subcarriers have no data, that is, empty subcarriers. Then these empty subcarriers can be removed, and finally, we get 54 subcarriers. That is a $54 \times N$ matrix, where N is the total number of sampling points. The video shooting resolution of the rear camera is 1920×1080, and the frame rate is 30 fps.

7.2. Data collection

We recruit 10 volunteers, including 7 men and 3 women, aged between 24–27. In the conference room, there are 12 sampling positions. As shown in Figs. 9(b) and 9(c), tables are placed at three different distances of 1 m, 3 m, and 5 m from the transmitter, and at each distance, there are four directions: east, west, south, and north, a total of 12 sampling locations. Furthermore, at a distance of 1 m, we also collect the data of test subjects who tap the password and hold the phone by one hand and evaluate the contrast accuracy with the two-handed operation. During the collection, each volunteer maintains his usual natural posture, holds the smartphone to run the Android APP, and then taps the password configured by themselves, and the APP starts to collect the CSI data. By collecting 10 groups at each location for each volunteer, a total of $13 \times 10 = 130$ CSI and video tuples are collected. To prevent slight changes in user posture with different time, we collect users' features data on different dates and conduct training together. Among the 10 volunteers, one user is regarded as the current legal user of the smartphone, while the other nine users are attackers. In the model training stage after data augmentation, the current legal user's data is labeled as 1, while the same amount of data randomly sampled from all nine other users is labeled as 0. These two parts constitute the current legal user dataset. When we train the current legal user's model, 70% of the dataset is training set, 10% is used as the validation set, and the last 20% is for testing.

7.3. Authentication accuracy

Overall Performance. We used the current target user's data collected on the west side as positive samples and all other users' current direction data as negative in the training phase. Because the system aims to authenticate users, it can be attributed to a one-classification problem that uses an end-to-end deep neural network for automatic feature extraction and fusion, avoiding the manual feature extraction work of the traditional OC-SVM (One-Class SVM). Therefore, for each system user, a corresponding individual model is trained to constitute a model library. The overall confusion matrix is shown in Fig. 10. As it can see that the minimum certification accuracy is 92.44%, and the overall average accuracy reaches 96.86%. It is worth noting that this confusion matrix is different from the traditional one. The x-axis represents different users, while the ordinate represents different individual models. Each row in the figure denotes the predicted probability produced after inputting different user data into models. The output probability of user 10 of model 8 reaches 30.37%, but it is far away from the probability of being incorrectly authenticated as user eight and is not a false accept. To better represent the balance of precision

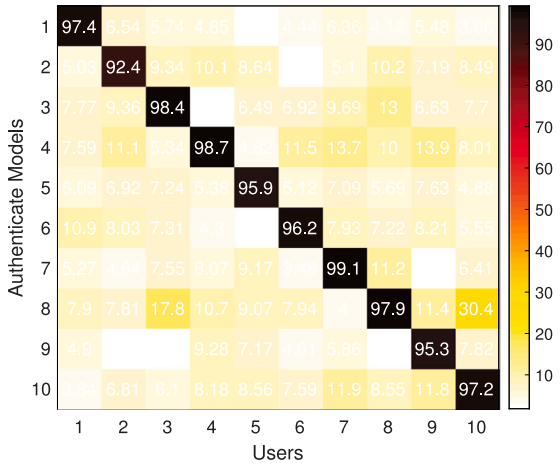


Fig. 10. The fusion matrix of overall accuracy.

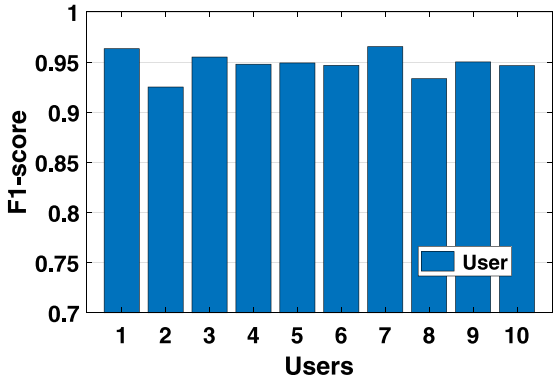


Fig. 11. The F1-score of *WiCapose*.

