ThumbUp: Secure Smartwatch Controller for Smart Homes using Simple Hand Gestures

Xiaojing Yu*, Zhijun Zhou*, Lan Zhang†, Xiang-Yang Li†

Abstract—The development of creative applications and intelligent gadgets requires a secure and straightforward interface with human users. We propose, design, and implement ThumbUp, a smartwatch-based two-factor real-time identification and authentication system in which smartwatch users can identify and authenticate themselves using some simple hand and finger movements, such as thumb-up. ThumbUp leverages the signal from the Inertial Measurement Unit (IMU) in in Commercial-Off-The-Shelf (COTS) smart devices to discover the unique pattern generated by each user's simple gestures using a carefully constructed deep learning model. Smart homes provide a comfortable, safe, and efficient living environment, epecially help the sick and aged. We propose strategies for convenient and reliable control in smart homes with gesture command recognition. We build an Auto-Encoder-based filter that reconstructs the raw data to improve the representation of gesture features. Moreover, we adopt the random forest method to analyse the contextual command correlation. And we employ the authentication system based on smartwatch for personalized command feedback and ensure that illegals cannot use the device. We implement our system and undertake rigorous studies to determine its usefulness and efficiency over a three-month period with 65 users. It achieves a 97% accuracy for user classification and an EER of 0.014 for authentication task with a single simple gesture. And our method achieves 91% accuracy for command recognition and 96% command accuracy with contextual informations. Additionally, we conduct a study of user acceptability of our system and explain how gesture proficiency influences authentication accuracy.

Index Terms—Smart Home, Smartwatch, Hand Gestures, Security, IMU Signals

1 Introduction

Smart homes integrate automation control, computer network and communication in one of the intelligent home control systems. The goal of smart homes is to provide a comfortable, safe, convenient, and efficient living environment. With solid product offerings such as SmartThings and Google's Nest, the smart home market develops by increasing suffer approbate, which is expected to reach 1.4 billion U.S. dollars in spending by 2023 [1]. In-home control can be divided into active and passive control. Active control means that users control functions voluntarily, such as playing music. Passive control is based on the analysis of the user's behaviors, initiated by the control system, such as automatic temperature control. This work focuses on the active control of smart homes. We believe that an excellent control system for smart homes ought to be accurate (accurately recognize commands), fast (have a low latency), user-friendly (be convenient to be used frequently), secure (unavailable to illegal users), and lightweight (not consume an excessive amount of resources). However, existing systems are incapable of balancing these characters. At present, most smart home components are controlled by the keyboard on mobile phones, remote controls, etc. Among these traditional methods, speech recognition technology has been widely used in the smart home, but it has low efficiency in difficult acoustic environments [2]. Biometric control method has attracted substantial attention [3]. However, sensing humans automatically with wireless signals such as WiFi [4] is greatly affected by the environmental change. Besides, existing studies lack the protection of personalized information and cannot provide personalized responses.

The popularity of smart wearable devices has facilitated rapid and convenient interaction with the physical world. Smartwatches are used for various applications, including instant messaging, online shopping, and rapid mobile payment, which are an excellent way to control the smart home. Numerous previous studies utilizing IMU have concentrated on big motion patterns produced by arm movements [5], [6], which maybe be inconvenient in daily applications. Luna *et al.* [3] propose an interaction method with smart TVs via gestures performed by a persons wrist using a smartwatch and provide recognition along three axes of IMU. Taprint [7] extends a virtual number pad on the back of hands with smart wristbands. Tapping vibrometry as biometrics is used to authenticate users with an accuracy of 96% for 128 users. It requires both hands to operate, which is not convenient enough for daily fast control. Kundu *et al.* [8] design a common hand gesture recognition system for wheelchair control with a classification accuracy of 94% based on electromyography (EMG) sensor , which is not commonly adopted in commercial devices.

In this work, we propose a convenient and reliable smart home control system based on smartwatches with accurate command recognition and secure user authentication, which provides personalized command feedback and ensure that illegals cannot use the device. The most noteworthy feature of our design is that we utilize simple gestures, mainly performed with the fingers, such as twisting the fingers or thumbing up (as illustrated in Fig. 1). Specifically, we need to address several

• *School of Computer Science and Technology, University of Science and Technology of China, Hefei, China. E-mail:*{yxjing, zhouzj18}@mail.ustc.edu.cn, †{zhanglan, xiangyangli}@ustc.edu.cn*

Fig. 1. Illustration of nine gestures studied in ThumbUp (from top, left to right): G1 (Snapping), G2 (Twisting finger), G3 (Beckon), G4 (Hand-waving), G5 (Fist-making), G6 (Victory-gesture), G7 (Gun-gesture), G8 (Thumb-up), and G9 (Finger-bending).

significant technical challenges:

- **Limited Training data:** The motion signals collected by the IMU for small gestures are much weaker than those collected for arm motions. Thus, the features derived from minuscule hand/finger movement may be masked by the inherent noise generated by the IMU sensors. Likewise, requiring a user to act during the user-training phase repeatedly is not user-friendly. It is complicated to extract helpful activity and identify features that uniquely characterize each user from weak signals with small sample sizes.
- **Reliability and Robustness:** In order to ensure the safety, it is essential that our system can accurately authenticate valid users and defend against attackers who may attempt to mimic authorized users maliciously. The authentication features chosen by our system should account for both the diversity of different users and the consistency of a single user. As biometrics of behavior, users' gestures will alter slightly over time. The system should be flexible to minor changes in users' hand gestures and avoid frequent model resets.
- **Energy-efficiency and Real-time Ability:** We need to implement a lightweight system that uses the limited storage and computational power available on smartwatches while maintaining high stability and real-time capability.

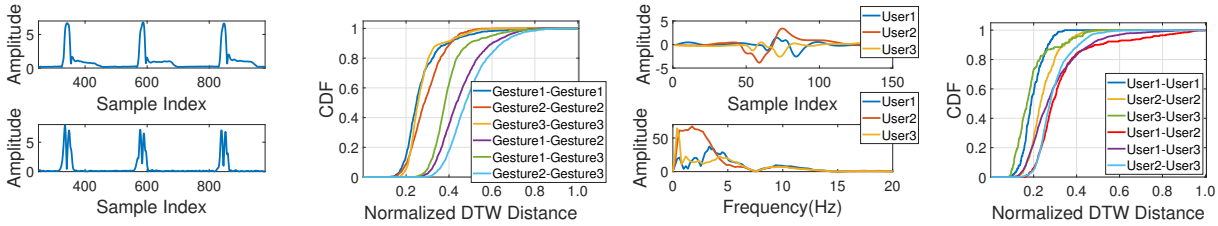
To address these issues, we design, implement, and evaluate **ThumbUp**, a system capable of authenticating users and controlling smart home services based on a small gesture. We analyze the anatomy of hand movement in human kinematics. Meanwhile, we investigate the stability and diversity of motion sensor signals using an extra verification signal, EMG. After pre-processing and detecting motion signals, 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 for users and gestures. ThumbUp involves an updating strategy that allows the model to evolve continuously in response to behavioral changes and system initialization for domain adaptation. In addition, we build an Auto-Encoder (AE) based privacy-preserving filter that outputs reconstructed STFT spectrogram instead of original data to improve the feature representation ability. Combined with the smart home scenario, we employ the random forest method to analyze user behavior and further improve the accuracy of command recognition. We demonstrate that ThumbUp can precisely identify users with a mean accuracy exceeding 95.7% and successfully verify a legitimate user with a mean error rate of 0.025 using a prototype implementation on COTS wearable devices and 65 participants. And ThumbUp can accurately recognize gesture command with an accuracy of 96% with contextual behaviour analysis.

