Deep Fisher Discriminant Learning for Mobile Hand Gesture Recognition

Ce Li, Chunyu Xie, Baochang Zhang, Chen Chen, Jungong Han

PII: S0031-3203(17)30519-8
DOI: 10.1016/j.patcog.2017.12.023
Reference: PR 6408

To appear in: *Pattern Recognition*

Received date: 4 July 2017
Revised date: 9 December 2017
Accepted date: 30 December 2017

Please cite this article as: Ce Li, Chunyu Xie, Baochang Zhang, Chen Chen, Jungong Han, Deep Fisher Discriminant Learning for Mobile Hand Gesture Recognition, *Pattern Recognition* (2017), doi: 10.1016/j.patcog.2017.12.023

Highlights

- we collect a large mobile gesture database using an Andriod Huawei device, which is the largest database in published studies for mobile gesture recongnition systems.
- we incorporate Fisher criterion into BiLSTM network and propose F-BiLSTM and F-BiGRU to improve the traditional softmax loss training function.
- Extensive experiments on our MGD, BUAA Mobile Gesture database, and a public database are conducted to verify the superior performance of the proposed networks.

Deep Fisher Discriminant Learning for Mobile Hand Gesture Recognition

Ce Li[b,1], Chunyu Xie[a,1], Baochang Zhang[a,*], Chen Chen[c], Jungong Han[d]

[a] *Department of Automation, Beihang University, Beijing, China*
[b] *China University of Mining and Technology, Beijing, China*
[c] *University of Central Florida, Orlando, FL, USA.*
[d] *Lancaster University, Lancaster, UK.*

Abstract

Gesture recognition becomes a popular analytics tool for extracting the characteristics of user movement and enables numerous practical applications in the biometrics field. Despite recent advances in this technique, complex user interaction and the limited amount of data pose serious challenges to existing methods. In this paper, we present a novel approach for hand gesture recognition based on user interaction on mobile devices. We have developed two deep models by integrating Bidirectional Long-Short Term Memory (BiLSTM) network and Bidirectional Gated Recurrent Unit (BiGRU) with Fisher criterion, termed as F-BiLSTM and F-BiGRU respectively. These two Fisher discriminative models can classify user's gesture effectively by analyzing the corresponding acceleration and angular velocity data of hand motion. In addition, we build a large Mobile Gesture Database (MGD) containing 5547 sequences of 12 gestures. With extensive experiments, we demonstrate the superior performance of the proposed method compared to the state-of-the-art BiLSTM and BiGRU on MGD database and two other benchmark databases (*i.e.*, BUAA mobile gesture and SmartWatch gesture). The source code and MGD database will be made publicly available at https://github.com/bczhangbczhang/Fisher-Discriminant-LSTM.

*Corresponding author
Email address: bczhang@buaa.edu.cn (Baochang Zhang)
[1] Ce Li and Chunyu Xie have equal contribution to the paper.

Keywords: Fisher Discriminant, Hand Gesture Recognition, Mobile Devices

1. Introduction

Human-computer interaction (HCI) is of great interest to researchers in biometrics. As an emerging HCI technology, gesture recognition demonstrates promising performance for extracting and analyzing the characteristics of user movement and is widely used in many applications, including behavioral biometric authentication, user verification, etc. [1, 2, 3]. With the emergence of modern smartphones, gesture recognition receives increasing attention, because it can easily obtain user's interaction with mobile devices by monitoring the combined activities captured by touch screen, camera, and microphone [4, 5, 6, 7]. However, due to complex surrounding environment, such methods may not perform well in practical scenarios. For example, video-based methods do not work well in the night time due to the camera limitation.

Alternatively, inertial sensors, such as accelerometer and gyrometer, are built in smartphones and can be used to record the hand motion signal [8, 9, 10]. The personalized gesture can be automatically acquired by accelerometer-based recognition solution [11]. Compared to vision-based solutions for gesture recognition [12], inertial sensors (*e.g.* accelerometer and gyrometer) are more robust under various lighting conditions [13]. However, the accuracy of these inertial sensors can be affected by different factors, including signal intensity differences (intense versus weak gestures), temporal variations (slow versus fast movements) and physical differences (users' physical conditions, etc.). In addition, noises from the sensing hardware pose extra challenges to the recognition task. To resolve these problems, different methods have been proposed, such as Support Vector Machine (SVM), Hidden Markov Model (HMM) and Dynamic Time Warping (DTW) [14, 15].

Recently, deep learning techniques have been successfully applied to the task of language modeling [16, 17], image captioning [18, 19], image classification [20], video analysis [21, 22], pose recovery [1, 23], and human activity

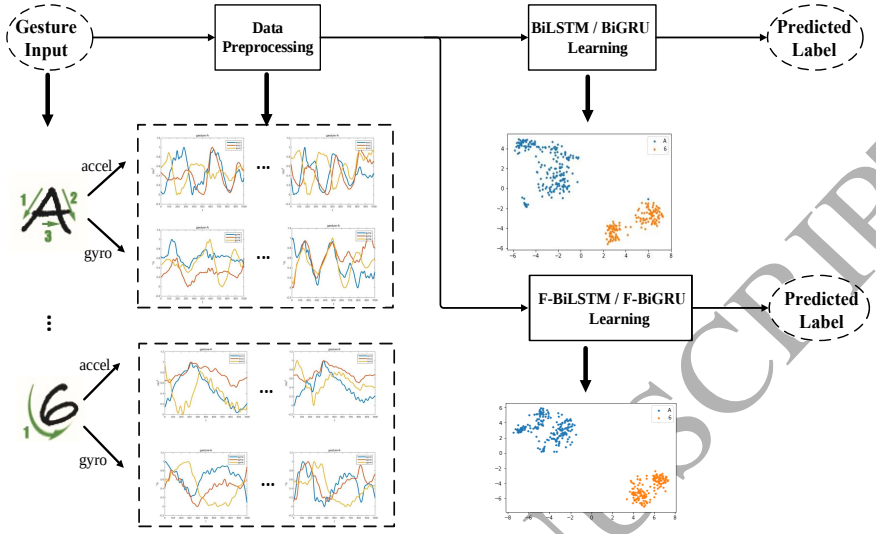


Figure 1: Flowchart of the proposed gesture recognition approach. We introduce Fisher criterion into BiLSTM and BiGRU network to improve the traditional softmax loss training function, which is able to minimize the intra-class variations and maximize the inter-class variations in the deep framework. For ease of display, we show the BiLSTM learned features and F-BiLSTM learned features of two classes gestures in two right subfigures.

recognition [7, 24, 25, 26, 27, 28, 29], etc. In particular, the effectiveness of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) [30] on modeling human gesture structure and temporal dynamics has been validated for automatic representing and classifying the complex sequential data simultaneously. Furthermore, to enhance the discrimination capability, different gating mechanisms are incorporated in LSTM, leading to GRU [31], BiLSTM [32, 33], and BiGRU [34], etc. In this paper, we use BiLSTM and BiGRU models considering the high performance and low memory requirement, to implement the gesture recognition on mobile devices by analyzing the sequential data streams captured from inertial sensors.