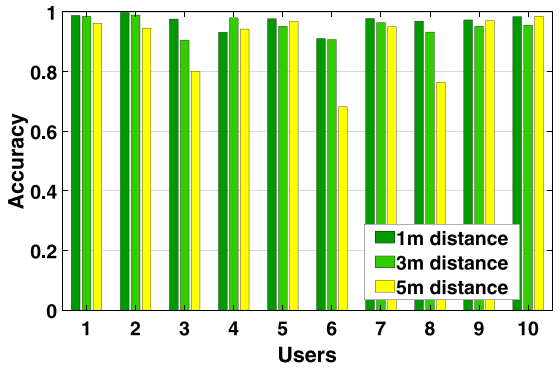


Fig. 12. Accuracy at different distance.

and recall, we employed the F1-score, as shown in Fig. 11. The highest F1-score is 96.5%, and the lowest is 92.5%, while the average is 94.8%.

Impact of Distance. Different distances between the transceiver have different effects on the CSI. The data collected at different distances of 1 m, 3 m, and 5 m are divided into three parts and input into the network to train and test the model. Because it is the data from the same direction, even if it crosses different Fresnel boundaries, the network can extract the finger behaviors' unique features. As shown in Fig. 12, it can be observed that the accuracy of the test at 1 m distance is the highest, that is when the direction is the same, the quality of the data collected at a close distance is the best, which can fully reflect the features of the subject's finger behaviors. As the distance increases, the

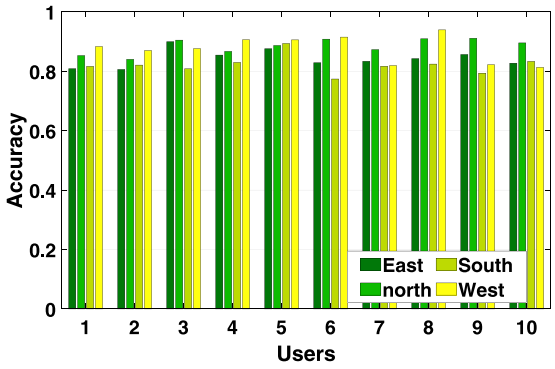


Fig. 13. Accuracy at different directions.

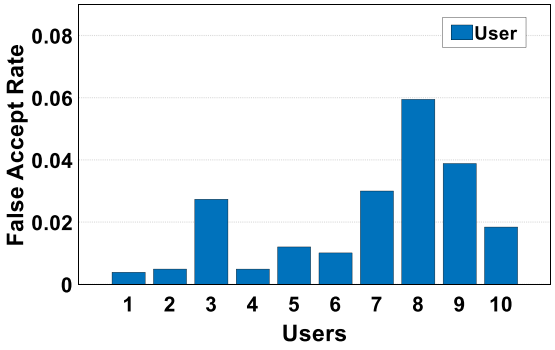


Fig. 14. Resistance capability to mimic attack.

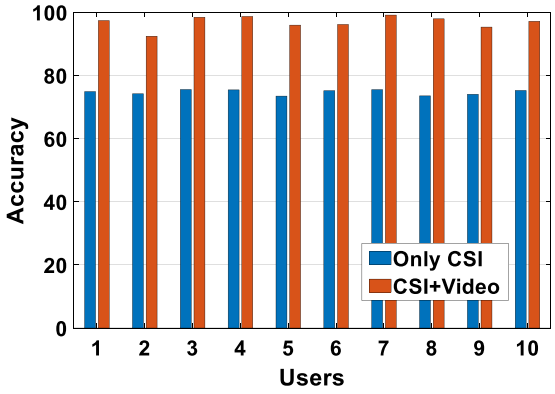


Fig. 15. Single factor vs. two factors.

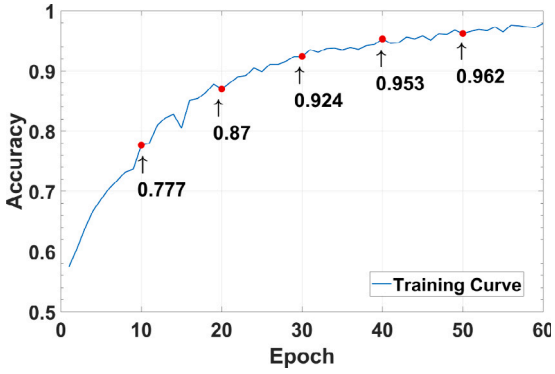


Fig. 16. Accuracy in different training epoch.

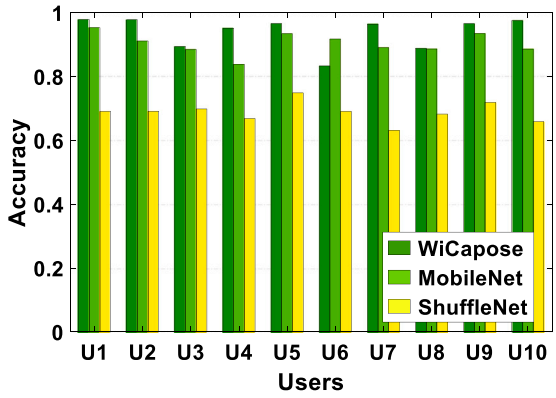


Fig. 17. Accuracy in different backbone NN.