Our work provides a novel and reliable approach to control the smart home used in a highly convenient way. Besides, to the best of our knowledge, ThumbUp is the first solution to leverage basic finger-movement gestures for user identification/authentication with IMU on COTS smartwatches. We expect that ThumbUp has potential applications in (a) enabling access to smart wearable devices; (b) quick payment by easy interaction; and (c) operating mobile devices secretly and reliably.

To summarize, we make the following contributions:

- We develop and implement a reliable authentication mechanism for wearable devices based on small gestures. We investigate the feasibility of using gestures as certification elements and establish that hand gestures include unique signatures of users. We design a model that extracts features and classifies user gesture patterns. Moreover, we suggest a self-calibration and transfer learning method to increase practicability and validity.
- We propose a user identity filter that reduce the user information and improve the feature expression of the gestures. Also, we propose a contextual command analysis method in smart home scenarios to improve the accuracy of gesture command recognition.
- We evaluate ThumbUp through extensive studies covering three months and involving 65 people. Experiments demonstrate that even simple finger gestures such as the victory gesture can yield reliable identification results. Furthermore, we test our system's security against imitated attacks, which demonstrates that the system can withstand such attacks with an average EER of 0.025.
- In terms of friendliness, the gestures utilized in our system are well-designed based on the research on biological kinematics systems. We interview participants regarding the comfort with which the gestures are performed. Then, we recommend gestures relying on both authentication performance and participant perceptions.

The remainder of the paper is organized as follows: In Sec. 2. We present the foundation for hand movement and feasibility analysis. In Sec. 3, we provide a high-level summary of ThumbUp's primary design. The details of our design are described in Sec. 4, 5, and 6. In Sec. 7, we present experimental evaluation results and user study. We introduce the related works in Sec. 8. In the end, we discuss the limitations of our work in in Sec. 9 and conclude in Sec. 10.



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

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

2 Basis of Hand Movement and Feasibility Study

In this section, we discuss the fundamental biological kinematics in order to theoretically establish the feasibility of hand movements as distinct features theoretically, investigates the stability and diversity of the motion sensor data, and concludes with the use of the EMG as an auxiliary verification.

2.1 Hand Movement's Anatomy

Intuitively, subtle movements are easier to be imitated. Muscle motions are controlled by the subconscious and are difficult to modify consciously. Even if two users perform identical movements, the biological kinematics of muscles are sufficiently distinct, allowing identification based on minor gestures. Additionally, because muscles are an interior component of the hand surface, they are quite resistant to changes in humidity and temperature [9].

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. The hand's movements are regulated by intrinsic muscles in the hand as well as muscles within muscles in the forearm (extrinsic muscles), providing for exceptional control of both precise and strong movements [10]. The motion signals would be perceived by a motion sensor on the forearm. It was shown in [11] 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, motion signals would capture both the biological and behavioral characteristics of muscles as a kind of authentication information.

2.2 Feasibility Study

Previous work [12] discusses the possibility of motion signals as unique certification conditions using EMG as an auxiliary verification. Here, we investigate gesture motion signals' diversity, consistency, and originality of gesture motion signals to understand ThumbUp's viability.

Diversity and Consistency of Motion Signals: To begin, we asked one participant to do ten times each of two different hand movements with a smartwatch. As illustrated in Fig. 2(a), the profiles for distinct movements vary significantly. We define the motion sensor signal as $S = \sqrt{G^T G + L^T L}$, where G and T signify the integration of angular and linear accelerations, respectively. To visualize the difference digitally, we compute the normalized Dynamic Time Warping (DTW) distance between signals from the same and different movements (as illustrated in Fig. 2(b)). The figure demonstrates unequivocally that it is possible to distinguish between various gestures. Meanwhile, the result (shown in Fig. 2(a)) supports the consistency of gesture motion signals by revealing that motion signals from repetitions of the same gesture are fairly similar.

Uniqueness of Motion Signals: The purpose of this study is to determine whether motion signals generated by various users for the same gesture are distinct. Three participants are asked to snap fingers 20 times each. As illustrated in Fig. 2(c), the profile is notably different in both the time and frequency domains. Similarly, we calculate the nomalized DTW distance (see Fig. 2(d)), which demonstrates that motion signals are unique for each user.

3 Design Scope and Overview

This section describes the design scope and system overview of ThumbUp.

3.1 Objective and Design Scope

We divide the security control into two parts: user identification and command recognition.

User Identification represents challenges involving both multi-user categorization and one-to-one identification against malicious attacks. Multi-user categorization aims to classify distinct users and provide personalized command responses when the family shares the wristbands. The objective of one-to-one identification is to correctly distinguish attackers and legitimate users, of which unauthorized users cannot use the control device.

Command Recognition represents challenges involving command signal detection and gesture recognition. To provide instant feedback, our system should detect the command signal quickly, and the classify model in real-time. The core task for a control system is to recognize command gestures accurately.

Gesture Design: We prefer to use our method in cases where daily identification of device owners is required, such as unlocking a smartwatch. Our system's gestures must be sufficiently convenient. Gestures that incorporate simultaneous movements of the fingers, palm, and wrist require more information to distinguish, but they are far less convenient. Meanwhile, simplistic movements such as softly waving one finger are useless in uniquely identifying a user. We define nine typical gestures (shown in Fig. 1), factoring user-friendliness and the trade-off between complexity and distinguishability. The following experiments assess users' perceptions of gestures and make recommendations for improvement.

Availability: Our system needs to extract *consistent* and *distinct* biometric signal features from small motion signals,

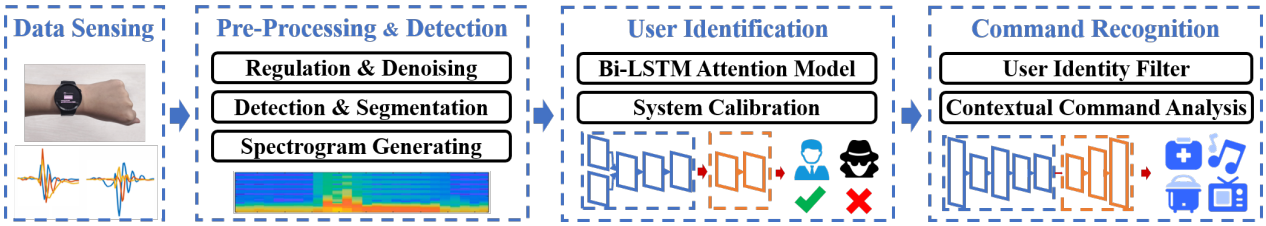


Fig. 3. System Workflow and Key Components of ThumbUp

i.e., the features must be durable to fulfill the requirements of long-term usage and diverse enough to withstand various types of attacks. Moreover, our system should have classifier mechanisms to give the corresponding results for two tasks.

Universal: We prefer to use standard commercial smartwatches equipped with accelerometers and gyroscopes to satisfy the universal requirement. As COTS smart-devices often have limited computational and storage resources, our solution requires a lightweight architecture.