For gesture recognition task, deep features can be learned automatically via current RNN and LSTM based methods which yield more abstract and

useful representations. However, typically no distribution prior is embedded into the learning of deep features, making such schemes uncontrollable for certain circumstances. For example, due to large intra-class variations (speed, pattern of gesture) caused by different performers and small inter-class variations caused by similar gestures, it is impractical to pre-collect all the possible testing identities for training samples with heterogeneous accelerometer and gyrometer signals, the conventional loss functions used by RNN and LSTM based methods are not always suitable. It is also noticed that the features of more compact structure suit better in representing the data. Particularly when the data variation is large, less compactness on the feature representation might cause an inaccurate classification in the real-world applications [35, 36]. The above observations inspire us to adopt Fisher criterion for minimizing intra-class variations and maximizing inter-class variations when integrated with softmax loss of LSTM network, obtaining more capacity to cope with external variations. Based on bidirectional LSTM and GRU (a variant of LSTM) models, two deep Fisher discriminant learning models termed F-BiLSTM and a variant F-BiGRU are proposed for hand gesture recognition on mobile devices. The framework of the proposed gesture recognition approach is shown in Fig. 1.

Furthermore, it is important to build a comprehensive hand gesture database for mobile devices that allows researchers to develop algorithms and conduct the relevant evaluation. Though there exist some gesture databases captured from mobile devices for various applications. The available data are often limited to particular scenarios and fail to serve general purposes. In this paper, we introduce a mobile-based gesture recognition benchmark, which helps researchers to conveniently evaluate and compare their estimation results. We also build a large mobile based hand gesture database consisting of 12 classes of gestures including 5547 samples in total performed by 32 participants (23 males and 9 females). Each class of gestures has about 460 samples at different performing speed, so they are with heterogeneous accelerometer and gyrometer signals. The sampling time of accelerometer and gyrometer sensors is 5ms corresponding to a frequency of 200Hz. To the best of our knowledge, it is the largest database

so far for mobile-based gesture recognition, which is of benefit to the research community. In summary, the contributions of this paper are as follows:

1. We incorporate Fisher criterion into the BiLSTM and BiGRU networks termed as F-BiLSTM and F-BiGRU to improve the traditional softmax loss function for training. Extensive experiments show superior performance of the proposed method compared to the state-of-the-art BiLSTM and BiGRU on three gesture recognition databases.

2. We build a large hand gesture database for mobile hand gesture recognition.

The rest of the paper is organized as follows. Section 2 introduces the related works, and Section 3 describes the details of the proposed method. Experiments and results are presented in Section 4. Finally, Section 5 concludes the paper.

2. Related Work

Gesture Recognition on Mobile Devices. Gesture recognition has been extensively investigated for the last two decades with remarkable advances for the problem on mobile devices using inertial sensors [37, 38, 39, 40, 6, 5]. Rekimoto *et al.* proposed a gesture recognition method to detect arm movement using a specific wearable device [37]. The human moving dynamics are estimated by analyzing the dominating force to predict a user's moving direction, however, users have to wear a large-size device which is not practical for real-world applications. Afterwards, more researchers captured the part of human gestures of three dimensional acceleration signal by a small wireless sensor-box [39], a combination of EMG and ACC sensors [41], five miniature inertial and magnetic sensors worn on the chest, the arms, and the legs [4], a wrist accelerometer [42], and a Kinect sensor [43]. Recently, Agrawal *et al.* presented a system called PhonePoint Pen to use the built-in accelerometer in mobile phones to recognize human writing [44]. The results of 15 subjects running on mobile devices indicated that the English characters can be identified with an average accuracy of

91%, which has presented a promising prospect for mobile-based gesture recognition. Lefebvre *et al.* also carried out gesture recognition experiments on a database captured by an Android Nexus S Samsung device with 22 participants performing 14 symbolic hand gestures, to validate the combination of both accelerometer and gyrometer sensors can achieve better performance than using each individual sensor [45].

Gesture Recognition Using Classical Machine Learning Methods. Mobile gesture recognition provides new directions and delivers compelling performance for machine learning applications. Hofmann *et al.* proposed a recognition scheme based on discrete HMM (dHMM) to identify dynamic gestures [46], which essentially divides the input data into different regions and assigns each of them to a corresponding codebook for dHMM classification. The experiments are carried out using 500 training gestures with 10 samples per gesture, yielding an accuracy of 95.6% for 100 testing gestures. Kallio *et al.* also trained the dHMM model for the gestures of the 3-dimensional acceleration signal and measured the recognition accuracy of a system using four degrees of complexity [39]. Kela *et al.* tested an HMM model with five states and achieved the accuracy 96.1% for classifying 8 gestures [47]. Pylvanainen *et al.* proposed a method based on continuous HMM (cHMM) to achieve reliable performance with 96.67% of correct classification on a database of 20 samples for 10 gestures [48]. In the recent works, Zhang *et al.* utilized multi-stream HMM as a decision fusion function to recognize 18 classes of hand gestures, and got the average recognition accuracy 91.7% in real application. [41].

Besides the aforementioned HMM-based methods, a few other techniques are used in gesture recognition. Akl *et al.* employed Dynamic Time Warping (DTW) to define a dictionary of 18 gestures, and achieved classification accuracy 90% in the experiment [15]. David *et al.* compared Naive Bayes and DTW methods to recognize four gesture types from five different subjects, and demonstrated the advantage of Bayesian classification compared to DTW in the experiment [49]. Wu *et al.* used multi-class Support Vector Machine (SVM) for user-

independent gesture recognition and validated that SVM significantly outperforms other methods including DTW, Naive Bayes and HMM [50]. Wang *et al.* combined LCS and SVM to perform the classification task and achieve the classification accuracy of 93% [51]. Based on these works, Kerem *et al.* compared different classical machine learning methods for classifying human activities [4], in which the implemented and compared methods consisted of Bayesian Decision Making (BDM), Rule-Based Algorithm (RBA), Least-Squares Method (LSM), k-Nearest Neighbor algorithm (k-NN), DTW, SVM, and Artificial Neural Networks (ANN). Besides, some researchers focused on the application of feature selection and feature fusion, such as Principle Component Analysis (PCA) [52], fusion of the feature extracted from inertial and depth sensor [5], and hybrid features combining short-time energy with Fast Fourier Transform (FFT) [53].

