ThumbUp: Identification and Authentication by Smartwatch using Simple Hand Gestures

Xiaojing Yu*, Zhijun Zhou*, Mingxue Xu*, Xuanke You*, Xiang-Yang Li†
*†*School of Computer Science and Technology*, *University of Science and Technology of China*, Hefei, China
*{ yxjing, zhouzj18, xmx18, yxkyong }@mail.ustc.edu.cn, †xiangyangli@ustc.edu.cn

***Abstract*—The widespread creative application and smart devices call for convenient and secure interaction with human users. We propose, design, and implement a smartwatch-based two-factor real-time identification and authentication system named ThumbUp, where smartwatch users can identify and authenticate themselves by some *simple* hand and finger gestures, such as thumb-up. ThumbUp leverages the signal collected from the Inertial Measurement Unit (IMU) in Commercial Off-The-Shelf (COTS) smart devices and discovers the unique fingerprint pattern produced by each user's simple hand gestures using a carefully crafted deep learning model. We implement our system and conduct extensive experiments to evaluate its efficacy and efficiency with 65 different users over a period of more than 3 months. It reaches an accuracy of 97% for identification, and EER 0.014 for authentication using only one simple gesture. We also survey the users' acceptance of our system and discuss how the proficiency of gestures affects authentication accuracy.**

***Index Terms*—Smartwatch, Authentication, Hand Gestures**

I. Introduction

The widespread smart wearable devices have provided a fast and convenient way to interact with the physical world. Specifically, global smartwatch shipments are forecast to reach 113 million units by 2022. Smartwatches provide various services such as instant messaging, online shopping and quick mobile payment. While smartwatches provide users convenience, they also pose a looming threat to impinge on users security [3]. Therefore, it is crucial to implement reliable and convenient identification and authentication schemes for smart wearable devices.

Existing authentication methods have a different emphasis on security and usability. The traditional methods, for example, PINs lack both *usability* (inconvenient to input passwords on smart devices with small-sized screen) and *security* (shoulder-surfing [30] attacks). Another popular method, face recognition, may be bypassed using images or counterfeit videos [4]. While other methods (*e.g.*, iris and fingerprint) are more secure but require additional specialized sensors. Recently, researchers also proposed many different authentication schemes by exploiting users' behavior patterns and motion habits, *e.g.*, gait [28], head movements [16], handwriting [1], and even tongue movement [22]. Compared with other means of authentication, behavior-based authentication methods may provide weaker security but offer greater convenience. Among these behaviors, gestures as biometric authentication have attracted considerable interest. Much work using IMU [12], [34] so far has focused on large motion patterns by arm movements for distinguishing different users. Using large motion patterns for authentication could be inconvenient in many daily applications. Furthermore, it could suffer the risk of being impersonated as attackers can observe and learn these motions.

Fig. 1. Illustration of nine gestures studied in ThumbUp (from top, left to right): G1(finger-snapping), G2(finger-turning in circles), G3(beckon), G4(hand-waving), G5(fist-making), G6(victory-gesture), G7(gun-gesture), G8(thumb-up), G9(finger-bending).

We found that current solutions cannot achieve usability, security, and user-friendliness simultaneously. In this work, we design schemes for convenient and reliable identification and authentication for wearable devices by using simple hand gestures which are mainly performed by fingers, including turning fingers or thumbing up (as shown in Fig. 1). Furthermore, the training phase of our model only requires a small number of samples (*e.g.*, 10 samples in our system) for each user to improve usability. Specifically, we need to address several critical technical challenges:

- **Limited Training Set:** The motion signals of tiny gestures collected by IMU are significantly weaker than that by arm movements. Also, asking a user to repeat the gesture many times in the user-training phase is not user-friendly. We aim to limit the number of training samples to 10. It is extremely challenging to extract useful features that can uniquely represent each user from weak signals with the limited samples. Furthermore, the valid features from tiny hand/finger movement could be buried in the inherent noises produced by the IMU

978-1-7281-4657-7/20/$31.00 ©2020 IEEE

sensors.

- **Reliability and Robustness:** It is necessary that our system can classify different users, authenticate legitimate users accurately and defend against attackers that may attempt to impersonate the legitimate users maliciously. Thus, the features adopted by our system for authentication should embrace both *diversity* of different users and *consistency* for the same user. As behavior biometrics, the gestures of users change slightly after a long period inevitably. For better usability, the system should be adaptive to slight changes in the hand gesture by legitimate users.
- **Energy-efficiency and Real-time Ability:** We need to implement a light-weight identification/authentication scheme with the limited storage and computing power of smartwatches, while simultaneously maintain high stability and real-time ability that required by a convenient system.

To meet these challenges, we design, implement and evaluate **ThumbUp**, a two-factor identification and authentication system that can authenticate users by a tiny gesture like thumb up. We analyze the anatomy of hand movement in human kinematics. Meanwhile, we explore the stability and diversity of the motion sensor signals with the electromyography (EMG) signal as auxiliary verification. After motion signals pre-processing and detection, we design a novel light-weight deep neural network model with multilayer Bidirectional Long Short-Term Memory (BiLSTM) and an attention mechanism for automatic feature extraction and classification. Through a prototype implementation on COTS wearable devices and evaluation on 65 participants, we demonstrate that ThumbUp can precisely identify users with a mean accuracy exceeding 95.7% and correctly authenticate a legitimate user with a low mean error rate of 0.025.

To the best of our knowledge, ThumbUp is the first system to use simple finger-movement gestures for user identification/authentication with IMU on COTS smartwatches. We expect that ThumbUp has potential applications in (1) giving access to smart wearable devices, (2) quick payment through simple interaction, (3) operating mobile devices privately and reliably.

To summarize, our contributions are as follows:

- We design and implement a reliable authentication scheme for wearable devices using tiny gestures. We explore the feasibility of hand-gesture-based biometrics as certification factors and demonstrate that hand gestures contain unique signatures of users. We design the model which extracts features and classifies gesture patterns of users. Moreover, we propose a self-calibration and transfer learning method to enhance the practicability and validity.
- We evaluate ThumbUp through comprehensive experiments over different system design parameters which last for 3 months and involve 65 participants. The experiments show that even finger movements like victory-gesture can generate accurate identification results. Furthermore, we test the security of our system under imitate attacks, which shows the system can resist such attack with a mean EER of 0.025.
- In consideration of friendliness, the gestures utilized in our system are well designed based on the study of biological kinematics mechanisms. We investigate the participants about the convenience of the gestures. Then we determine the recommended gesture choices combining both the authentication performance and participants' views.

The remainder of the paper is organized as follows. We present the basis of hand movement and feasibility study in Sec. II. In Sec. III, we give a brief overview of the main design of ThumbUp. The details of data pre-processing are illustrated in Sec. IV, followed by the model description in Sec. V and calibration mechanism of the system in Sec. VI. We present experimental evaluation results and user study in Sec. VII. We review the related work in Sec. VIII and end up with the conclusion in Sec. IX.