We aim to develop a highly secure and reliable real-time smart home control system based on 3D simple gestures with commercially available smart devices.

3.2 Overview of System

As detailed in Fig. 3, ThumbUp is composed of three components: The first part is *pre-processing & detection* (Sec. 4), which aims to eliminate noises from continuous motion signals, detect the command signal, and extract sequential features. The second part is *user identification* (Sec. 5), which uses carefully built deep learning methods to extract representations from spectrograms and determine the user's identity. Besides, we introduce a continuous model evolution strategy to respond to user behavioural changes and system initialization for domain adaptation. The third part is *command recognition* (Sec. 6). We describe how to improve the classification ability by filtering user identity and contextual command analysis.

4 Pre-processing and Detection

Due of the noisy, partial, and even erroneous signals gathered by motion sensors, the first phase of ThumbUp is to filter out the noise and segment the signal to match the genuine motions.

4.1 Data Regulation & Denoising

To ensure that the accelerometer and gyroscope are sampled uniformly, we interpolate the data to $100Hz$ of the sampling rate. The amplitude of the signals is normalized using the Z-score technique, this is, the processed signals follow a typical normal distribution (mean= 0 and standard deviation= 1). Following that, we use a Savitzky-Golay smoothing filter [13], commonly known as a least-square smoothing filter, to eliminate random noise. The core idea behind this filter is to perform a least-square fit with a high-degree polynomial for each data point, spanning an odd-sized window centered on that data point [14], which not only minimizes noise but also preserves the shape and height of waveform peaks.

4.2 Detection and Segmentation

The IMU continuously collects motion signals; we need to detect possible samples and split signals into a given size. One popular way is to empirically establish a constant threshold and consider the portion of the signal as target sample whose short-term energy surpasses this threshold. But the threshold is difficult to select in practical varying noise scenarios.

We employ a method similar to the Constant False Alarm Rate (CFAR) algorithm [15] to detect the gestures. The central concept is to use dynamic thresholds to establish the start and finish points of a single gesture. We use X to denote the long time-series signal, while $x(i)$ 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, which is set to 128 in our setting. Besides, the average power and standard deviation at the i_{th} sample index, denoted as $E(i)$ and $D(i)$ respectively, are defined as:

$$E(i) = \frac{1}{W}\sum_{k=i-W+1}^{i} x(k)^2, D(i) = \sqrt{\frac{1}{W}\sum_{k=i-W+1}^{i}(x(k)^2 - E(i))^2}. \quad (1)$$

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. 4, and it shows satisfactory efficiency. The orange line represents the motion signals we concern.

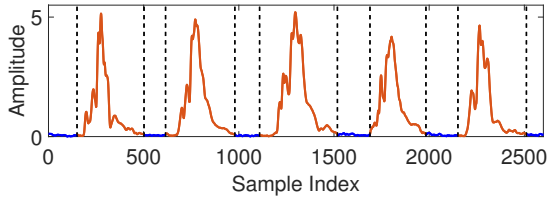


Fig. 4. Gesture Detection and Segmentation

4.3 Spectrogram Generating

For motion sensor signals, time-domain features indicate the sequential relationship of gestures, while frequency features reflect different hand muscles motions. The Short Time Fourier Transform (STFT) [16] has time domain sensitivity for both high and low-frequency signals and contains more frequency and time-domain information for the next step than DWT. The power band represents the frequency spectrum generated by STFT. The spectrogram is the STFT's magnitude squared, this is, $|X(m,w)|^2$. The discrete-time STFT of a signal $x[n]$ is calculated as $X(m,w) = \sum_n x[n]w[n-m]exp(-jwn)$. Hamming window is applied for the window function $w[n]$. By evaluating

$X(m; w)$ for a larger number of $(m; w)$ points, high-resolution information is obtained at the expense of reduced total information and greater computing cost. We achieve a satisfactory trade-off through experiments (shown in Sec. 7). We observe that the vibration caused by human mobility is mostly less than 15 Hz, and a cut-off frequency of 17 Hz for F is sufficient to preserve information for the motion sensor signal. We concatenate all channels and generate spectrograms to represent high-dimensional signals. The spectrogram is represented by a two-dimensional array with the dimension of 102×17.

5 User Identification

In this subsection, we will introduce user identification solution.

5.1 BiLSTM Attention Model

We propose a deep neural network (Fig. 5) for extracting subtle and stable representations from spectrograms and classifying users for identification and authentication. The model is divided into three parts: the BiLSTMs layer, the attention layer, and classifier layer. To begin, the input data are STFT-derived spectrograms Then, motion features are extracted using three-layer BiLSTMs. We add an attention mechanism based on Squeeze-and-Excitation Networks to significant aggregate information extracted from the motion representations generated by the BiLSTM layers. Finally, we use a Multilayer Perceptron as the classifier in our model with a softmax activation function. Moreover, we conduct ablation studies (described in Sec. 7.2.1) to better understand various parts of the proposed model.

5.1.1 Representation

Due to the structural properties, Recurrent Neural Networks (RNNs) store the memory based on historical information, making them well-suited for processing sequential data [17]. Long Short-Term Memory (LSTM) is purpose-built to address the problem of long-term reliance by utilizing memory cells that work better in longer sequences. For the spectrogram of the sequential temporal signal, BiLSTMs [18] have a greater capacity to extract representations than LSTMs with both before 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:

$$\overrightarrow{\mathbf{h}}_t = \overrightarrow{LSTM_F}(\overrightarrow{\mathbf{h}_{T-1}}, \mathbf{s}_t), \overleftarrow{\mathbf{h}}_t = \overleftarrow{LSTM_B}(\overleftarrow{\mathbf{h}_{T-1}}, \mathbf{s}_t). \quad (2)$$

After that, at each time step, these hidden state outputs from the forward LSTM $\overrightarrow{\mathbf{h}}_t$ and the backward LSTM $\overleftarrow{\mathbf{h}}_t$ are concatenated to enable encoding of information from past and future contexts respectively. With such a small number of training examples, models will quickly overfit. Especially since we only have a small number of training samples, models will easily overfit on these samples. The dropout layer randomly discards neural network units from the network. We send these concatenated hidden states to a dropout layer to avoid complex co-adaptations on training samples and achieve network model averaging.

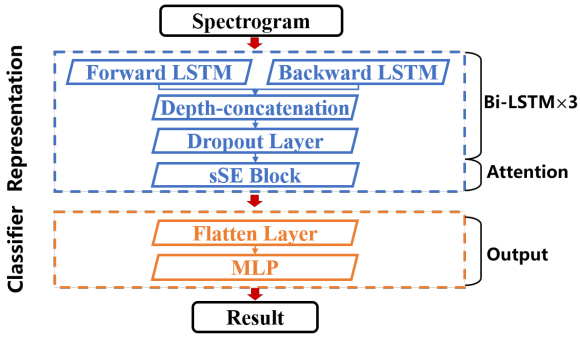


Fig. 5. Model Architecture.

We splice three layers of the network structure introduced above. In detail, we set the dimensionality of the Bi-LSTM output space as 64, 32, and 32. The fraction of the input units to drop is 0.5 in our setting. Then, a Batch Normalization layer is used to prevent gradient disappearance and explosion during *backpropagation* and to provide consistency between the updating stages of different scales.