Gesture Recognition Using Deep Learning Methods. Driven by the tremendous success of deep learning, the research paradigm has been shifted from traditional machine learning methods to deep learning methods for mobile gesture recognition, such as ANN [4, 45], RNN [54, 43], LSTM [55], and Convolutional Neural Network (CNN) [42]. Shin *et al.* developed a dynamic hand gesture recognition technique using recurrent neural network (RNN) algorithm, which was evaluated based on the gesture database captured by SmartWatch [54, 56]. Especially, for each gesture sequence containing 3-dimensional data of accelerometer, LSTM achieved the best performance with 128 neuro units in the experiment of SmartWatch gestures database. Gjoreski *et al.* compared deep CNN and Random Forest (RF) on two wrist gesture databases, and the results turn out that CNN slightly outperformed RF with sufficient data and achieved significantly better accuracy than other classical machine learning methods, including Naive Bayes, k-NN, Decision tree, and SVM [42]. Recently, Lefebvre *et al.* carried out gesture recognition experiments on a database consisting of both accelerometer and gyrometer sensors [45], and showed that the BiLSTM based method achieves an accuracy of 95.57% on the database of 1540 gestures. To the best of our knowledge, the BiLSTM based method is currently

the state-of-the-art baseline and performs better than previous approaches such as cHMM, DTW, SVM, and LSTM.

3. Deep Fisher Discriminant Learning

In this section, we first describe the network structures of Bidirectional Long-Short Term Memory (BiLSTM) and its variant with Bidirectional Gate Recurrent Unit (BiGRU). Then, we explain how our approaches, termed F-BiLSTM and F-BiGRU, incorporate the Fisher criterion to improve the discriminative power of these deep models, termed F-BiLSTM and F-BiGRU. For ease of explanation, we summarize the main variables and briefly describe them in Table 1.

Table 1: A brief description of variables used in the paper.

i_t: a sigmoidal input gate	f_t: a forget gate	o_t: an output gate
z_t: a update gate	r_t: a reset gate	$\tilde{h}_t$: a candidate output
c_t: a cell state	x_t: an input vector	h_t: a final output
W_*: all diagonal or weight matrices	b_*: all bias terms	μ_i: the ith class mean of output vectors
$\mathcal{L}_f$: the Fisher criterion loss	$\mathcal{L}_s$: the softmax loss	δ, θ, α: the scalar parameters

3.1. BiLSTM

We briefly describe the LSTM unit which is the basic building block of the proposed F-BiLSTM model. The neurons of LSTM contain a constant memory cell name, which has a state c_t at time t. A LSTM neuron unit is presented in detail at the bottom of Fig. 2. Each LSTM unit is controlled by a sequence of gates: a sigmoidal input gate i_t, a forget gate f_t and an output gate o_t. At each time step t, LSTM unit receives inputs from two external sources at each of the three gates. The external two sources are the current sample x_t and the previous hidden state h_{t-1}. The cell state c_{t-1} in the cell block is an internal source of each gate. The gates are passed through the tanh non-linearity and activated by the logistic function. After multiplying the cell state by the forget gate f_t, the

final output of the LSTM unit h_t is computed by multiplying the activation o_t of the output gates with updated cell state. We use W_* to represent all diagonal matrices and b_* to represent all bias terms. The updating procedure in a layer of LSTM units is summarized as follows:

$$\begin{aligned} i_t &= \sigma\left(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i\right), \\ f_t &= \sigma\left(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f\right), \\ c_t &= f_t c_{t-1} + i_t tanh\left(W_{xc}x_t + W_{hc}h_{t-1} + b_c\right), \\ o_t &= \sigma\left(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o\right), \\ h_t &= o_t tanh\left(c_t\right). \end{aligned} \tag{1}$$

The BiLSTM model using LSTM units is able to effectively model temporal data in many applications [16]. We consider the gesture data using 3 dimensional accelerometer and 3 dimensional gyrometer signals synchronized into an input vector through sampling time-steps. As shown in Fig. 2, the forward and backward LSTM hidden layers are fully connected to the input layer and consist of multiple LSTM neurons with full recurrent connections. Experiments are conducted with different hidden neuron sizes and 128 neurons yield satisfactory results. The output layer has a size equivalent to the number of neurons to classify (*i.e.* $M = 128$). $G = \{G_1, ..., G_T\}$ is a gesture sequence of T size; $G_t = (x_1(t), ..., x_N(t))$ is a vector at time step t; N denotes the sensor number; $(y_1, ..., y_n)$ is the BiLSTM output set with n being the number of gestures to be classified. The softmax activation function is used for this layer to give network a response between 0 and 1. Classically, these outputs can be considered as posterior probabilities of the input sequence belonging to a specific gesture class. The softmax loss function is defined as

$$\mathcal{L}_s = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{e^{W_{y_i}^T O_i + b_{y_i}}}{\sum_{j=1}^{n} e^{W_j^T O_i + b_j}}, \tag{2}$$

where $O_i = (o_1, ..., o_M)$ denotes the ith output belonging to the y_ith class. W_j denotes the jth column of the weights W in the last layer; b is the bias term; m is the size of mini-batch and n is the number of classes.