II. Basis of Hand Movement & Feasibility Study

In this section, we discuss the basic biological kinematics to demonstrate the feasibility of hand movements as unique features theoretically, explore the stability and diversity of the motion sensor signal, and further utilize the EMG for auxiliary verification.

A. Hand Movement's Anatomy

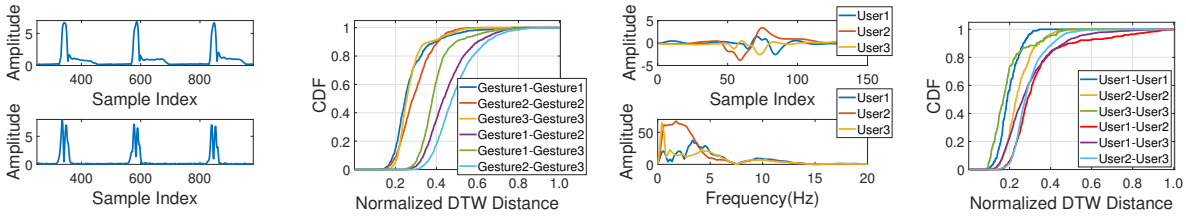
Subtle movements are intuitively easier to be imitated. Nevertheless, the muscle movements are controlled by the subconscious, and it is hard to be consciously modified. Even if two users complete very similar movements, the biological kinematics mechanisms of muscles are fairly distinguishable, which allows identification based on small gestures. Besides, being an internal part of the hand surface, muscles are quite stable concerning changes in humidity and temperature [13].

The forearm muscles inside the position where we wear the smartwatches, act upon hands. The bulk of these muscles form the fleshy roundness of the forearm, with tendons extending into the wrist and hand. Movements of the hand are controlled by the hand itself (intrinsic muscles) as well as muscles within muscles in the forearm (extrinsic muscles) allowing excellent control of precise movements and powerful movements [19]. The motion signals would be perceived by a motion sensor on the forearm. It was shown in [32] that the forearm muscles are good representations of the hand movements and finger gestures.

Moreover, the small gesture without arm movements involves tiny jitters of people's peculiar habits. Thus, both the biological and behavior patterns of muscles would be captured in motion signals as a kind of authentication information.

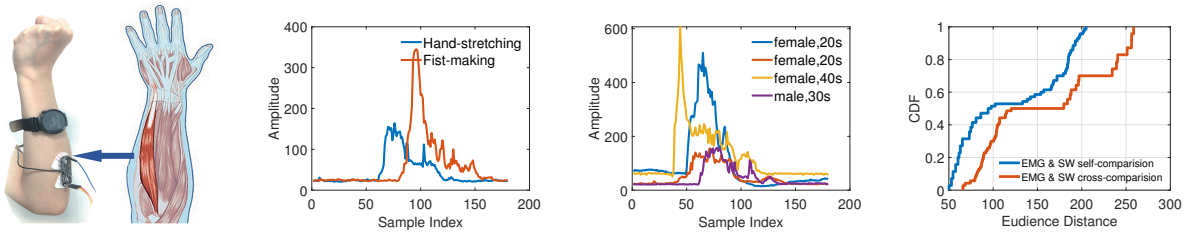
B. Feasibility Study

To gain a better understanding of the feasibility of ThumbUp, we explore the diversity, consistency, and uniqueness of gesture motion signals.



(a) (b) (c) (d)

Fig. 2. The motion signals of (a)different gestures and (c)different users in the time and frequency domains, the DTW distance among (b)gestures and (d)users.



(a) (b) (c) (d)

Fig. 3. (a) placement of EMG measuring devices, (b) EMG signals of Hand-stretching and Fist-masking for same user, (c) EMG signals of 4 people (3 female, 1 male) over 5s, (d) CDF diagram of Euclidean distance between EMG and motion signals.

Diversity & Consistency of Motion Signals: Firstly, we have asked one participant to perform two different hand movements 10 times for each gesture with a smartwatch. In Fig. 2(a), we observe that the profiles for separate gestures differ significantly. To illustrate the difference digitally, we compute the normalized Dynamic Time Warping (DTW) distance of the same gesture and different gestures (shown in Fig. 2(b)). The figure clearly shows that it is feasible to differentiate between different gestures. Meanwhile, the result (shown in Fig. 2(a)) reveals that motion signals from repetitions of the same gesture are fairly similar, which demonstrates the consistency of gesture motion signals.

Uniqueness of Motion Signals: We intend to find out whether motion signals of the same gesture generated by different users are distinct. Three participants are asked to snap fingers for 20 times. From Fig. 2(c), the profile differs significantly both in time and frequency domain. Similarly, we calculate the nomalized DTW distance (the result is shown in Fig. 2(d)), and it reveals that motion signals are unique for different users.

C. Motion Signal Correlation with EMG

We aim to explore the connection between muscle movement and motion signals more intuitively. However, muscle movements for gestures cannot be digitalized directly. The EMG signal is superimposed action potential generated from the contraction of muscle fibers, which can reflect muscle movement more directly [20]. Hence, we verify the correlation between muscle movements and motion signals by analyzing the relationship between EMG signals and motion sensor signals.

We place the EMG sensor on the center of the measured muscle belly between an innervation zone and the distal tendon for better accuracy, as shown in Fig. 3(a). We ask volunteers to try their best to stretch their five fingers and make fists for activating the muscle. The signal waveforms of these two gestures are depicted in Fig. 3(b). The figure shows that the signals between different gestures are fairly different. To emphasize the individual differences, we ask 4 volunteers of different ages and genders to stretch fingers. EMG results are shown in Fig. 3(c). Significant differences exist in EMG signals among diverse people, even with the same age and gender.

To obtain the correlation of EMG and motion signal, we ask four participants to stretch fingers and make fist 20 times with EMG device and the smartwatch simultaneously to obtain quantitative description (as shown in Fig. 3(a)). We use $S = \sqrt{G^T G + L^T L}$ to describe the motion sensor signal, where G, T denote the integration of angular acceleration and linear acceleration respectively. Similarly, we calculate the integration of EMG signal noted as E, and we quantify the similarity between S and E by DTW. The result is shown in Fig. 3(d), where self-correlation is the similarity of two types of signals for the same person and the cross-correlation is for different people. The distance of cross-correlation (AVG=159.11) is larger than self-correlation (AVG=120.01). It can be derived from these statistics that there is a certain similarity between the motion signal and EMG signal from the same person, and the similarity decreases when the signal source changes. Considering the individual difference of EMG, the results demonstrate the feasibility of motion signals as unique conditions for certification.

III. Design Scope & Overview

In this section, we describe our design scope and the system overview of ThumbUp.