Convolutional Neural Networks (CNNs) make use of convolutions to efficiently extract meaningful information. The features vectors created by BiLSTMs can be treated as an image. For image segmentation, the spatial information at the pixel level is instructive. To enhance the features' representational abilities, we add an attention mechanism with Channel Squeeze and Spatial Excitation Block (sSE) [19] to the model rather than utilizing CNNs directly, which'squeezes' the feature along the channels and 'excites' it spatially.

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 $\mathbf{U} = [\mathbf{u}^{1,1}, ..., \mathbf{u}^{i,j}, ..., \mathbf{u}^{H,W}]$, where $\mathbf{u}^{i,j} \in \mathbb{R}^{1 \times 1 \times C}$ 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 $\mathbf{W}_{sq} \in \mathbb{R}^{1 \times 1 \times C \times 1}$. 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

$$\hat{\mathbf{U}}_{sSE} = \mathbf{F}_{sSE}(\mathbf{U}) = [\sigma(q_{1,1})\mathbf{u}^{1,1}, ..., \sigma(q_{H,W})\mathbf{u}^{H,W}]. \quad (3)$$

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.

5.1.2 Classification

In the end, we use a Multilayer Perceptron (MLP) layer with the softmax activation as the classifier in our model. C_u represents the classification model. We put the feature into the classifier and obtain the finial result $C_u(x_i)$. The softmax function calculates the cross entropy, which is defined as

$$\text{loss} = \frac{1}{|\mathcal{D}|} \sum_{x_i \in \mathcal{D}} \sum_{m=1}^{M} d_{i,m} \log C_u(x_i). \quad (4)$$

The softmax function 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 user-defined or adaptively learnt threshold that defines the system's strictness.

5.2 System Evolving

5.2.1 Self Calibration

As time passes a legitimate user's gesture may shift slightly, necessitating that our model be adaptable to the transition in order to avoid frequent model resets. As discussed previously, we specify a σ threshold for a legal user, which will determine whether or not the sample is legitimate in the authentication task. We increase the self-calibration threshold σ_e and include a sample in the training set of user i if the model determines that the sample is valid for the related user and $P_i > \sigma_{e,i} > \sigma_i$.

Considering the computational overhead associated with iterative updates, we additionally include a "cache" concept: if the newly added data exceeds half of the cache storage, we will temporarily utilize the cache as the database and authenticate again. If the new data's authentication impact is superior to the original, it will be added to the database. If not, the cache will be cleared and the samples re-added. This approach updates the model's training set continually when a new positive sample is added or when a specified number of new positive samples is accumulated, ensuring improved authentication accuracy and increasing the system's reliability and adaptability.

5.2.2 Domain Transfer

When users' domain changes, retraining the model will be time- and resource-intensive, and the capacity to extract features will be insufficient. We opt to retrain the new user's model using *pre-trained* and *fine-tuning* [20] transferable approaches, which are frequently utilized in transfer learning. We truncate the pre-trained softmax layer in the pre-trained model and replace it with the softmax layer from the new datasets when adding new datasets. To preserve 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 proposed fine-tuning strategy effectively addresses the abovementioned issues while preserving the model's validity on new datasets. Although fine-tuning process saves a lot of computational overhead and time consumption compared with the original training, it still puts an unbearable burden on the smartwatch. So the fine-tuning part is recommended to be carried out on the cloud side with abundant computational resources. The smartwatch only deploys the trained lightweight model and performs the data collection function in the self-calibration.

6 Command Recognition

This section will introduce the key technologies of command recognition for smart homes based on gesture recognition and its contextual command analysis.

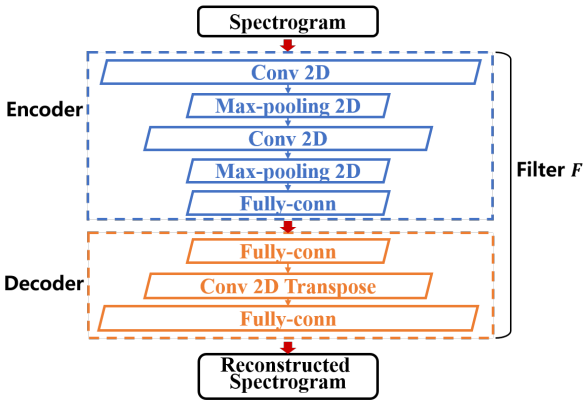


Fig. 6. Model Architecture of filter F

6.1 User Identity Filter

In Sec. 5.1, we establish the BiLSTM Attention model for user identification. The objectives of gesture classification and user classification are similar. We can adopt the same model architecture to classify gestures. However, the challenges of the two tasks are different. Users use the same actions in user identification tasks. The challenge we need to face is to extract unique and sustainable characteristics of users from weak signals. In the task of gesture recognition, in addition to extracting the features of different actions of the same user, we need to consider the diversity of different users, which greatly increases the difficulty of gesture recognition. Therefore, we designed a user identity filter to reconstruct the STFT spectrogram. The filter can effectively remove the user's identity information from the noise in a gesture recognition task and retain the action features to improve the accuracy.

6.1.1 Filter Architecture

The Auto-Encoder (AE) network is a commonly used unsupervised learning network that has been extensively used for anomaly detection and noise reduction [21]. AE includes an encoder that compresses the dimensions and extracts the representation from input samples and a decoder that reconstructs the data from encoded characteristics as nearly as possible to the original. The loss function typically employed in the training process is the MAE between inputs and reconstructed outputs. Inspired by the CNN-based AE networks introduced in [22], we develop a privacy-preserving filter (denoted as F, as shown in Fig. 6) that treats users' identities as noise and extracts activity features. Instead of uploading the actual IMU signals to the server, the filter generates the reconstructed STFT spectrogram.

As mentioned in Sec. 5, CNNs have a high capacity for representing image data. We leverage a two-layer CNN to extract gesture features with a 2D kernel filter for the encoder. Following each convolution layer is a max-pooling layer that reduces the dimensionality of the data. The numbers of output filters in the convolution layers are 32 and 64, respectively. And the kernel size, i.e., the height and width of the 2D convolution window, is 3. We set the pool size of the max-pooling layer as (2,2) in the implementation. Finally, we produce the compressed feature using a fully connected layer with the rectified linear unit (ReLU) activation function. The impact of dimensionality of fully connected layer output is discussed in Sec. 7.3. The

decoder part consists of one fully connected layer with a ReLU function that initially introduces nonlinearity. The data is upsampled using the transposed convolution layer with a 2D kernel. We set filters=64, kernel size=(3, 3), and strides=2. The final fully connected layer generates the reconstructed spectrogram using ReLU as the activation function.

6.1.2 Loss Definition

The critical issue is how to train the filter discussed above. The loss function combined with multiple loss functions for different tasks is widely used in the multi-goal learning process, such as transfer learning [23]. We consider the independent loss function for each task and summary the overall loss.