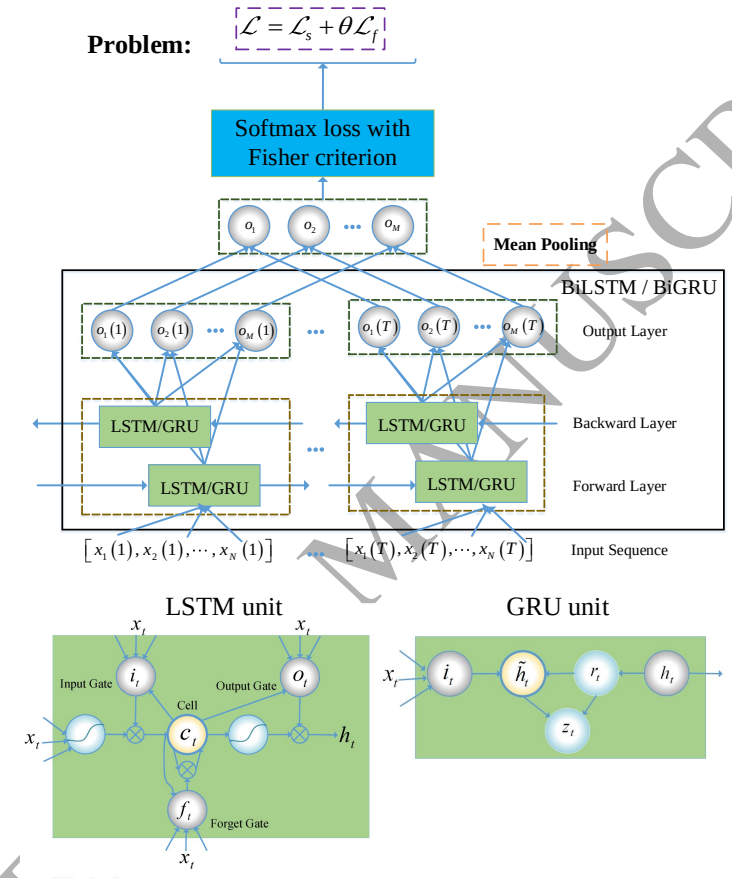


Figure 2: The architecture of F-BiLSTM/F-BiGRU: The input gesture vectors are learned and represented as the sequences via BiLSTM or BiGRU, then the Fisher criterion is proposed to be a new loss function in the fully connected layer, leading to a better performance without affecting the training convergence and model size.

3.2. F-BiLSTM

To further enhance the performance of BiLSTM, we incorporate the Fisher criterion into the softmax loss function, which is shown in Fig. 2. First, the input layer consists of the concatenation of 3-dimensional accelerometer and 3-dimensional gyrometer signals synchronized in time (*i.e.* $N = 6$). The sensor data is normalized between 0 and 1 according to the maximum value that sensors can capture. In order to minimize the intra-class variations and maximize the inter-class variations of gesture data, we propose a new Fisher criterion based on Fisher Linear Discrimination as follows:

$$\mathcal{L}_f = \frac{1}{m}\sum_{i=1}^{m} \|O_i - \mu_{y_i}\|_2^2 - \frac{\delta}{n\,(n-1)} \sum_{j=1,k=1}^{n} \|\mu_j - \mu_k\|_2^2 \tag{3}$$

where μ_{y_i} is the y_ith class mean of output vectors, and δ is the discriminative factor. To learn BiLSTM, the Fisher criterion utilizes the whole training set and mean vectors μ_{y_i} of each class in each iteration as the mean vector updates. We propose to augment the loss in Eq. (2) with the additional Fisher criterion term in Eq. (3) as follows:

$$\mathcal{L} = \mathcal{L}_s + \theta\mathcal{L}_f \tag{4}$$

where θ is bounded within [0,1] to control the Fisher criterion in Eq. (4), and δ is restricted in a more subtle interval [1e-5,0.1] to balance the intra-class distance and inter-class distance in the Fisher criterion. These two parameters are used to balance the three parts of the loss function. In forward and backward processes, we set output vector O_i, mean vector μ_j, loss parameter W, scalar parameters θ, δ and learning rate λ, BiLSTM parameters H_f and iteration number e, respectively. In each iteration, we compute the loss of F-BiLSTM by Eq. (3) and Eq. (4), and the backpropagation error by

$$\frac{\partial\mathcal{L}^e}{\partial O_i^e} = \frac{\partial\mathcal{L}_s^e}{\partial O_i^e} + \theta\frac{\partial\mathcal{L}_f^e}{\partial O_i^e}. \tag{5}$$

Then, we update the parameter W, mean vector μ_j and BiLSTM parameter H_f in the $e+1$ iteration by the following formulas until a convergence is reached.

$$\begin{aligned} W^{e+1} &= W^e - \lambda^e \cdot \frac{\partial \mathcal{L}_f^e}{\partial W^e}, \\ \mu_j^{e+1} &= \mu_j^e - \alpha \cdot \Delta \mu_j^e, \\ H_f^{e+1} &= H_f^e - \lambda^e \sum_i^m \frac{\partial \mathcal{L}^e}{\partial O_i^e} \cdot \frac{\partial O_i^e}{\partial H_f^e}. \end{aligned} \tag{6}$$

With optimized parameters θ, δ and α, the discriminative power of F-BiLSTM can significantly enhance hand gesture recognition. This network is learned by the online backpropagation through time with momentum. To classify a testing gesture sequence, we use a rule of keeping only the most probable class $argmax_{i\in[1,n]}O_i$ to determine the final gesture class. The details of parameter analysis on θ, δ and α are presented in Section 4.3.

3.3. F-BiGRU

We also investigated Fisher criterion into Bidirectional Gated Recurrent Unit (BiGRU) as shown in Fig. 2. BiGRU organizes the recurrent units in the way that each unit adaptively captures dependence of different time scales [57, 58]. Similar to the BiLSTM unit, the BiGRU has the output of the GRU h_t, candidate gate $\tilde{h}_t$, update gate z_t and reset gate r_t units to modulate the information flow without separate memory cells, as shown at the bottom right of Fig. 2. The updating flows of GRU in BiGRU differ with the one described in Eq.(1), and can be summarized as follows:

$$\begin{aligned} z_t &= \sigma\left(W_z x_t + W_{zf} h_t + b_z\right), \\ r_t &= \sigma\left(W_r x_t + W_{rf} h_t + b_r\right), \\ \tilde{h}_t &= tanh\left(W x_t + U\left(r_t \odot h_{t-1}\right) + b_h\right), \\ h_t &= \left(1 - z_t\right) h_{t-1} + z_t \tilde{h}_t \end{aligned} . \tag{7}$$

where the output h_t at time t is a linear interpolation between the previous forget gate h_{t-1} and the candidate gate $\tilde{h}_t$ computed in the same way as traditional recurrent unit. The update gate z_t determines the number of units for updating its forget gate, and the reset gate r_t.