A. Objective and Design Scope

We divide the certification task into two parts: identification and authentication. **Identification** represents multi-user classification tasks; the main goal is to classify different users. **Authentication** represents one-to-one identification

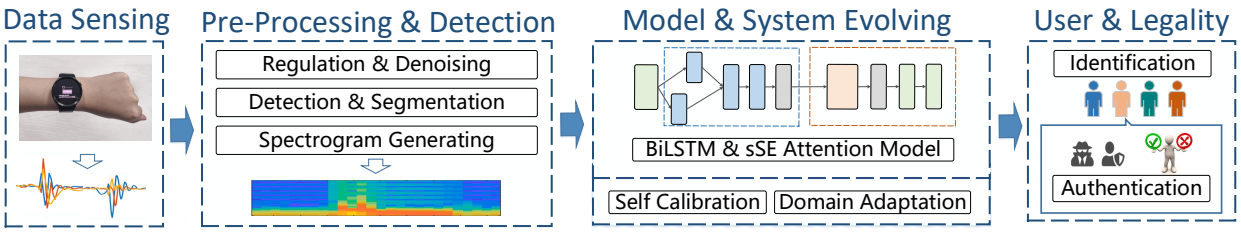


Fig. 4. System Workflow and Key Components of ThumbUp

with malicious attacks; the main goal is to distinguish between malicious and legitimate users correctly.

Gesture Design: We prefer to implement our system in situations that require daily authentication of device owners, such as smartwatch unlocking. The gestures that our system adopts need to be convenient enough. The sophisticated gestures that include movements of fingers, palm, and wrist at the same time involve more information to discriminate, but they are much less convenient. Meanwhile, too simple gestures that only contain weak movement such as slightly waving one finger are ineffective in uniquely identifying a user. We define nine typical gestures (shown in Fig. 1), considering user-friendliness according to user preference, and the trade-off between complexity and distinguishability. In the following experiments, we test the users' impressions of these gestures for authentication and give recommendations for improvement.

Availability: Our system needs to extract *consistent* and *distinguishable* bio-metrical signal features from subtle motion signals, *i.e.*, the features have to be persistent to meet the needs of long-term usage, and high diversity to be resistant to different forms of attacks. Moreover, our system should contain classifier mechanisms and give the corresponding results accurately for two tasks.

Universal: For universal requirement, we prefer to use typical commercial smartwatches with accelerometers and gyroscopes. As COTS smart-devices often have limited computing and storage resources, lightweight design for our system is in need.

Our goal is to build a highly secure and reliable real-time authentication system based on 3D simple hand and finger gestures using commercial smart-devices without any additional hardware.

B. Overview of System

As shown in Fig. 4, ThumbUp consists of three parts: The first part is *pre-processing & detection* (Sec. IV), which aims to remove noises, segment, and extract sequential features from continuous motion signals. The second part is *model description* (Sec. V), which extracts representations from spectrograms with carefully crafted deep learning methods. Also, we describe our design of the classification algorithm. The third part is the *system evolving* (Sec. VI), which introduces the strategy for continuous model evolving to cope with the behavioral changes of user and system initialization for domain adaptation.

IV. Pre-processing & Detection

As the signals collected from motion sensors are noisy, incomplete, and even erroneous, the first step of ThumbUp is to filter the noise and segment the signal to match the actual gestures.

A. Data Regulation & Denoising

To ensure uniform sampling of the accelerometer and gyroscope, we interpolate the data to $100Hz$ of the sampling rate. We use the Z-score normalization technique to normalize the amplitude of the signals. The processed signals obey the standard normal distribution, with the mean value 0 and the standard deviation 1. Afterwards, we reduce random noise by a Savitzky-Golay smoothing filter [5], also called least-square smoothing filter. The basic idea behind this filter is to find a least-square fit with a polynomial of high degree for each data point, over an odd sized window centered around that data point [21], which not only reduces noise but also maintains the shape and height of waveform peaks.

B. Detection and Segmentation

IMU continuously collects motion sensor signals; we need to detect the possible samples and segment signals into a given size. One common way is to set a constant threshold empirically, and the part of the signal whose short-term energy exceeds this threshold would be regarded as a gesture. However, the threshold is hard to choose with different noise levels in realistic scenarios.

We use a method similar to the Constant False Alarm Rate (CFAR) algorithm [37] to detect the gestures. The main idea is to determine the start and endpoint of one single gesture by dynamic thresholds. We use X to denote the long time-series signal, and $x(i)$ is the square root of the squared sum for six axes collected from the accelerometer and gyroscope at the i_{th} sample index. Let W denote the sliding window size, set as 128. Besides, $E(i)$ and $D(i)$ denote the average power and standard deviation at the i_{th} sample index, respectively. Here $E(i) = \frac{1}{W} \times \sum_{k=i-W+1}^{i} x(k)^2$, and $D(i) = \sqrt{\frac{1}{W} \times \sum_{k=i-W+1}^{i} (x(k)^2 - E(i))^2}$. Thus, a potential start point of a gesture is detected if $x(i)^2 > E(i) + \gamma_1 \times D(i)$ and a potential endpoint is detected if $x(i)^2 < \gamma_2 \times \bar{E}$, among them, γ_1 and γ_2 are both the constant, $\bar{E}$ is the average noise power detected before the first gesture.

The segmentation result is depicted in Fig. 5, and it shows satisfactory efficiency. The orange line represents the motion signals we concern.

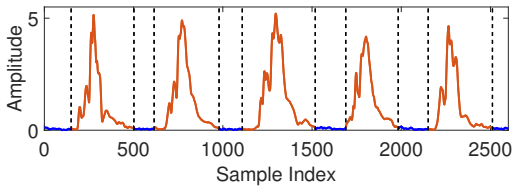


Fig. 5. Gesture Detection and Segmentation

C. Spectrogram Generating

For motion sensor signals, time-domain features reflect the sequential relationship inside gestures, while frequency features reflect different motions of muscles in hand. The Short Time Fourier Transform (STFT) [7] is a widely used tool for signal processing and has sensitivity in the time domain for both high and low-frequency signal, contains more frequency and time-domain features for next step than Discrete Wavelet Transform (DWT). We use the power band produced by STFT to represent the frequency spectrum. The spectrogram is the magnitude squared of the STFT $|X(m,w)|^2$. For a given signal $x[n]$ and a window function $w[n]$, the discrete-time STFT of $x[n]$ is calculated as $X(m,w) = \sum_n x[n]w[n-m]exp(-jwn)$. Hamming window is applied for the window function $w[n]$. Evaluating $X(m;w)$ for more $(m;w)$ points will provide high-resolution information with decreased overall information and increased computation overhead. We achieve a satisfactory trade-off by experiments (shown in Sec. VII). We observe that the vibration caused by human mobility is mostly less than 15 Hz, a 17 Hz cut-off frequency for F is sufficient to retain enough information for the motion sensor signal. We concatenate all channels and generate spectrograms to represent the signals in the high dimension.