Table 1
Accuracy of holding phone in one-hand.

Users	1	2	3	4	5
Accuracy	0.8517	0.8446	0.8695	0.8627	0.8533
Users	6	7	8	9	10
Accuracy	0.8580	0.8619	0.8556	0.8559	0.8450

signal attenuates, and the action features become less notable, with the prediction accuracy of the model decreasing.

Impact of Different Directions. CSI is sensitive to the user's actions and location. According to the Fresnel zone theory, the effect of the CSI signal generated by crossing through the Fresnel zone at different positions is varied. In the experiment, the smartphone is the receiver, and all tap behaviors are performed on it. When people sit in different directions, finger movements cross different Fresnel zones to produce different CSI reception signals. As shown in Fig. 13, it can be observed that the authentication accuracy of 10 volunteers in four different directions, the accuracy reach highest, 93.94%, when the subject sits on the west of the table, and decreases in the order of north, east, and south. The principal reason is that the WiFi antenna is in the lower part of Nexus 5, so if volunteers hold the smartphone in the left hand and sit on the west and north sides of the table, the effects of crossing Fresnel boundary is the most prominent. In the east, the Fresnel zone's outer side is cross, and the accuracy is slightly reduced. When sitting on the south side of the table, the human body blocks the signal transmission, making the CSI signal of the finger movement insignificant, resulting in a decrease of accuracy.

Impact of Hold Posture. In daily life, people holding their smartphones with one hand are also typical scenes. It is necessary to analyze the authentication accuracy when holding a smartphone in one hand. Volunteers sit 1 meter away from the transmitter, collect authentication data with one hand with Nexus 5, and feed into the same FC neural network for training. The results are shown in Table 1, and the average accuracy is 85.58%. The main reason is that when the smartphone is operated on one hand, the smartphone needs to be flipped at a large angle due to palm size and finger length limitation. At this time, the obtained smartphone rotation and translation matrices' quality is dropped.

7.4. Security and effectiveness

Resilience to Mimic Attack. Mimicry attacks are the most prevalent attack method. When an attacker witnesses the entire process of the victim's system authentication, the first step is to imitate the attacker's behavior to attempt a system intrusion. Traditional 1-factor authentication cannot effectively prevent such attacks, but *WiCapose* can effectively block them. It is assumed that the attacker can observe the entire process of regular user authentication and record the user's input process and behavior pattern. To verify the system's effectiveness, let a volunteer as the target user, and other users imitate the target user's action features to carry out trial attacks on the system. Fig. 14 shows the average false accept rate.

Table 2
Total training time of models.

Backbone	MobileNet	ShuffleNet	Ours
Trainning time	1 h 46 m	22 m	12 m

It indicates that even if the user's entire input process is recorded, the intrusion success probability is below 2%. Even if the user's gender and behavior pattern are similar, the imitating attack's success probability is only 6%.

Ablation Experiment. To verify that two-factor authentication is sufficient, we independently use the CSI data collected in the experiment to authenticate users. As shown in Fig. 15, the accuracy reaches 75.5% by just using CSI. While using both CSI and Video information for authentication, the accuracy reaches 99.12%. The CSI encourages mimicry-resistant micro-scale behaviors sensing, while pose information extracted from video serves as reinforcement in CSI modal under environment-dependent noise and assistant in enhancing security and thus preventing additional attacks. Using nexmon and Nexus 5 smartphones to collect CSI data, the signal's quality and stability are moderate due to software and hardware defects. With different distances, different directions, and angles, the features of CSI are different. Using the FC neural network to extract features of the signal, the available feature information is limited. It is not sufficient to use CSI to characterize the user's behavior features alone so that the authentication accuracy is much lower.

Impact of Train Epoch. The epoch of training has an important impact on the accuracy of the model output. As the training epoch increases, the accuracy also increases in the case of a fixed dataset. As shown in Fig. 16, the overall accuracy increase has a monotonically increasing relationship with epoch. It can be observed that the accuracy increased rapidly before 30 epochs. After that, the accuracy growth trend slowed down. When it reaches 60 epochs, the accuracy reaches 97.9%. Our model chooses to train 60 epochs based on practice to achieve a balance between time and accuracy.