Similar to F-BiLSTM, we also apply the Fisher criterian for BiGRU and learn a new variant named F-BiGRU to recognize hand gestures. The learning process for F-BiGRU is similar to F-BiLSTM with the same loss function as Eq.(4). The parameter updating procedures for W and mean vector μ are same as F-BiGRU (*i.e.*, same as the first and second formulas in Eq.(6)), while the BiGRU parameter H_B (the set of all output h_t) in the $(e+1)$th iteration is updated as:

$$H_B^{e+1} = H_B^e - \lambda^e \sum\nolimits_i^m \frac{\partial \mathcal{L}^e}{\partial O_i^e} \cdot \frac{\partial O_i^e}{\partial H_B^e}. \tag{8}$$

4. Experiments

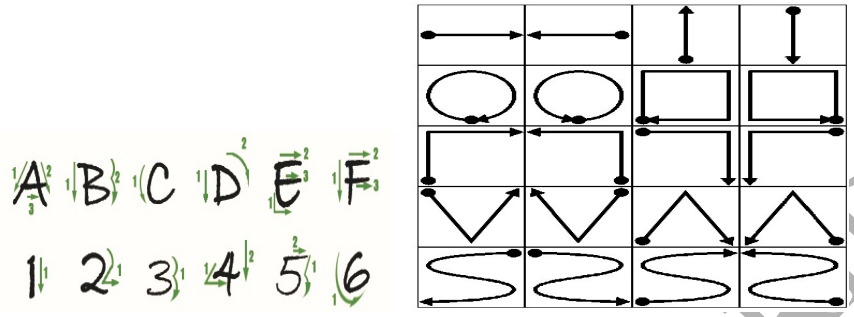
4.1. Hardware Device

Our mobile hand gesture database is collected using the Android system on a Huawei mobile phone, which has a 3D accelerometer and a gyrometer. According to [45], we collect the data from both accelerometer and gyrometer, and record each gesture by pressing, holding and releasing the "Sensor" button on the touch screen.

4.1.1. Data Collection

As shown in Fig. 3(a), the gesture database is composed of two categories: Arabic numerals (1, 2, 3, 4, 5, 6) and English capital letters (A, B, C, D, E, F). Furthermore, the stroke order of gestures is set in advance to ensure the consistency of gestures captured on the left or right hand of each participant.

The collected MGD consists of 12 gestures performed by 32 participants (23 males and 9 females) with about fifteen times per gesture. Each class of gestures has about 460 samples at different performing speeds, and there are a total of 5547 gesture sequences with heterogeneous accelerometer and gyrometer signals.The sampling time of accelerometer and gyrometer sensors is 5ms corresponding to a frequency of 200Hz. To the best of our knowledge, it is the largest database so far for mobile-based gesture recognition, which is of benefit to the research community.



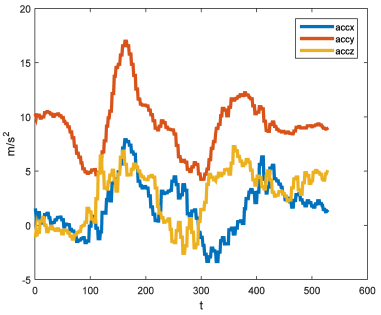
(a) **MGD database** (b) **SmartWatch Database**

Figure 3: Examples of hand gestures in **MGD database** and **SmartWatch Database**.

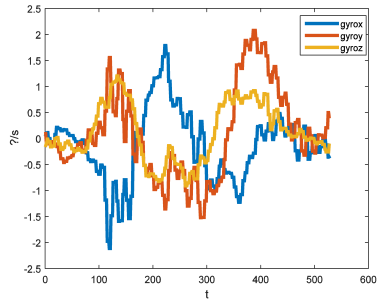
4.2. Implementation Details

We use Tensorflow toolbox as the deep learning platform, a Intel (R) Core (TM) i5-6500@3.20GHz, and an NVIDIA GTX 1070 GPU to perform the experiments. In order to validate the effectiveness of our proposed Fisher criterion in LSTM for modeling temporal sequences, we compare our methods, F-BiLSTM and F-BiGRU, with the state-of-the-art baselines (BiLSTM and BiGRU [58]) on three benchmarks including our collected database (MGD), and two previous databases: the BUAA Mobile Gesture database [59] and the SmartWatch Gestures database [58]. Some examples of hand gestures are shown in Fig. 3(a) and Fig. 3(b). We comprehensively evaluate the performance of the proposed models under different parameter settings of δ, α and θ in Sec. 4.3, and provide extensive experimental comparison results in Sec. 4.6.

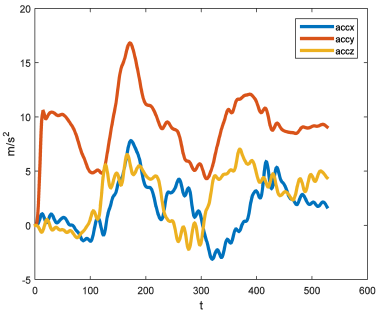
Data preprocessing. The main objective for data preprocessing is to facilitate gesture recognition. In real-world applications, the sensor data often contain a lot of noise due to complex environmental conditions and hardware limitations. Therefore, we first apply a filtering process to suppress noise (*i.e.*, data smoothing) by using Average Filter, Median Filter, and Butterworth Filter. Through experiment comparison, we select the Average Filter in terms of its good performance and computational efficiency. Fig. 4 shows the original accelerometer and gyrometer signals and the processed signals using the Average Filter.



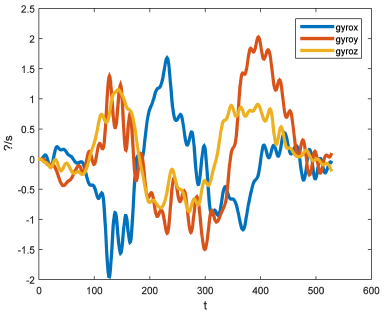
(a) Original accelerometer data



(b) Original gyrometer data

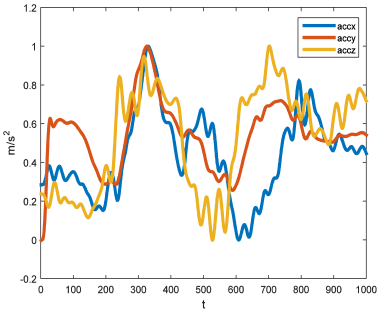


(c) Filtered accelerometer data

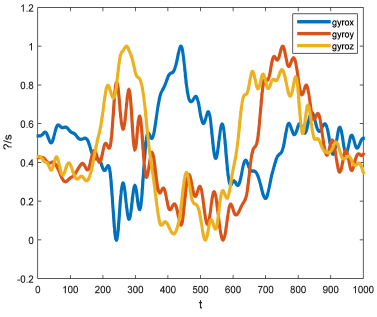


(d) Filtered gyrometer data

Figure 4: The original accelerometer and gyrometer data vs. the processed data by the Moving Average Filter.