V. Model Description

Next, we propose a deep neural network (as shown in Fig. 6) to extract subtle and stable representations from spectrograms and classify users for identification and authentication. The model is composed of four parts: spectrogram input layer, BiLSTMs, sSE Networks, and classifier layer. Firstly, the input data of the model is the gesture spectrograms obtained by STFT. Then, a three-layer BiLSTMs are used to extract motion features. For improving the model performance, we add an attention mechanism with Squeeze-and-Excitation Networks to significant aggregate information from the motion representations generated by the BiLSTM layers. Finally, we adopt a Multilayer Perceptron followed by a softmax activation as the classifier in our model. Moreover, we perform ablation studies (in Sec. VII-B2) to understand aspects of the proposed model architecture, which results in an improvement.

A. BiLSTM Layer

Owing to the structural property, Recurrent Neural Networks (RNNs) maintain the memory based on historical information, which is suitable for processing sequential data [18]. Long Short-Term Memory (LSTM) is explicitly designed to address the long-term dependency problem through

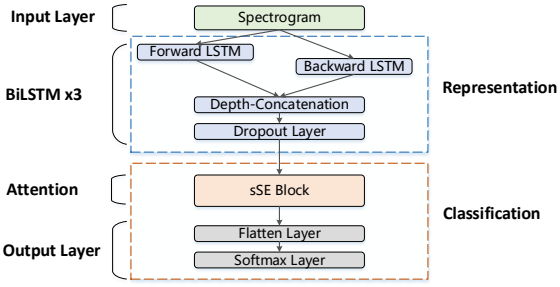


Fig. 6. Model Architecture

purpose-built memory cells, which performs better in longer sequences. For the spectrogram of the sequential timing signal, BiLSTMs [6] have stronger representation extraction ability than LSTM with both previous and subsequent information.

First, we get the input spectrogram $\mathbf{s} = [\mathbf{s}_1, ..., \mathbf{s}_T], s_t \in \mathbb{R}^d$ from STFT. The BiLSTMs layer computes the forward hidden sequence $\overrightarrow{\mathbf{h}}$, the backward hidden sequence $\overleftarrow{\mathbf{h}}$ and the output sequence $\mathbf{h_t}$ by iterating the backward layer from $t = T$ to 1, the forward layer from $t = 1$ to T. Then the layer updates corresponding hidden states at each time-step:

$$\overrightarrow{\mathbf{h}}_t = \overrightarrow{LSTM_F}(\overrightarrow{\mathbf{h}_{T-1}}, \mathbf{s}_t), \overleftarrow{\mathbf{h}}_t = \overleftarrow{LSTM_B}(\overleftarrow{\mathbf{h}_{T-1}}, \mathbf{s}_t), \quad (1)$$

where $\overrightarrow{LSTM_F}$ and $\overleftarrow{LSTM_B}$ in our system are implemented by the composite function proposed in [8].

Next, these hidden state outputs from the forward LSTM $\overrightarrow{\mathbf{h}}_t$ and backward LSTM $\overleftarrow{\mathbf{h}}_t$ are concatenated at every time-step to enable encoding of information from past and future contexts respectively. Especially since we only have a small number of training samples, models will easily overfit on these samples. In order to prevent complex co-adaptations on training samples and perform model averaging with networks, we feed these concatenated hidden states to the dropout layer that temporarily discard neural network units from the network with a certain probability. Then a Batch Normalization layer is adopted for avoiding the gradient disappearance and explosion in the process of *backpropagation* and makes the updating steps of different scales more consistent.

B. Attention Layer

Convolutional Neural Networks (CNNs) take advantage of convolutions, which help to extract relevant information at a low computational cost. We can treat the features vectors generated by BiLSTMs as an image. For image segmentation, the pixel-wise spatial information is informative. In order to improve the representational ability of the features, we add an attention mechanism with Channel Squeeze and Spatial Excitation Block (sSE) [26] network to the model instead of using CNNs directly, which 'squeezes' the feature along the channels and 'excites' spatially (shown in Fig. 7).

We note the output feature map generated by representation block as $\mathbf{U} \in \mathbb{R}^{H \times W \times C}$. First, sSE slices the input tensor

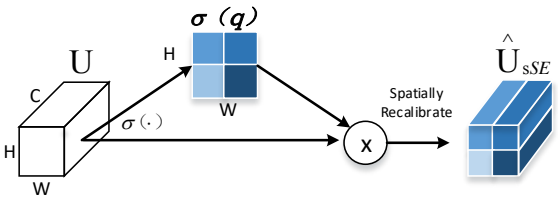


Fig. 7. Channel Squeeze and Spatial Excitation Block

$\mathbf{U} = [\mathbf{u}^{1,1}, ..., \mathbf{u}^{i,j}, ..., \mathbf{u}^{H,W}]$, where $\mathbf{u}^{i,j} \in \mathbb{R}^{1\times1\times C}$ corresponding to the spatial location (i, j) with $i \in \{1, ..., H\}$ and $j \in \{1, ..., W\}$. The spatial squeeze operation is achieved through a convolution $\mathbf{q} = \mathbf{W}_s q \star \mathbf{U}$ with weight $\mathbf{W}_{sq} \in \mathbb{R}^{1\times1\times C\times1}$. Each $q_{i,j}$ represents the linearly combined representation for all channels C for a spatial location (i, j). Then $\mathbf{q}$ is passed through a sigmoid layer $\sigma(.)$ to recalibrate or excite $\mathbf{U}$ spatially

$$\hat{\mathbf{U}}_{sSE} = \mathbf{F}_{sSE}(\mathbf{U}) = [\sigma(q_{1,1})\mathbf{u}^{1,1}, ..., \sigma(q_{H,W})\mathbf{u}^{H,W}]. \quad (2)$$

Each value $\sigma(q_{i,j})$ corresponds to the relative importance of spatial information (i, j) of the given feature. This recalibration provides more importance to relevant spatial locations and ignores irrelevant ones.

C. Classifier

In the end, we use a Multilayer Perceptron (MLP) layer with the softmax activation as the classifier in our model. The softmax function calculates the cross entropy and converts the *logits* into a probability distribution. The probability of T-th sample for i class is calculated as $P_i = exp({\theta_i}^T)/\sum_{k=1}^{K} exp({\theta_k}^T)$, where $\boldsymbol{\theta}^T$ is the output logits from the previous linear layer.

For identification tasks, we choose the user class with maximum P_i as the prediction for identification. For authentication tasks, the sample is labeled as a true sample for j user if $j = \text{argmax}_i \{P_i \mid P_i > \sigma_i\}$, where σ_i is the threshold that is empirically set by the user or adaptively learned from the training set, which characterizes the strictness of the system.