Impact of Backbone Neural Network. We analyze the effectiveness of the FC Neural Network (NN) model. We chose two other regularly used classic neural network models, MobileNet V2 [33] and ShuffleNet [34], both of which can be deployed on smart mobile devices. We use the same data, containing both CSI and video calibration matrices, and train from scratch to obtain certification results. It can be observed from Fig. 17 that most of the authentication accuracy of our model is higher than that of MobileNet and ShuffleNet. It is mainly because ShuffleNet has fewer parameters and cannot well express data features. As mentioned earlier, there is no distinct contour feature in the data like in the picture, so the convolution operation used by MobileNet cannot extract contour features from the matrices data. But user 6 has a large change in motion when holding the device, and the accuracy is lower than that of MobileNet. Considering the power consumption and real-time issues of smartphone terminal equipment, considering the duration of the training, our model training time is much less than the others. As we can see from Table 2, *WiCapose* training only takes 12 min, while MobileNet takes 1 h and 46 min for training 40 epochs, and ShuffleNet takes 22 min. It shows that our model's deployment efficiency is better than the others. It is worth noting that, like many present deep learning models, although model training takes a long time, once the deployment is completed, the inference time is very fast.

8. Conclusion

In this paper, we propose *WiCapose*, which fuses inimitable non-copyable location-dependent CSI side-channel information and camera

pose information for two-factor authentication without additional devices and steps. The CSI is collected through the nexmon kernel model, combined with the pose information extracted by the ORB algorithm, and fed into a deep neural network. It can fuse both side-channel data and extract the user finger behaviors' unique features for authentication, effectively preventing multiple attacks and strengthening the security capabilities of systems such as the mobile scenarios.

CRediT authorship contribution statement

Zhuo Chang: Conceptualization, Methodology, Software, Investigation, Formal analysis, Writing – original draft. **Yan Meng:** Data curation, Writing – original draft. **Wenyuan Liu:** Visualization, Investigation. **Haojin Zhu:** Resources, Supervision. **Lin Wang:** Conceptualization, Funding acquisition, Resources, Supervision, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Sharif M, Bhagavatula S, Bauer L, Reiter MK. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. New York, NY, USA: Association for Computing Machinery; 2016, p. 1528–40.

[2] Li Y, Xu K, Yan Q, Li Y, Deng RH. Understanding OSN-based facial disclosure against face authentication systems. In: Proceedings of the 9th ACM symposium on information, computer and communications security. New York, NY, USA: Association for Computing Machinery; 2014, p. 413–24.

[3] Wang Y, Cai W, Gu T, Shao W. Your eyes reveal your secrets: An eye movement based password inference on smartphone. IEEE Trans Mob Comput 2020;19:2714–30.

[4] Diao W, Liu X, Zhou Z, Zhang K. Your voice assistant is mine: How to abuse speakers to steal information and control your phone. In: Proceedings of the 4th ACM workshop on security and privacy in smartphones & mobile devices. 2014. p. 63–74.

[5] Coli P, Marcialis GL, Roli F. Fingerprint silicon replicas: Static and dynamic features for vitality detection using an optical capture device. Int. J. Image Graph. 2008;08:495–512.

[6] yubico. Yubico. 2021, URL https://www.yubico.com/.

[7] Niu K, Zhang F, Jiang Y, Xiong J, Lv Q, Zeng Y, et al. WiMorse: A contactless morse code text input system using ambient WiFi signals. IEEE Internet Things J 2019;6:9993–10008.

[8] Wu C, Zhang F, Hu Y, Liu KJR. GaitWay: Monitoring and recognizing gait speed through the walls. IEEE Trans Mob Comput 2021;20:2186–99.

[9] Zeng Y, Wu D, Xiong J, Liu J, Liu Z, Zhang D. MultiSense: Enabling multi-person respiration sensing with commodity WiFi. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020;4.

[10] Xie Y, Xiong J, Li M, Jamieson K. MD-Track: Leveraging multi-dimensionality for passive indoor Wi-Fi tracking. In: The 25th annual international conference on mobile computing and networking. New York, NY, USA: Association for Computing Machinery; 2019.

[11] Yue S, Yang Y, Wang H, Rahul H, Katabi D. BodyCompass: Monitoring sleep posture with wireless signals. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020;4.

[12] Tian Y, Lee G-H, He H, Hsu C-Y, Katabi D. RF-Based fall monitoring using convolutional neural networks. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018;2.

[13] Zhang L, Wang C, Zhang D. Wi-PIGR: Path independent gait recognition with commodity Wi-Fi. IEEE Trans Mob Comput 2021;1. http://dx.doi.org/10.1109/TMC.2021.3052314.