(a) Preprocessed accelerometer data



(b) Preprocessed gyrometer data

Figure 5: The preprocessed accelerometer and gyrometer data.

The gesture execution speed of different participants may vary considerably, which leads to different signal lengths when using a fixed sampling frequency (200HZ) of accelerometer and gyrometer in the mobile phone. For example, gestures captured with fast motion may have fewer sampling points. Also, the signal strength of gesture sequences may vary. To cope with signal strength and speed variations, we apply amplitude and sequence normalization to the original signal sequences. Specifically, we first normalize a signal $x_i^n(t)$ by

$$x_i^n(t) = \frac{x_i(t) - \min_{t=1}^{T} x_i(t)}{\max_{t=1}^{T} x_i(t) - \min_{t=1}^{T} x_i(t)}, \quad \forall i \in \{1, ..., 6\}. \tag{9}$$

Then, we use cubic spline interpolation to normalize the length of a sequence to a fixed size (we set this size as 1000 in our experiments). Fig. 5 shows the preprocessed accelerometer and gyrometer data, where the sequence is filtered and normalized.

4.3. Parameters Evaluation

There are several parameters affecting the performance of gesture recognition, *i.e.*, the parameter α is restricted in [0,1] to control the update rate of mean μ, the parameter θ is bounded in [0,1] to balance the Fisher criterion and softmax in Eq. (4), and the parameter δ is restricted in a more subtle interval [1e-5,0.1] to balance the intra-class distance and inter-class distance in the Fisher criterion. The model BiLSTM with only the softmax loss can be considered as a special case of F-BiLSTM when θ is set to 0 in the loss function Eq. (4). In the following experiments, we pick up values in each interval to obtain an optimized parameter configuration for the best performance according to [35, 36]. We conduct experiments on the MGD dataset based on F-BiLSTM. Three parameters are used together in our F-BiLSTM model. For simplicity, we iteratively keep any two parameters with fixed values and test the third one for the optimal parameter setting.

Experiment 1. We fix α to 0.5, δ to 0.01 and vary θ from 0 to 1 to investigate the effect of θ. Fig. 6(a) shows the classification accuracy on

the testing set. The result shows that the model trained with only softmax loss has sub-optimal performance.

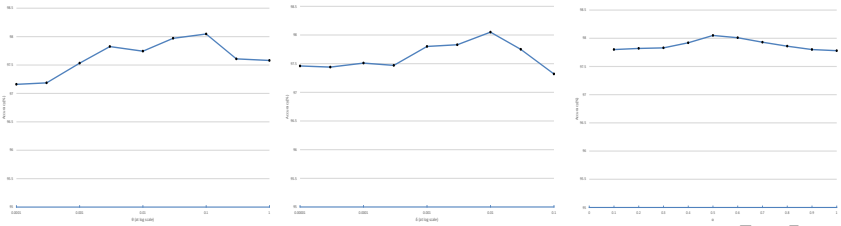
Experiment 2. We fix α to 0.5, θ to 0.1 and vary δ from 1e-5 to 0.1 to verify that the term of inter-class distances can promote the classification performance. As shown in Fig. 6(b), δ balances the intra-class distance and inter-class distance in the Fisher criterion.

Experiment 3. We fix θ to 0.1, δ to 0.01 and vary α from 0 to 1 to test the performance of our method. The results are illustrated in Fig. 6(c). We find that the performance of our model remains relatively stable across a wide range of α, but a moderate value of $\alpha = 0.5$ has the best performance.

4.4. Analysis of Model Effect

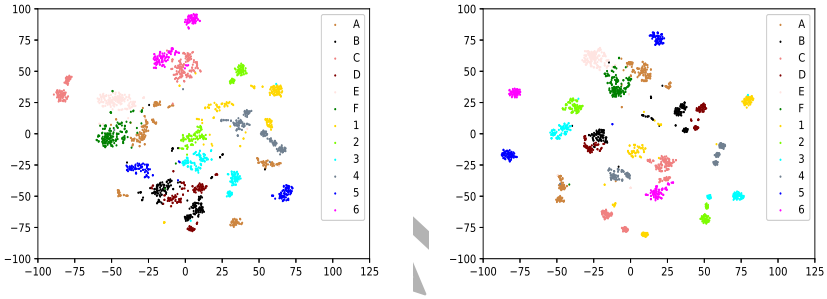
In the parameter tuning experiment, we show that F-BiLSTM and F-BiGRU have better discriminative ability than the baseline BiLSTM and BiGRU. In this section, we further discuss how a better feature distribution is achieved. We set θ to 0.1, δ to 0.01 and α to 0.5 for the F-BiLSTM model, and set the parameters to 0.3, 0.01, 0.5 for the F-BiGRU model, respectively.

Fig. 7 shows the feature visualizations of the MGD database. In Fig. 7(a) and Fig. 7(b), the BiLSTM and BiGRU features of 12 classes are visualized by the supervised t-SNE [60], while the F-BiLSTM and F-BiGRU features are illustrated in Fig. 7(c) and Fig. 7(d), respectively. The supervised t-SNE method plots the 2-dimensional features calculated based on the 128-dimensional features of BiLSTM, BiGRU, F-BiLSTM, and F-BiGRU, given the ground truth labels as shown in Fig. 7, From this figure, more compactness represents better deeply learned features, *i.e.*, minimizing the intra-class variations and maximizing the inter-class variations. Clearly, the distribution of F-BiLSTM and F-BiGRU features are more discriminative than the baseline BiLSTM and BiGRU features. Especially the F-BiGRU features in Fig. 7(d) are better than

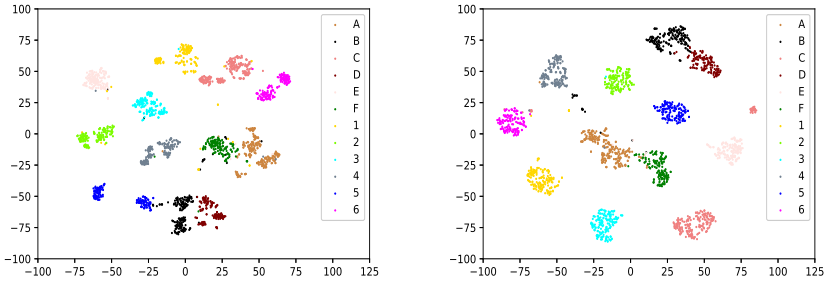


(a) parameter θ (b) parameter δ (c) parameter α

Figure 6: Influence of parameters θ, δ, and α on recognition accuracy.