VI. System Evolving

A. Self Calibration

As time passes, the gesture of a legal user may change slightly, which requires our model to be adaptive to the transition to avoid frequent model reset. As we mentioned above, we set a threshold σ for a legal user, and it will decide whether the sample is legitimate or not in the authentication task. We set a higher threshold σ_e for self-calibration and add a sample into the training set of user i when the model determines that the sample is legitimate for corresponding user and $P_i > \sigma_{e,i} > \sigma_i$. According to this process, the training set of the model is automatically updated continuously when adding a new positive sample or when a specified number of new positive samples are accumulated, which ensures higher authentication accuracy and makes the system more reliable and adaptive.

B. Transfer Learning

When the domain of users changes, retraining the model will bring intensive time-consuming and resource-consuming, and the ability of feature extraction would be insufficient. We choose to utilize *pre-trained* and *fine-tuning* [36] methods to retrain the new user's model, which is commonly used in transfer learning [23]. For the addition of new datasets, we truncate the pre-trained softmax layer in the pre-trained model and replace it with the softmax layer of the new datasets. In order to maintain the training effect of the original large-scale data, the parameters are updated using a learning rate of one-tenth of the train from scratch. The *fine-tuning* method effectively solves the previously mentioned problems and maintains the validity of the model on new datasets.

VII. Evaluation

We conduct a comprehensive evaluation of ThumbUp through laboratory studies. We first collected motion sensor signals from 65 participants to determine the accuracy of ThumbUp for identification task and micro & macro benchmark of our model. Then, we explore the robustness of authentication with imitation attacks. Moreover, we show the performance of our system about the real-time ability and power consumption. Additionally, we perform the user study and illustrate how to choose gestures for better performance.

A. Implementation

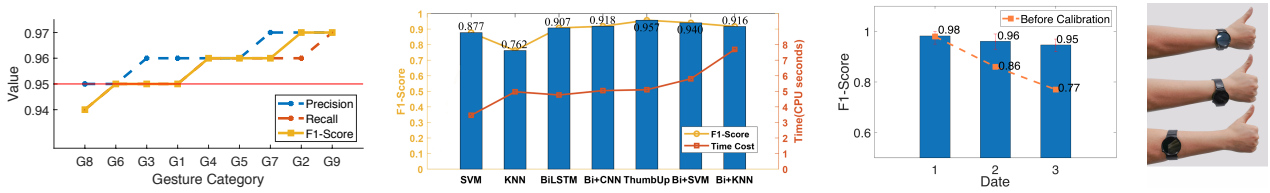
Motion sensing: We conduct all our experiments using the HUAWEI-WATCH with Android Wear 2.0.0 and Android Operating System 7.1.1. For the motion signal collection, we utilize the built-in accelerometer and gyroscope in the smartwatches and use the motion readings through existing Android Wear APIs to detect signals. The sampling rates of the accelerometer and gyroscope are both 100Hz.

Algorithm model: We use TensorFlow for construction and training for the neural networks off-line. We train the deep learning model offline on a PC with 12 Intel i7-8700K CPU kernels, 64GB memory, and 4 Titan X GPUs. After that, we build the trained model in the TensorFlow Lite framework and employ our system on the Android mobile platform for real-time evaluation.

B. Identification Accuracy

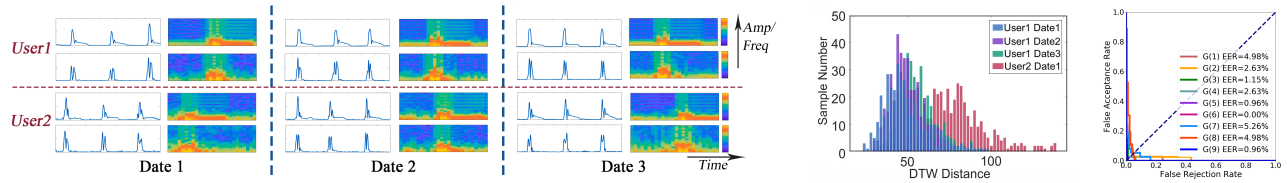
We recruit 65 volunteers and perform extensive studies on the collected dataset for over three months. 35 participants are male, and 30 are female. Their ages range from 19 to 57 (AVG=28.6, 6 > 50s). 75% are students, and the rest are non-students. 41 of them is fairly experienced with smart-phones and computers. 23 of them are familiar with wearables.

1) Baseline Classification Accuracy: We first investigate the accuracy of our system across multiple users. Before the experiments, we explain our system briefly and show the example photos of 9 gestures (illustrated in Fig. 1) to the participants. We ask participants to wear the smartwatch on their dominant hands, maintaining a comfortable tightness. Before the data collection, the participants are asked to



(a) (b) (c) (d)

Fig. 8. (a)Identification accuracy, (b)Comparisons, (c)3 Periods w & w/o calibration, (d)Placement



(a) (b) (c)

Fig. 9. (a)Temporal stability of users for different gestures, (b)DTW distance, (c)ROC for Imitation Attack.

TABLE I
RESULTS OF THE PROPOSED MODEL WITH DIFFERENT SWITCH CONFIGURATIONS.

Models	Highest F1-score	Lowest F1-score	Average F1-score
1-layer BilSTMs	0.90	0.96	0.935
2-layer BilSTMs	0.90	0.96	0.941
4-layer BilSTMs	0.92	0.97	**0.946**
without sSE	0.92	0.96	0.937
cSE	0.92	0.97	0.945
scSE	0.93	0.97	**0.949**
All(Full model+sSE)	0.94	0.97	**0.957**

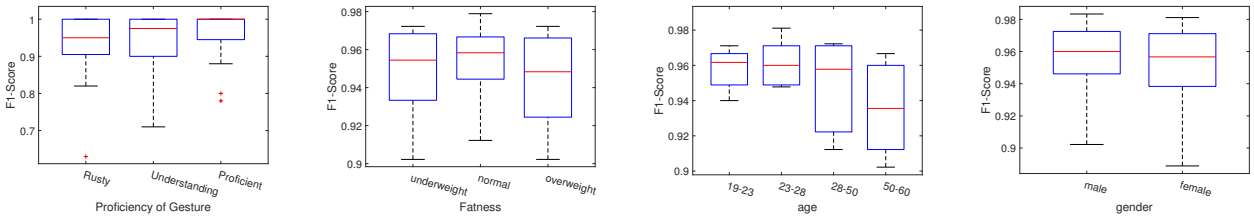
practice the gestures a few times. Once comfortable, each participant is asked to perform 9 gestures with 20 repetitions. That is, we have $65 \times 9 \times 20$ gestures in the dataset.

We evaluate the identification quality of the 9 gestures by precision, recall, and F1-Score. For each gesture, we repeat the training process for 10 times by randomly selecting 10 of 20 samples as the training samples and compute the average results in the rest 10 samples. The results (shown in Fig. 8(a)) demonstrate that our system obtains average accuracy of 95.7% for nine gestures. The accuracy of G2 (finger-turning in circles) and G9 (finger-bending) is up to 97%, which confirm the ability of identification.