First, the reconstructed data should achieve great performance on the activity recognizer. We use the cross-entropy as the loss function of activity classifier (C_a), which is given as:

$$\text{loss}_a = -\frac{1}{|\mathcal{D}|}\sum_{x_i\in\mathcal{D}}\sum_{n=1}^{N} d_{i,n}\log C_a(F(x_i)), \tag{5}$$

where N is the number of gestures, $d_{i,n}$ is a binary variable that indicates whether the sample i belongs to the class n. We train the activity classifier by minimizing loss_a so that the classifier can predict the gestures correctly. Second, we aim to preserve users' identity information, of which the goal is to prevent distinguishing users. We amend the loss function of identification classifier (C_u), which is given as:

$$\text{loss}_u = \frac{1}{|\mathcal{D}|}\sum_{x_i\in\mathcal{D}}\sum_{m=1}^{M} d_{i,m}\log C_u(F(x_i)), \tag{6}$$

where M is the number of users, $d_{i,m}$ is a binary variable that indicates whether the sample i belongs to the user m. When minimizing loss_u, the user authentication model can not work with the generating data. Combining the Equation (5) and (6), we have the overall loss function of privacy preserving network:

$$\text{loss} = \text{loss}_a + \alpha\text{loss}_u, \tag{7}$$

where α is the trade-off parameters.

In the implementation, we adopt the same network of ThumbUp as the G_a and G_u. More specifically, we train a model as G_a, and we train a G_u for each gesture; we have N G_u models. In G_a, we only splice two layers of the Bi-LSTM structure and set the dimensionality of the Bi-LSTM output space as 64 and 32. We save the parameter weights for models and use the fine-tuning method introduced in Sec. 5.2.2 in the training process, in which we freeze the loaded models and only train the layers in the filter to minimize overall loss.

Numerous wristwatch programs capture and upload IMU signals to the server as the primary input data, including fall detection, gait counting, and other activity trackers [24]. In Sec. 5.1, our user identification model demonstrates that IMU signals expose the user's personal information, and so pose a danger of privacy disclosure. For instance, if a user uses our system as the unlock scheme, an attacker can tap the data package used for online activity recognition programs and utilize the data to compromise the smartwatch. With our identity filter, we inverse the reconstructed STFT spectrogram to sequence signal and then upload the reconstructed signal instead of raw data for other activity trackers, which successfully protect users' privacy and ensure activity trackers function correctly.

6.2 Contextual Command Analysis

People have potential behavior patterns in daily life, such as fixed mealtimes, the loop of switches, and the sequence of actions like opening curtains and windows. When using gesture commands to control smart homes, mining the contextual correlation between commands improves command recognition precision. However, to achieve real-time control, a significant problem is finding an effective way to achieve accurate recognition and reduce computational complexity simultaneously.

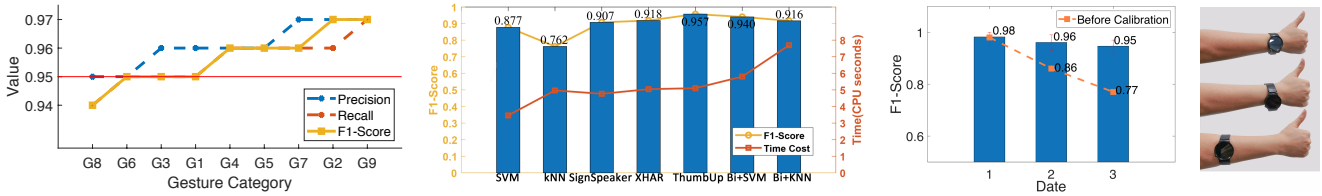
Regarding contextual inference methods, Hidden Markov Model (HMM) is the most commonly used method and has achieved great results in many scenarios. However, in our command process, we do not infer the current command only based on the history of the command alone. We aim to identify whether the result predicted by the deep model is plausible with assisted past commands. Decision-tree-based algorithms are widely used in the field of HAR due to the advantages of simplicity, accuracy, and interpretability [25], which suits our scenarios very well. Caros *et al.* [26] present a decision-tree-based light-weight approach for real-time human activity classification. Random Forest (RF) contains multiple decision trees and outputs the class label based on the results from trees, which determines the importance of inputs without dimensionality reduction and feature selection [27]. RF algorithm calculates the number of votes received by each prediction target. The prediction target that gets the highest number of votes is used as the final prediction, of which the RF algorithm has a high resistance to over-fitting. We adopt the RF classifier to analyze the contextual sequential pattern of the commands. Without additional sensors, we combine two inputs: the deep model recognition results as the recognition feature and t historical commands as the contextual feature. The model is quite a lightweight method with a fast training speed. Besides, we use the process introduced in Sec. 5.2.1 to update the recognition model and decision tree. It is worth mentioning that contextual analysis is applicable in the case of command correlations such as home control and industrial production. In the open scenario, we only adopt the gesture recognition model introduced above when the effect of sequential classification is not satisfactory.

7 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 in user identification and gesture recoginition with micro and macro benchmarks. Then, we explore the robustness of authentication with imitation attacks. We evaluate ThumbUp in home control on the real dataset in smart home. 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.

7.1 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



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

Fig. 7. (a) Identification accuracy, (b) Comparisons, (c) 3 Periods w & w/o calibration, and (d) Placement.

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. We build the trained model in the TensorFlow Lite framework and employ our system on the Android mobile platform for real-time evaluation.

7.2 User Identification

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, $6 > 50$s). 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.

7.2.1 Classification Accuracy

We first investigate the accuracy of our system across multiple users. Before the experiments, we briefly explained our system and showed the participants the example photos of 9 gestures (illustrated in Fig. 1). We ask participants to wear the smartwatch on their dominant hands, maintaining a comfortable tightness. Before the data collection, the participants are asked to practice the gestures a few times. Once comfortable, each participant is asked to perform 9 gestures with 20 repetitions. Participants are free to sit or stand while trying to avoid large body movements. We have $65 \times 9 \times 20$ gestures in the dataset.

(a) Gesture Collection (b) Imitate Attack

Fig. 8. Illustrations of Dataset Collection.

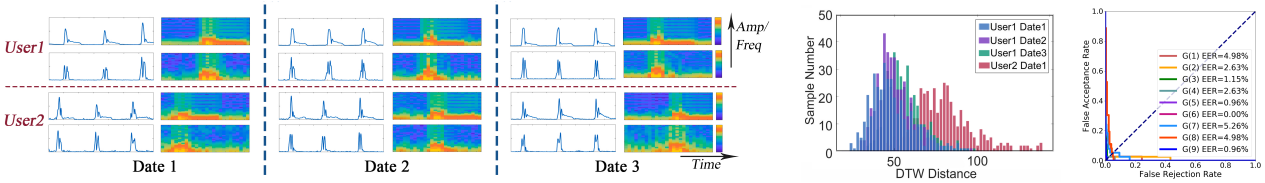
We evaluate the identification quality of the 9 gestures by precision, recall, and F1-Score. For each gesture, we repeat the training process for 10 times by randomly selecting 10 of 20 samples as the training samples and compute the average results in the rest 10 samples. The results (shown in Fig. 7(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. We compare our design with state-of-the-art baselines: SVM [28], kNN [29], SignSpeaker [30], and XHAR [23]. Moreover, we use the features extracted by BiLSTM as the input of traditional classification algorithms (SVM, KNN), which achieves higher accuracy than the original spectrogram and shows the effectiveness of our model for feature extraction. The result (shown in Fig. 7(b)) displays that our model achieves the highest accuracy on our dataset with acceptable computation cost.