(a) BiLSTM (b) BiGRU



(c) F-BiLSTM (d) F-BiGRU

Figure 7: Feature visualization of 12 classes of the MGD database. Different colors mean different classes.

Table 2: Comparison of computational time (unit: second) of different methods on **MGD database**.

Method \ Database	**MGD database**
HMM	575.77
RNN	1390.48
LSTM	458.95
GRU	443.39
BiLSTM	1009.84
BiGRU	929.68
F-BiLSTM (proposed)	**1009.96**
F-BiGRU (proposed)	**928.66**

the BiLSTM features in Fig. 7(a). As another verification, the quantitative evaluation is performed based on three databases in the next section.

4.5. Analysis of Computational Time

We implement HMM [50], RNN [43], LSTM [55], GRU [31], BiLSTM [33], and BiGRU [34] for comparison. We first compare the total computational time of these methods on the MGD database in Table 2. HMM tests on CPU with 575.77 seconds, while deep methods run on GPUs with a similar computation cost. LSTM and GRU are much faster than RNN, due to the improved unit with a high performance and low memory requirement as described in Section 1. We can also observe that both BiLSTM and BiGRU are nearly twice as expensive as LSTM and GRU in terms of the computational burden, because more neurons are used to denote the bidirectional memory. It is worth noting that the time cost of F-BiLSTM and F-BiGRU are similar to BiLSTM and BiGRU, which validate the efficiency of the proposed method.

4.6. Comparison with the State-of-the-arts

Experiment on MGD Database. For the proposed database, we select 3500 sequences to train our model and 2047 sequences for testing. After preprocessing, the length of each data sequence is set to 1000. Thus each input sample (3-axis accelerometer and gyrometer signals) is a matrix of 1000×6. Here, we train the network by using adaptive moment estimation, with the learning rate of 0.002 and the batch size of 200. For the F-BiLSTM model, we set θ to 0.1, δ to 0.01 and α to 0.5. We complete the training of BiLSTM and F-BiLSTM models with 1.5K iterations. The parameters of F-BiGRU model are set to 0.3, 0.01, 0.5 respectively. The training of BiGRU and F-BiGRU is completed with 1.2K iterations.

Table 3: Average accuracy(%) of BiLSTM, BiGRU and our proposed F-BiLSTM, F-BiGRU on **MGD database**.

Method / **Gesture**	BiLSTM	F-BiLSTM	BiGRU	F-BiGRU
A	97.41	97.85	97.09	98.09
B	94.17	96.50	97.24	98.78
C	98.95	99.40	99.85	100.00
D	96.88	99.04	98.02	98.87
E	96.88	97.40	98.48	98.61
F	96.86	98.59	97.62	99.54
1	93.80	95.33	96.62	98.53
2	98.60	98.82	99.03	99.35
3	96.69	97.56	98.29	99.42
4	98.77	98.97	99.28	99.29
5	96.55	98.16	99.77	100.00
6	99.10	99.32	99.77	99.61
Overall	97.05	**98.04**	98.38	**99.15**

Table 4: Comparison of overall accuracy(%) of different methods on **MGD database**.

Method \ Database	**MGD database**
HMM	91.11
RNN	94.22
LSTM	96.46
GRU	97.78
BiLSTM	97.05
BiGRU	98.38
F-BiLSTM (proposed)	**98.04**
F-BiGRU (proposed)	**99.15**

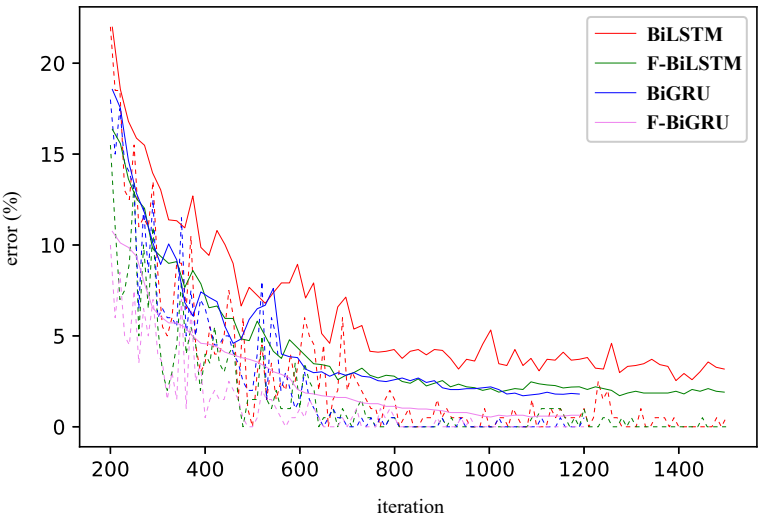


Figure 8: Training on **MGD database**. Dotted lines denote training errors, and solid lines denote testing errors.

Table 5: Average accuracy(%) of BiLSTM, BiGRU and our proposed F-BiLSTM, F-BiGRU on **BUAA mobile gesture database**.

Method / Gesture	BiLSTM	F-BiLSTM	BiGRU	F-BiGRU
A	100.00	99.17	98.34	99.58
B	97.29	98.92	97.84	98.37
C	100.00	100.00	100.00	100.00
D	99.26	97.42	96.77	99.35
1	97.87	99.57	100.00	100.00
2	100.00	100.00	100.00	100.00
3	97.06	100.00	100.00	100.00
4	95.83	97.50	97.08	97.08
Overall	98.44	**99.06**	98.75	**99.25**

In Table. 3, we report the classification accuracy of different methods on the testing set based on the average over 5 runs. It is clear that by incorporating the Fisher criterion to the baseline models (BiLSTM and BiGRU), the recognition performance can be improved. In Fig. 8, we analyze the training convergence for F-BiLSTM and F-BiGRU. Dotted lines denote training errors, while solid lines denote testing errors for different methods. As shown in this figure, F-BiLSTM and F-BiGRU converge faster, and gain better performance than BiLSTM and BiGRU. More specifically, F-BiLSTM converges more quickly (iteration #800 V.S. #1200) than BiLSTM and the error rates drop from 2.95% to 1.96%. F-BiGRU converges faster (iteration #1000 V.S. #1100) than BiGRU and the error rates drop from 1.62% to 0.85%. The results show that the introducing of Fisher criterion into the loss function can speed up the convergence and gain the lower error rates.