2) Micro & Macro Benchmark: **Impact of parameters configuration:** For the input spectrogram of extracted features, our model reaches the best performance with the 128 width of a sliding window (choose from [256,128,64,32]), 8 for increment (choose from [32,16,8,4]). To know the importance of various components in the model, we perform ablation studies (shown in Table I). We verify that the good performance of our model mostly results from using sSE network and using 3-layer BiLSTM. We observe that 4-layer BiLSTMs achieve comparable accuracy of 3-layers. To balance the high-precision and computation cost, we adopt 3-layer BiLSTMs in our system. Meanwhile, we compare two commonly used attention mechanisms: cSE [10] and scSE [26]. Also, we find that Batch Normalization layer and Dropout layer have a significant effect on the stability and generalization ability of the model. We compare ThumbUp with BiLSTM, BiLSTM+CNN, SVM, and kNN, which are commonly used in IMU signal classification [9] [39]. Moreover, we use the features extracted by BiLSTM as the input of traditional classification algorithms (SVM, KNN), which achieves higher accuracy than using original spectrogram and shows the effectiveness of our model for feature extraction. The result (shown in Fig. 9(b)) displays that our model achieves the highest accuracy on our dataset with acceptable computation cost.

Impact of time horizon: To evaluate the similarity and repeatability of authentication over time, we test the performance of ThumbUp over 3 months. We recruit 20 participants ranging in age from 19 to 29 (AVG: 24.8, SDT: 2.5) included in the list of the above 65 participants. Each participant repeats 9 gestures 20 times in each session (Date1, Date2, and Date3). The gap between two sessions is 3-4 weeks. Fig. 9(a) intuitively shows the temporal stability of the accelerometer signals and its spectrogram for two users over time. Fig. 9(b) shows the difference in DTW distance among different periods. The signals undergo some changes after a long interval of 3 periods but still similar. We notice that the user remembers the type of gestures but might forget the specific details after months, which are essential factors of user uniqueness, especially for tiny gestures. In order to maintain the usability of the model, we add the evolving mechanism described in Sec. VI. We train the initial model with 10 samples collected at the first session. In each subsequent period, we split 10 samples from 20 samples and select them to update the model based on the evolution mechanism. The remaining samples are used to test the accuracy of the model. We compare the authentication accuracy of this method with/without the update mechanism. The result (Fig. 8(c)) shows that our system is less effective as the time gap increases between two separate authentication attempts, which leads to a problem that it cannot be used in applications with long intervals without the opportunity to update itself. However, ThumbUp achieves high accuracy with evolving mechanism. As our experiments show, our method displays 0.95 F1-value even after two periods and increases 0.18 than the one without updating, which shows that our system with an evolving mechanism is effective.

Impact of placement: In everyday life, users may not



(a) (b) (c) (d)

Fig. 10. Impact of proficiency, fatness, age, and gender

wear the smartwatch at the standard position. In order to test the impact of the wearing position, we asked 5 participants to wear the smartwatch at two atypical positions shown in Fig. 8(d). Evaluations show that the average F1-score is 0.93 for the standard position, 0.44 for the loose band, 0.13 for the forearm, which reveals that the system can not identify users with smartwatch at atypical positions. We think that the signal is too weaker at the forearm and it will add much noise with a loose band.

C. Authentication Robustness

For exploring the security against attackers, we focus on the imitation attack which we believe is the most threatening attack type.

We asked 10 participants (attackers) outside the list of 65 participants in the training set to imitate motion patterns of 10 participants (targets) who are included in the training set. Then we calculate their chances of successful imitation, *i.e.*, ThumbUp mistakenly accepts samples from the attackers. The attackers' ages range from 19 to 50; 5 are male. All participants are relatively proficient with computers and smartwatches and familiar with these gestures. We take video footage when the five target participants perform the gestures. Each attacker mimics 9 gestures eight times to their best effort while watching the targets' videos. In summary, we collect 40 samples for each gesture each target user. We also ask the target users to repeat each gesture 40 times in order to balance the number of positive and negative samples in evaluation.

We calibrate the threshold σ in the authentication mechanism to observe the False Rejection Rate (FRR) and False Acceptance Rate (FAR). The Receiver Operating Characteristic (ROC) curve of one user is shown in Fig. 9(c). We summarize the average Equal Error Rates (EER) for the nine gestures in Table II. We observe that under an appropriate threshold, we can make a proper distinction between attackers and legitimate users. As the table shows, G2 (finger-turning) and G4 (hand-waving) perform best against imitation attacks, while G7 (gun-gesture) is close behind. Moreover, we compare ThumbUp with state of the art one-class classifiers: GAN [27] and Autoencoder+SVM [40]. Comparing to 0.221 for GAN and 0.173 for AE+SVM, ThumbUp achieves the lowest average EER, which is 0.025. We suspect that the number of training samples is too little for a one-class deep neural network classifier. ThumbUp is trained by the amount of training samples from different users, which take advantage of the feature extraction that happens in the front layers of the network without developing the network from scratch. Moreover, as shown in Fig. 9(b), the DTW distance (used in [3]) from different users cannot be clearly distinguished.

D. Delay and Power Consumption

We deploy our system on a HUAWEI-WATCH to explore the real-time ability of ThumbUp. We estimate the delay of 5000 times. The average latency from the time when the user finishes their gestures to the time that authentication is finished is 0.085s. The result indicates the real-time ability of our system.

We use the Android Debug Bridge (ADB) tool for evaluating power consumption. We compare two states of the smartwatch: idle display and running the authentic system 5 times per second. Then we estimate the power consumption of the screen-on smartwatch for one hour. With our system running, the power capacity of the smartwatch drops to 213mAh, while the initial is 264mAh before running our system. Meanwhile, when the system is idle, the power capacity of the smartwatch drops to 231mAh with the same initial battery capacity.

E. User Study and Discussion

Impact of proficiency: We ask the participants to record the proficiency of gestures at the end of experiments and divide samples from 34 participants into these 3 categories (*Rusty*, *Understanding*, *Proficient*) (32:100:183). We calculate the average F1-score for each category. As shown in Fig. 10(a), for a certain gesture, the more proficient the user is, the higher stability the authentication process has.

Impact of fatness: We record the Body Mass Index (BMI) values and waist circumference of participants, which is used to quantify the amount of tissue mass in an individual [11]. We divide participants into 3 categories (underweight, normal, and overweight) (13:24:12) according to [2]. As shown in Fig. 10(b), we observe that the F1-score decreases slightly as fatness rises. And we make the assumption that the abilities to control muscles decline as fatness rises, which may affect the accuracy of authentication.