[14] Fan L, Li T, Fang R, Hristov R, Yuan Y, Katabi D. Learning longterm representations for person re-identification using radio signals. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020.

[15] Liu J, Chen Y, Dong Y, Wang Y, Zhao T, Yao Y-D. Continuous user verification via respiratory biometrics. In: IEEE INFOCOM 2020 - IEEE conference on computer communications. 2020. p. 1–10.

[16] Zhou B, Lohokare J, Gao R, Ye F. EchoPrint: Two-factor authentication using acoustics and vision on smartphones. In: Proceedings of the 24th annual international conference on mobile computing and networking. New York, NY, USA: Association for Computing Machinery; 2018, p. 321–36.

[17] Meng Y, Wang Z, Zhang W, Wu P, Zhu H, Liang X, et al. Wivo: Enhancing the security of voice control system via wireless signal in IoT environment. In: Proceedings of the eighteenth ACM international symposium on mobile ad hoc networking and computing. 2018. p. 81–90.

[18] Douhou S, Magnus JR. The reliability of user authentication through keystroke dynamics. Stat Neerl 2009;63:432–49.

[19] Sanchez-Reillo R, Sanchez-Avila C, Gonzalez-Marcos A. Biometric identification through hand geometry measurements. IEEE Trans Pattern Anal Mach Intell 2000;22:1168–71.

[20] Li M, Meng Y, Liu J, Zhu H, Liang X, Liu Y, et al. When CSI meets public WiFi: Inferring your mobile phone password via WiFi signals. In: Proceedings of the 2016 ACM SIGSAC conference on computer and communications security. New York, NY, USA: Association for Computing Machinery; 2016, p. 1068–79.

[21] Chen H, Li F, Du W, Yang S, Conn M, Wang Y. Listen to your fingers: User authentication based on geometry biometrics of touch gesture. Proc ACM Interact Mob Wearable Ubiquitous Technol 2020;4.

[22] Li H, Yang W, Wang J, Xu Y, Huang L. WiFinger: Talk to your smart devices with finger-grained gesture. In: Proceedings of the 2016 ACM international joint conference on pervasive and ubiquitous computing. New York, NY, USA: Association for Computing Machinery; 2016, p. 250–61.

[23] Qian K, Wu C, Yang Z, Liu Y, Jamieson K. Widar: Decimeter-level passive tracking via velocity monitoring with commodity Wi-Fi. In: Proceedings of the 18th ACM international symposium on mobile ad hoc networking and computing. New York, NY, USA: Association for Computing Machinery; 2017.

[24] Roth V, Richter K, Freidinger R. A PIN-Entry method resilient against shoulder surfing. In: Proceedings of the 11th ACM conference on computer and communications security. New York, NY, USA: Association for Computing Machinery; 2004, p. 236–45.

[25] Yu S, Guo S, Stojmenovic I. Fool me if you can: Mimicking attacks and anti-attacks in cyberspace. IEEE Trans Comput 2015;64:139–51.

[26] Schulz M, Wegemer D, Hollick M. Nexmon: The C-based firmware patching framework. 2017, URL https://nexmon.org.

[27] Zhang F, Zhang D, Xiong J, Wang H, Niu K, Jin B, et al. From Fresnel diffraction model to fine-grained human respiration sensing with commodity Wi-Fi devices. Proc ACM Interact Mob Wearable Ubiquitous Technol 2018;2.

[28] Dragomiretskiy K, Zosso D. Variational mode decomposition. IEEE Trans Signal Process 2014;62:531–44.

[29] Zhang F, Wu C, Wang B, Lai H-Q, Han Y, Liu KJR. WiDetect: Robust motion detection with a statistical electromagnetic model. Proc ACM Interact Mob Wearable Ubiquitous Technol 2019;3.

[30] Mur-Artal R, Montiel JMM, Tardós JD. ORB-SLAM: A versatile and accurate monocular SLAM system. IEEE Trans Robot 2015;31:1147–63.

[31] Mur-Artal R, Tardós JD. ORB-SLAM2: An open-source SLAM system for monocular, stereo, and RGB-D cameras. IEEE Trans Robot 2017;33:1255–62.

[32] Jung AB, et al. Imgaug. 2020, https://github.com/aleju/imgaug. Online; [Accessed 1 February 2020].

[33] Sandler M, Howard A, Zhu M, Zhmoginov A, Chen L-C. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018.

[34] Zhang X, Zhou X, Lin M, Sun J. Shufflenet: An extremely efficient convolutional neural network for mobile devices. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018. p. 6848–56.