Impact of parameters configuration: For the input spectrogram of extracted features, our model reaches the best performance with the 128 widths of a sliding window (choose from [256,128,64,32]), 8 for increment (choose from [32,16,8,4]). We perform ablation studies to know the importance of various components in the model (shown in Table 1). We verify that the good performance of our model mostly results from using the sSE network and using 3-layer BiLSTM. We observe that 4-layer BiLSTMs achieve comparable accuracy to 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 [31] and scSE [19]. Also, we find that the Batch Normalization layer and Dropout layer have a significant effect on the stability and generalization ability of the model.

TABLE 1
Results of the proposed model with different switch configurations.

Models	Highest F1-score	Lowest F1-score	Average F1-score
DT	0.90	0.96	0.935
RF	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**

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,



(a) (b) (c)

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

we add the evolving mechanism described in Sec. 5.2.2. 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. 7(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 body motion: The body motion of users creates non-zero acceleration readings. We asked 5 participants to wear the smartwatch and perform gesture G9 under different motion states: stilling the body's locomotion (sitting) and sustained movement (walking and running). Evaluations show that when the user body is at a static position, the average F1-score is 0.93. However, with body sustained movement, the F1-score comes down to 0.67 (walking) and 0.42 (running). The results reveal that large body motion brings a huge disruption to our system.

Impact of placement: In everyday life, users may not 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. 7(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 with a loose band.

7.2.2 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.

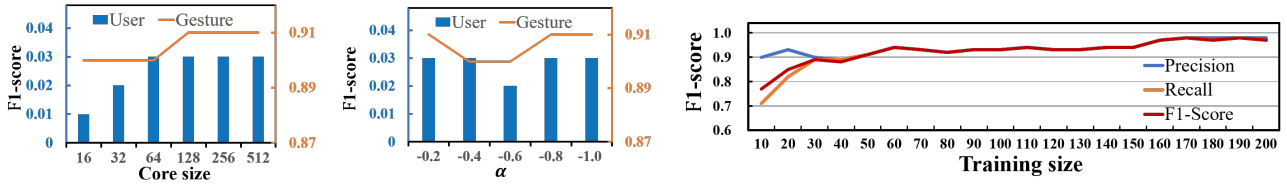
We calibrate the threshold σ in the authentication mechanism to observe the False Rejection Rate (FRR) and False Acceptance Rate (FAR). The Receiver Operating Characteristic (ROC) curve of one user is shown in Fig. 9(c). We summarize the average Equal Error Rates (EER) for the nine gestures in Table 4. 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.

We compare our authentication performance with state of the art one-class classifiers: GAN [32] and Autoencoder+SVM [33]. Compared to 0.221 for GAN and 0.173 for AE+SVM, our design achieves the lowest average EER, which is 0.025. We suspect the number of training samples is too few for a one-class deep neural network classifier. ThumbUp is trained by the number 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. Then, we compare our work with state-of-the-art authentication methods based on IMU signals. As shown in Fig. 9(b), the DTW distance from different users, which is used in [7] and needs fewer pre-detected samples, cannot be distinguished. The authentication system proposed in [34] achieved an EER of 0.054, which offers comparable authentication performance.

We explore the effectiveness of our model when valid users perform unknown gestures. In our design, samples of unknown gestures should be determined as an illegal sample, even from valid users. We collect 20 unknown gesture samples from 5 legitimate users and 20 samples from G1 to G9 as the inputs of corresponding authentication models. The average EER among 9 models is 0.008, which proves the effectiveness of our method for determining unknown gestures.

7.3 Command Recognition

First, we evaluate the accuracy of our model for gesture recognition. We choose four gestures (G1-G4) from the dataset collected from 65 participants as described above and use half of the samples from each user as the training dataset and the rest as the test dataset. We compare our method with the state of art DNN methods: the LSTM based classification model proposed in [35] and the CNN-IMU model introduced in [36]. We also compare with traditional algorithms such as SVM and random forest (RF) classifier. The results shown in Table 2 demonstrate that our model achieves the highest accuracy of 91%.



(a) (b) (c)

Fig. 10. Impact of (a) core size and (b) loss parameter α for user identity filter. And (c) the impact of training size in contextual command analysis

TABLE 2
Comparisons with state of art methods.

Models	Precision	Recall	F1-score
LSTM [35]	0.80	0.80	0.80
CNN-IMU [36]	0.83	0.83	0.83
RF	0.84	0.83	0.84
SVM	0.87	0.86	0.87
ThumbUp w/o Filter	0.88	0.87	0.88
ThumbUp	**0.91**	**0.91**	**0.91**

Besides, the experiments indicate that our filter process effectively increases the accuracy from 88% to 91%. We evaluate the performance of the filter over different core sizes. The result, shown in Fig. 10(a), illustrates that the medial compressed feature with a larger core size provides more information of both user and activity. Next, we study the impact of loss parameter, i.e., α, in the model training process with core size equals 128, as shown in Fig. 10(b). In our dataset, when $\alpha = -0.6$, the filter achieves the lowest user accuracy (F1-score=0.02) with a F1-score of 0.93 for gesture classification. In the user privacy protection task, adding noise is the most common method [37]. We compare our filter with a commonly used noise method: adding uniform noise. The result shows that when the Fi-score of user classification is 0.02, the noise method achieves 0.37 accuracy for gesture classification (0.91 for ours).

TABLE 3
Performance of Contextual Command Analysis.

Models	Command Latency t						
	1	2	3	4	6	8	10
SeqDT [25]	0.85	0.88	0.85	0.91	0.81	0.86	0.78
HMM	0.41	0.66	0.73	0.77	0.77	0.77	0.72
LR	0.91	0.90	0.90	0.87	0.86	0.86	0.82
DT	0.94	0.93	0.94	0.92	0.94	0.93	0.91
RF	0.95	0.95	**0.96**	0.94	0.94	0.93	0.93

To obtain a further evaluation of our models in smart home control, in this work, we have experimented with the dataset proposed in [38]. We map the collected gesture samples to four activity labels (Toileting, Breakfast, Lunch and Dinner), representing the control of the related activities. After that, we have a sequential command whose length is 400. We compare the performance of our method (RF) with state of the art sequential prediction methods: SeqDT [25], HMM, Logistic Regression (LR) and Decision Tree (DT). We evaluate the impact of command latency t, and the results of F1-Score is shown in Table 3, where we observe that RF always performs the best for different numbers of command latency over other the baseline methods. Specifically, the RF increases the accuracy from 91% to 96% when $t = 3$. Besides, the decision tree algorithm works well for different command latency. Moreover, Fig. 10(c) presents the comparison results on the training size on our model with the same test dataset. The result shows that when training size reaches 50, our model leads to a 90% accuracy. As the number of training samples increases, the accuracy of our model increases drastically.

7.4 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.

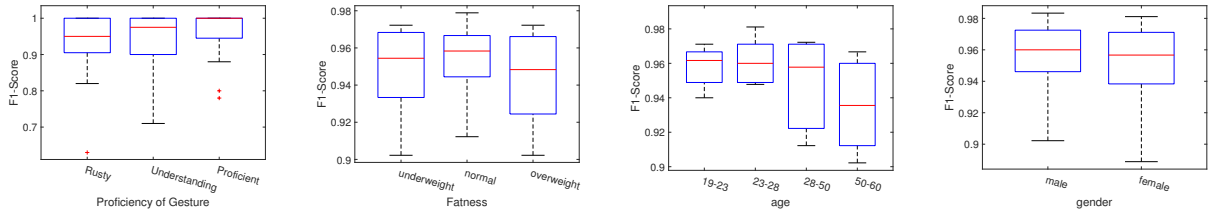
7.5 User Study

We analyzed the impact of user factors using our system in user identification task.