Impact of age and gender: We divided participants into 4 categories by age (6:10:32:16). According to the result shown in Fig. 10(c), users between 23 and 28 have higher F1-score, which may be related to the stiffness or fatigue of muscles caused by increasing age. Another reason may be that younger participants are more adept at these gestures according to the user survey. Besides, we choose 30 males

TABLE II
F1-SCORE OF IDENTIFICATION, EER OF IMITATE-ATTACK FOR EACH GESTURE AND THE RANK OF USER-FRIENDLINESS

Gesture	F1-score of **Identification**	EER of **Imitation Attack**	Rank of **Friendliness**
G5	0.96	0.028	1
G8	0.94	0.033	2
G2	**0.97**	**0.014**	3
G6	0.95	0.026	4
G9	**0.97**	0.027	5
G7	0.96	0.020	6
G4	0.96	0.018	7
G1	0.95	0.032	8
G3	0.95	0.025	9

and 30 females that cover the age range from 19 to 60 separately. In Fig. 10(d), the F1-Scores of male and female are roughly the same.

At the end of the study, we survey the participants' opinions about the usability, applicability, and usefulness of the system. 84% of participants think that ThumbUp is convenient and reliable. They are willing to use our system as the approach to get access to the smartwatches in daily life. Furthermore, participants are asked to choose three gestures that they are most willing to use in their daily lives. The result is depicted in Table II. The popularity decreases from top to bottom, '1' represents the gesture with the highest user satisfaction, and '9' represents the least. Our survey shows that most users prefer relatively simple gestures like G5(fist-making) and G8(thumb-up). Combining the statistical results of previous experiments, we generally recommend the top five gestures (listed in Table II) with both user-friendliness and usability. Considering the relationship between proficiency and accuracy of gestures (shown in Fig. 10(a)), users can redefine their personal unlock gestures with the most familiar gestures for better security.

VIII. Related Work

Existing biometrics authentication approaches can be divided into two categories: *physiological* and *behavioral* techniques. Physiological techniques take advantage of the physical characteristics of human body, such as heart [17], body electric potentials [33] and acoustics [39]. Behavioral techniques utilize unique manners, such as gait [28], head movements [16], and even tongue movement [22], which are closely related to personal behavioral habits.

Gesture-based authentication and recognition have drawn great attention in academia and industry, with sensor signal based on depth camera [38], capacitance [31], electromagnetic [14], wireless signals [24], *etc.* Zhao et al. [38] propose a depth camera-based dynamic hand gesture authentication method, which achieves 95.21% accuracy for the complicated gesture and 91.38% for the simple gesture. Yang et al. [35] present the study of mobile authentication using free-form touchscreen gesture generated by participants instead of text passwords. In this method, they find that participants generate new passwords and authenticate faster with comparable memorability while being more willing to retry.

There are some gesture-based studies using inertial measurement unit. Chris et al. [18] use motion patterns when users are entering passwords on smartwatches. Object Hallmarks [25] utilizes data from IMU embedded on the wearable wrist to generate fingerprints of users' behaviors when using objects like fridge and freezer, taking advantage of Euclidean distance of signal peak values. Sun et al. [29] propose a 3-D hand gesture signature-based biometric authentication system with an on-phone accelerometer, and the results tested by 19 users show 4.65% FRR and 0.27% FAR. MotionAuth [34] uses the arm-movement based motion sensor signal measured by wrist-worn smart devices, which is similar to our design. It authenticates with a large scale arm-generated gestures like lifting the hand, while ThumbUp achieves a comparable secure authentication using a simple hand gesture.

There are also some relative works about subtle gesture kinematics analysis used for wrist-worn or other mobile devices. TwistIn [15] takes a smartwatch as an authentication token for access and control of other smart devices by twisting the phone a few times, and it achieves 95% accuracy for 12 users. Taprint [3] proposes a secure PIN input system, which extends a virtual number pad on the back of hands with smart wristbands. Also, it uses the tapping vibrometry as biometrics to authenticate the user with an accuracy of 96% for 128 users.

IX. Conclusion

In this work, we present ThumbUp to identify and authenticate users using *only one* simple gesture like thumb-up. We carefully design pre-processing methods for converting weak signals with noises to a spectrogram containing user characteristics. A light-weight robust deep neural network is designed to extract the unique representations from motion signals and identify users accurately. We demonstrate its utility by extensive experimental studies with 65 users over 3 months. The evaluations demonstrate the power of ThumbUp: it achieves extremely high accuracy (above 97%) for identification and EER 0.029 for authentication. We believe that our design will open up a wide range of exciting opportunities for convenient and secure authentication using wearable smart devices.

Nevertheless, this is only a first step towards cracking an extremely tough task. Several challenging issues are left for future research: 1) improving the system robustness in a more hostile environment when users engage in other daily activities like walking and running; 2) implementing our scheme in commercial products for further verifying system performance in realistic scenarios; 3) exploring the system adaptivity with gestures completely designed by users.

ACKNOWLEDGEMENT

The research is partially supported by National Key R&D Program of China 2018YFB0803400, China National Funds for Distinguished Young Scientists with No.61625205, China National Natural Science Foundation with No. ł61751211, No.61520106007, Key Research Program of Frontier Sciences, CAS. No. QYZDY-SSW-JSC002.

References

[1] Ballard, L., Lopresti, D., and Monrose, F. Evaluating the security of handwriting biometrics. In *Tenth International Workshop on Frontiers in Handwriting Recognition* (2006), Suvisoft.

[2] Chen, C., Lu, F., et al. The guidelines for prevention and control of overweight and obesity in chinese adults. *Biomedical and environmental sciences: BES 17* (2004), 1.

[3] Chen, W., Chen, L., Huang, Y., Zhang, X., Wang, L., Ruby, R., and Wu, K. Taprint: Secure text input for commodity smart wristbands. In *The 25th Annual International Conference on Mobile Computing and Networking* (2019), ACM, pp. 1–16.

[4] Duc, N. M., and Minh, B. Q. Your face is not your password face authentication bypassing lenovo–asus–toshiba. *Black Hat Briefings 4* (2009), 158.

[5] Gander, W., and Hrebicek, J. *Solving problems in scientific computing using Maple and Matlab®*. Springer Science & Business Media, 2011.

[6] Graves, A., Jaitly, N., and Mohamed, A.-r. Hybrid speech recognition with deep bidirectional lstm. In *2013 IEEE workshop on automatic speech recognition and understanding* (2013), IEEE, pp. 273–278.

[7] Griffin, D., and Lim, J. Signal estimation from modified short-time fourier transform. *IEEE Transactions on Acoustics, Speech, and Signal Processing 32*, 2 (1984), 236–243.

[8] Hochreiter, S., and Schmidhuber, J. Long short-term memory. *Neural computation 9*, 8 (1997), 1735–1780.

[9] Hou, J., Li, X.-Y., Zhu, P., Wang, Z., Wang, Y., Qian, J., and Yang, P. Signspeaker: A real-time, high-precision smartwatch-based sign language translator. In *ACM MobiCom* (2019).

[10] Hu, J., Shen, L., and Sun, G. Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 7132–7141.

[11] Janssen, I., Heymsfield, S. B., Allison, D. B., Kotler, D. P., and Ross, R. Body mass index and waist circumference independently contribute to the prediction of nonabdominal, abdominal subcutaneous, and visceral fat. *The American journal of clinical nutrition 75*, 4 (2002), 683–688.

[12] Kratz, S., and Rohs, M. A $3 gesture recognizer: simple gesture recognition for devices equipped with 3d acceleration sensors. In *Proceedings of the 15th international conference on Intelligent user interfaces* (2010), ACM, pp. 341–344.

[13] Kumar, A., Singh, T., and Kumar, A. Hand anatomy. *Biometrics Research Laboratory, Department of Electrical Engineering, Indian Institute of Technology Dehli, New Dehli, India* (2009), 1–11.

[14] Laput, G., Yang, C., Xiao, R., Sample, A., and Harrison, C. Em-sense: Touch recognition of uninstrumented, electrical and electromechanical objects. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology* (2015), ACM, pp. 157–166.

[15] Leung, H.-M. C., Fu, C.-W., and Heng, P.-A. Twistin: Tangible authentication of smart devices via motion co-analysis with a smartwatch. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2*, 2 (2018), 72.

[16] Li, S., Ashok, A., Zhang, Y., Xu, C., Lindqvist, J., and Gruteser, M. Whose move is it anyway? authenticating smart wearable devices using unique head movement patterns. In *2016 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (2016), IEEE, pp. 1–9.

[17] Lin, F., Song, C., Zhuang, Y., Xu, W., Li, C., and Ren, K. Cardiac scan: A non-contact and continuous heart-based user authentication system. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking* (2017), ACM, pp. 315–328.

[18] Lu, C. X., Du, B., Zhao, P., Wen, H., Shen, Y., Markham, A., and Trigoni, N. Deepauth: in-situ authentication for smartwatches via deeply learned behavioural biometrics. In *Proceedings of the 2018 ACM International Symposium on Wearable Computers* (2018), ACM, pp. 204–207.

[19] Marieb, E. N., and Hoehn, K. *Human anatomy & physiology*. Pearson Education, 2007.

[20] Merletti, R., Parker, P. A., and Parker, P. J. *Electromyography: physiology, engineering, and non-invasive applications*, vol. 11. John Wiley & Sons, 2004.

[21] Muaaz, M., and Mayrhofer, R. Smartphone-based gait recognition: From authentication to imitation. *IEEE Transactions on Mobile Computing 16*, 11 (2017), 3209–3221.

[22] Nguyen, P., Bui, N., Nguyen, A., Truong, H., Suresh, A., Whitlock, M., Pham, D., Dinh, T., and Vu, T. Tyth-typing on your teeth: Tongue-teeth localization for human-computer interface. In *Proceedings of the 16th Annual International Conference on Mobile Systems, Applications, and Services* (2018), ACM, pp. 269–282.

[23] Pan, S. J., and Yang, Q. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering 22*, 10 (2009), 1345–1359.

[24] Pradhan, S., Chai, E., Sundaresan, K., Qiu, L., Khojastepour, M. A., and Rangarajan, S. Rio: A pervasive rfid-based touch gesture interface. In *Proceedings of the 23rd Annual International Conference on Mobile Computing and Networking* (2017), ACM, pp. 261–274.

[25] Ranjan, J., and Whitehouse, K. Object hallmarks: Identifying object users using wearable wrist sensors. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2015), ACM, pp. 51–61.

[26] Roy, A. G., Navab, N., and Wachinger, C. Concurrent spatial and channel squeeze & excitationin fully convolutional networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* (2018), Springer, pp. 421–429.

[27] Sabokrou, M., Khalooei, M., Fathy, M., and Adeli, E. Adversarially learned one-class classifier for novelty detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3379–3388.

[28] Schürmann, D., Brüsch, A., Sigg, S., and Wolf, L. Bandanabody area network device-to-device authentication using natural gait. In *2017 IEEE International Conference on Pervasive Computing and Communications (PerCom)* (2017), IEEE, pp. 190–196.

[29] Sun, Z., Wang, Y., Qu, G., and Zhou, Z. A 3-d hand gesture signature based biometric authentication system for smartphones. *Security and Communication Networks 9*, 11 (2016), 1359–1373.

[30] Tari, F., Ozok, A., and Holden, S. H. A comparison of perceived and real shoulder-surfing risks between alphanumeric and graphical passwords. In *Proceedings of the second symposium on Usable privacy and security* (2006), ACM, pp. 56–66.

[31] Truong, H., Zhang, S., Muncuk, U., Nguyen, P., Bui, N., Nguyen, A., Lv, Q., Chowdhury, K., Dinh, T., and Vu, T. Capband: Battery-free successive capacitance sensing wristband for hand gesture recognition. In *Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems* (2018), ACM, pp. 54–67.

[32] Xu, C., Pathak, P. H., and Mohapatra, P. Finger-writing with smartwatch: A case for finger and hand gesture recognition using smartwatch. In *Proceedings of the 16th International Workshop on Mobile Computing Systems and Applications* (2015), ACM, pp. 9–14.

[33] Yan, Z., Song, Q., Tan, R., Li, Y., and Kong, A. W. K. Towards touch-to-access device authentication using induced body electric potentials. *arXiv preprint arXiv:1902.07057* (2019).

[34] Yang, J., Li, Y., and Xie, M. Motionauth: Motion-based authentication for wrist worn smart devices. In *2015 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)* (2015), IEEE, pp. 550–555.

[35] Yang, Y., Clark, G. D., Lindqvist, J., and Oulasvirta, A. Free-form gesture authentication in the wild. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (2016), ACM, pp. 3722–3735.

[36] Yosinski, J., Clune, J., Bengio, Y., and Lipson, H. How transferable are features in deep neural networks? In *Advances in neural information processing systems* (2014), pp. 3320–3328.

[37] Yu, T., Jin, H., and Nahrstedt, K. Writinghacker: audio based eavesdropping of handwriting via mobile devices. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing* (2016), ACM, pp. 463–473.

[38] Zhao, J., and Tanaka, J. Hand gesture authentication using depth camera. In *Future of Information and Communication Conference* (2018), Springer, pp. 641–654.

[39] Zhou, B., Lohokare, J., Gao, R., and 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* (2018), ACM, pp. 321–336.

[40] Zou, Y., Zhao, M., Zhou, Z., Lin, J., Li, M., and Wu, K. Bilock: User authentication via dental occlusion biometrics. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2*, 3 (2018), 152.