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. 11(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 [39]. We divide participants into 3 categories (underweight, normal, and overweight) (13:24:12) according to [40]. As shown in Fig. 11(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 and may affect the accuracy.

Impact of age and gender: We divided participants into 4 categories by age (6:10:32:16). According to the result shown in Fig. 11(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 and 30 females that cover the age range from 19 to 60 separately. In Fig. 11(d), the F1-Scores of male and female are roughly the same.



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

Fig. 11. Impact of proficiency, fatness, age, and gender

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 4. 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 4) with both user-friendliness and usability. Considering the relationship between proficiency and accuracy of gestures (shown in Fig. 11(a)), users can redefine their personal unlock gestures with the most familiar gestures for better security.

TABLE 4
F1-score of identification, EER of Imitate-Attack, and the rank of user-friendliness for each gesture.

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

8 Related Work

Existing biometrics classification approaches can be divided into two categories: *physiological* and *behavioral* techniques. Physiological techniques take advantage of the physical characteristics of the human body [28]. Behavioral techniques utilize unique manners, such as kinesiological movements [41] and even tongue movement [42], which are closely related to personal behavioral habits. PerAE [43] is an identity recognition system based on the electrocardiogram; It maintains an AE module to classify the heartbeats of other users as anomalies.

Gesture-based recognition have drawn great attention in academia and industry, with sensor signal based on capacitance [44], wireless backscattering [45], cameras [46], *etc.* Yang *et al.* [47] present the study of mobile authentication using free-form touchscreen gestures generated by participants instead of text passwords. Some gesture-based studies use IMU: Authors in [48] undertake an investigative analysis to study the feasibility and practical deployability of handwriting-based authentication techniques that utilize motion sensors. Sun *et al.* [49] propose a 3D 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 [6] uses the arm movement signals measured by wrist-worn smart devices, which is similar to our design. It authenticates with 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 [50] 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 [7] proposes a secure PIN input system, which extends a virtual number pad on the back of hands with smart wristbands. It uses tapping vibrometry as biometrics with an authentication accuracy of 96% for 128 users. WatchAuth [51] also shows the tap gesture's biometric capability to authenticate users and recognize intent-to-pay simultaneously. Li *et al.* [52] investigates the feasibility of authenticating users by sensing hand motions of signing their names in the air using fingers, which achieves 0.83% EER against insider adversaries.

9 Discussion

Nevertheless, this is only the first step toward completing an extremely difficult assignment. Several open research questions remain as follows:

Accuracy Improvement: Current recognition accuracy is restricted to two main issues. On the one hand, the signals of small gestures are weaker and more susceptible to perturbation than a large range of body movements; it is very difficult to extract user identity features from small gesture movements. On the other hand, we only collect adequate user samples in cold-start conditions. We designed the system considering that asking users to collect too much gesture sample data during the cold start phase would decrease user-friendliness, so we set the number of data used for initial model recognition to 10. We explore the effect of the number of training samples in Fig. 10(c). The result demonstrates that as the number of training samples increases, the accuracy of our model increases drastically. Thus, with the self-calibration section, we can perform user data updates with less computational overhead.

Body Motion: The evaluation in Sec. 7.2.1 reveals that the body motion of users brings a huge disruption to our system as it creates non-zero acceleration readings. Enhancing the system's robustness in a more hostile environment when users engage in

other daily activities, such as walking and running, is necessary to improve the usability of our system. Using the periodicity of body movements during continuous body motion to filter out noise signals is a potentially viable practice [53].

10 Conclusion

In this work, we present ThumbUp for identifying and authenticating users with only a single basic gesture, such as a thumbs-up, and we introduce the extension of ThumbUp in smart home control. We carefully design pre-processing methods for reducing noisy, weak inputs to a spectrogram containing user characteristics. A light-weight robust deep neural network is used to extract unique representations from motion signals. We illustrate its utility through extensive experimental investigations conducted over a three-month period with 65 users. We believe that our approach will open up a wide range of exciting opportunities for convenient and safe authentication using wearable smart devices.

Acknowledgments

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.62132018, Key Research Program of Frontier Sciences, CAS. No.QYZDY-SSW-JSC002, The University Synergy Innovation Program of Anhui Province with No.GXXT-2019-024.

References

[1] "Statista, howpublished = https://www.statista.com/statistics/920679/smart-home-device-shipments-worldwide-by-category/."

[2] J. Li, L. Deng, Y. Gong, and R. Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[3] M. M. Luna, T. P. Carvalho, F. A. A. Soares, H. A. Nascimento, and R. M. Costa, "Wrist player: a smartwatch gesture controller for smart tvs," in *41st Annual Computer Software and Applications Conference (COMPSAC)*, vol. 2. IEEE, 2017, pp. 336–341.

[4] F. Wang, J. Feng, Y. Zhao, X. Zhang, S. Zhang, and J. Han, "Joint activity recognition and indoor localization with wifi fingerprints," *IEEE Access*, vol. 7, pp. 80 058–80 068, 2019.

[5] S. Kratz and M. Rohs, "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*. ACM, 2010, pp. 341–344.

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

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

[8] A. S. Kundu, O. Mazumder, P. K. Lenka, and S. Bhaumik, "Hand gesture recognition based omnidirectional wheelchair control using imu and emg sensors," *Journal of Intelligent & Robotic Systems*, vol. 91, no. 3, pp. 529–541, 2018.

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

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

[11] C. Xu, P. H. Pathak, and P. Mohapatra, "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*. ACM, 2015, pp. 9–14.

[12] X. Yu, Z. Zhou, M. Xu, X. You, and X.-Y. Li, "Thumbup: Identification and authentication by smartwatch using simple hand gestures," in *2020 IEEE International Conference on Pervasive Computing and Communications (PerCom)*. IEEE Computer Society, 2020, pp. 1–10.

[13] W. H. Press and S. A. Teukolsky, "Savitzky-golay smoothing filters," *Computers in Physics*, vol. 4, no. 6, pp. 669–672, 1990.

[14] M. Muaaz and R. Mayrhofer, "Smartphone-based gait recognition: From authentication to imitation," *IEEE Transactions on Mobile Computing*, vol. 16, no. 11, pp. 3209–3221, 2017.

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

[16] D. Griffin and J. Lim, "Signal estimation from modified short-time fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 32, no. 2, pp. 236–243, 1984.

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

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

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

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

[21] Q. Wang, L. Ye, H. Luo, A. Men, F. Zhao, and Y. Huang, "Pedestrian stride-length estimation based on lstm and denoising autoencoders," *Sensors*, vol. 19, no. 4, p. 840, 2019.

[22] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," *arXiv preprint arXiv:1312.6114*, 2013.

[23] Z. Zhou, Y. Zhang, X. Yu, P. Yang, X.-Y. Li, J. Zhao, and H. Zhou, "Xhar: Deep domain adaptation for human activity recognition with smart devices," in *17th Annual IEEE International Conference on Sensing, Communication, and Networking (SECON)*. IEEE, 2020, pp. 1–9.

[24] S. Ashry, T. Ogawa, and W. Gomaa, "Charm-deep: continuous human activity recognition model based on deep neural network using imu sensors of smartwatch," *IEEE Sensors Journal*, vol. 20, no. 15, pp. 8757–8770, 2020.

[25] Z. He, Z. Wu, G. Xu, Y. Liu, and Q. Zou, "Decision tree for sequences," *IEEE Transactions on Knowledge and Data Engineering*, 2021.

[26] J. S. Caros, O. Chételat, P. Celka, S. Dasen, and J. CmÃral, "Very low complexity algorithm for ambulatory activity classification," in *3rd European Medical and Biological Conference EMBEC*. Citeseer, 2005, pp. 16–20.

[27] M. Javeed, A. Jalal, and K. Kim, "Wearable sensors based exertion recognition using statistical features and random forest for physical healthcare monitoring," in *International Bhurban Conference on Applied Sciences and Technologies (IBCAST)*. IEEE, 2021, pp. 512–517.

[28] B. Zhou, J. Lohokare, R. Gao, and F. Ye, "Echoprint: Two-factor authentication using acoustics and vision on smartphones," in *Proceedings of the 24th Annual International Conference on Mobile Computing and Networking*. ACM, 2018, pp. 321–336.

[29] P. Palimkar, V. Bajaj, A. K. Mal, R. N. Shaw, and A. Ghosh, "Unique action identifier by using magnetometer, accelerometer and gyroscope: Knn approach," in *Advanced Computing and Intelligent Technologies*. Springer, 2022, pp. 607–631.

[30] J. Hou, X.-Y. Li, P. Zhu, Z. Wang, Y. Wang, J. Qian, and P. Yang, "Signspeaker: A real-time, high-precision smartwatch-based sign language translator," in *The 25th Annual International Conference on Mobile Computing and Networking*, 2019, pp. 1–15.

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

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

[33] Y. Zou, M. Zhao, Z. Zhou, J. Lin, M. Li, and K. Wu, "Bilock: User authentication via dental occlusion biometrics," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 3, p. 152, 2018.

[34] G. Li, L. Zhang, and H. Sato, "In-air signature authentication using smartwatch motion sensors," in *2021 IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*. IEEE, 2021, pp. 386–395.

[35] O. Steven Eyobu and D. S. Han, "Feature representation and data augmentation for human activity classification based on wearable imu sensor data using a deep lstm neural network," *Sensors*, vol. 18, no. 9, p. 2892, 2018.

[36] F. Moya Rueda, R. Grzeszick, G. A. Fink, S. Feldhorst, and M. Ten Hompel, "Convolutional neural networks for human activity recognition using body-worn sensors," in *Informatics*, vol. 5, no. 2. Multidisciplinary Digital Publishing Institute, 2018, p. 26.

[37] M. Alaggan, S. Gambs, and A.-M. Kermarrec, "Heterogeneous differential privacy," 2015.

[38] F. Ordóñez, P. De Toledo, A. Sanchis *et al.*, "Activity recognition using hybrid generative/discriminative models on home environments using binary sensors," *Sensors*, vol. 13, no. 5, pp. 5460–5477, 2013.

[39] I. Janssen, S. B. Heymsfield, D. B. Allison, D. P. Kotler, and R. Ross, "Body mass index and waist circumference independently contribute to the prediction of nonabdominal, abdominal subcutaneous, and visceral fat," *The American journal of clinical nutrition*, vol. 75, no. 4, pp. 683–688, 2002.

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

[41] I. Olade, C. Fleming, and H.-N. Liang, "Biomove: Biometric user identification from human kinesiological movements for virtual reality systems," *Sensors*, vol. 20, no. 10, p. 2944, 2020.

[42] P. Nguyen, N. Bui, A. Nguyen, H. Truong, A. Suresh, M. Whitlock, D. Pham, T. Dinh, and T. Vu, "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*. ACM, 2018, pp. 269–282.

[43] L. Sun, Z. Zhong, Z. Qu, and N. Xiong, "Perae: an effective personalized autoencoder for ecg-based biometric in augmented reality system," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 6, pp. 2435–2446, 2022.

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

[45] M. Yin, X.-Y. Li, Y. Zhang, P. Yang, and C. Wan, "Back-guard: Wireless backscattering based user activity recognition and identification with parallel attention model," in *28th International Symposium on Quality of Service (IWQoS)*. IEEE, 2020, pp. 1–10.

[46] K. Guo, H. Zhou, Y. Tian, W. Zhou, Y. Ji, and X.-Y. Li, "Mudra: A multi-modal smartwatch interactive system with hand gesture recognition and user identification," in *IEEE Conference on Computer Communications*, 2022, pp. 100–109.

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

[48] R. Wijewickrama, A. Maiti, and M. Jadliwala, "Write to know: on the feasibility of wrist motion based user-authentication from handwriting," in *Proceedings of the 14th ACM Conference on Security and Privacy in Wireless and Mobile Networks*, 2021, pp. 335–346.

[49] Z. Sun, Y. Wang, G. Qu, and Z. Zhou, "A 3-d hand gesture signature based biometric authentication system for smartphones," *Security and Communication Networks*, vol. 9, no. 11, pp. 1359–1373, 2016.

[50] H.-M. C. Leung, C.-W. Fu, and P.-A. Heng, "TwistIn: Tangible authentication of smart devices via motion co-analysis with a smartwatch," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 2, no. 2, p. 72, 2018.

[51] J. Sturgess, S. Eberz, I. Sluganovic, and I. Martinovic, "Watchauth: user authentication and intent recognition in mobile payments using a smartwatch," *arXiv preprint arXiv:2202.01736*, 2022.

[52] G. Li and H. Sato, "Sensing in-air signature motions using smartwatch: A high-precision approach of behavioral authentication," *IEEE Access*, 2022.

[53] F. Adib, H. Mao, Z. Kabelac, D. Katabi, and R. C. Miller, "Smart homes that monitor breathing and heart rate," in *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, 2015, pp. 837–846.

Xiaojing Yu received the BE degree in 2018, from the College of Computer Science and Technology, University of Science and Technology of China (USTC), where she is currently working toward the PhD degree. Her research interests include mobile computing, human interaction, and GIS management.

Zhijun Zhou received the bachelor's degree in Internet of Things, in 2018, from the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics (NUAA). And she received the master's degree in computer application technology, in 2021, from the College of Computer Science and Technology, University of Science and Technology of China (USTC), China. Her research interests include intelligent sensing and ubiquitous computing.

Lan Zhang (IEEE Fellow) received the bachelors degree from the School of Software, and the PhD degree from the Department of Computer Science and Technology, Tsinghua University, China, in 2007, 2014, respectively. She is currently a research professor in the School of Computer Science and Technology, University of Science and Technology of China. Her research interests include data trading, privacy protection, and mobile computing, etc.

Xiang-Yang Li (IEEE Fellow, ACM Fellow) received the bachelors degree from the Department of Computer Science and the second bachelors degree from the Department of Business Management, Tsinghua University, P.R. China, both in 1995, and the MS and PhD degrees from the Department of Computer Science, University of Illinois at Urbana Champaign, 2000, 2001, respectively. He is currently a professor and executive dean at the School of Computer Science and Technology, University of Science and Technology of China. His research interests include artificial intelligence of things, edge computing, privacy and security, data sharing and trading, and algorithms.