We also implement HMM [50], RNN [43], LSTM [55], GRU [31], BiLSTM [33], and BiGRU [34], which are compared with our work under the same experimental setting on the MGD database. As shown in Table 4, it demonstrates that

Table 6: Comparison of overall accuracy(%) of different methods on **BUAA mobile gesture database**.

Method \ Database	**BUAA mobile gesture database**
HMM	95.00
RNN	95.80
LSTM	96.23
GRU	97.29
BiLSTM	98.44
BiGRU	98.75
F-BiLSTM (proposed)	**99.06**
F-BiGRU (proposed)	**99.25**

our proposed Fisher criterion with either BiLSTM or BiGRU achieves better performance than RNN based methods (*e.g.* RNN, LSTM, GRU, BiLSTM, and BiGRU), and also enhance significantly compared to state-of-the-art classical machine learning methods (*e.g.* HMM).

Experiment on BUAA Mobile Gesture Database [59]. This database has 1120 samples for gestures A, B, C, D, 1, 2, 3, 4. Each sample includes 3-dimensional acceleration and angular velocity of the mobile phone. The training and testing sets are divided randomly into 70% and 30%, respectively. We conduct the experiments by using the same setting for F-BiLSTM and F-BiGRU as before. We set θ to 0.1, δ to 0.03 and α to 0.5. Model training is completed with 400 iterations. Table. 5 shows that LSTMs with Fisher criterion still have better results than baselines on a smaller dataset.

The models converge faster and yield lower classification error rates with the Fisher criterion as shown in Fig. 9. From this figure, F-BiLSTM converges more quickly (iteration #300 V.S. #340) than BiLSTM and the error rates drop from 1.56% to 0.94%. F-BiGRU converges faster (iteration #210 V.S. #220) than BiGRU, and the error rates drop from 1.25% to 0.75%. The results show that

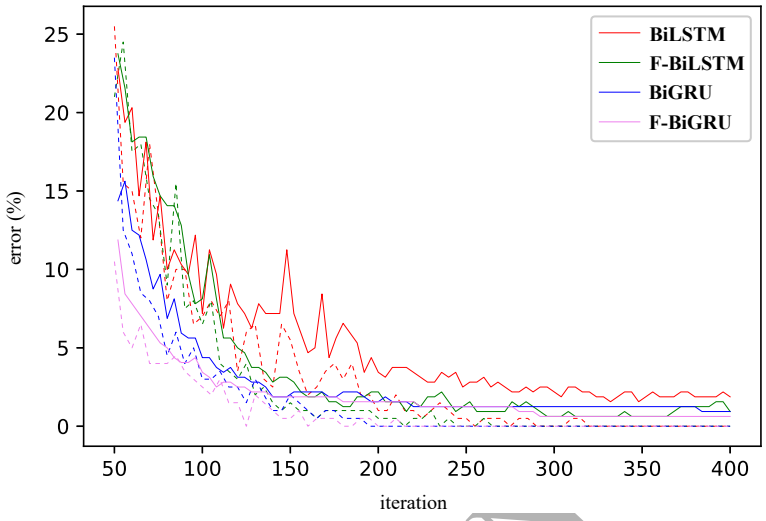


Figure 9: Training on **BUAA Mobile Gesture Database**. Dotted lines denote training errors, and solid lines denote testing errors.

the Fisher criterion can speed up the convergence and gain the lower error rate (*i.e.*, higher accuracy rate). We also compare the performance of our proposed framwork with the implemented HMM [50], RNN [43], LSTM [55], GRU [31], BiLSTM [33], and BiGRU [34] on BUAA mobile gesture database. In Table 6, the consistent improvements show that Fisher criterion can effectively improve the modeling ability of BiLSTM and BiGRU.

Experiment on SmartWatch Gesture Database [56]. In this database, eight different users perform twenty repetitions of twenty different gestures for a total of 3200 sequences as shown in Fig. 3(b). Different from the 6-dimensional sequences of the previous two databases, each sequence in this dataset only contains acceleration data from the 3-axis accelerometer of the first generation Sony SmartWatch. Furthermore, due to the lower sampling frequency, we set the length of each gesture sequence preprocessed to 50. We randomly select 2400 sequences as the training set and the rest 800 sequences as the testing

Table 7: Average accuracy(%) of BiLSTM, BiGRU and our proposed F-BiLSTM, F-BiGRU on **SmartWatch gesture database**.

Method / **Gesture**	BiLSTM	F-BiLSTM	BiGRU	F-BiGRU
1	94.58	97.91	97.08	97.50
2	95.00	97.22	95.56	95.56
3	86.90	87.59	93.10	93.10
4	95.91	97.27	97.27	97.73
5	96.88	98.13	96.88	98.13
6	93.33	94.07	96.30	100.00
7	96.44	96.89	98.22	99.56
8	97.62	98.57	100.00	100.00
9	93.49	96.74	96.74	97.67
10	94.84	98.06	100.00	100.00
11	89.76	94.15	94.15	95.12
12	92.89	92.44	96.00	97.33
13	90.42	95.00	94.17	95.42
14	94.88	96.30	96.30	97.21
15	95.14	95.14	100.00	97.84
16	92.20	89.27	93.17	93.17
17	96.52	95.65	99.13	100.00
18	96.22	97.30	96.76	95.68
19	94.29	94.76	94.76	96.67
20	97.21	98.60	100.00	100.00
Overall	94.30	**95.65**	96.80	**97.40**

Table 8: Comparison of overall accuracy(%) of different methods on **SmartWatch gesture database**.

Method \ Database	**SmartWatch gesture database**
HMM	82.50
RNN	89.98
LSTM	93.80
GRU	96.62
BiLSTM	94.30
BiGRU	96.80
F-BiLSTM (proposed)	**95.65**
F-BiGRU (proposed)	**97.40**

set. The parameters of Fisher criterion adopt the same setting in the previous experiment. Adaptive moment estimation is used to train the network, and the initial learning rate λ is set to 0.0001. The batch size is 1000. Training for BiLSTM and F-BiLSTM is terminated after 1.4K iterations and BiGRU and F-BiGRU with 2K iterations.

Fig. 10 shows the training and validation errors. Similar to Fig. 8 and Fig. 9, dotted lines denote training errors, and solid lines denote testing errors. In Fig. 10, F-BiLSTM converges more quickly (iteration #510 V.S. #750) than BiLSTM and the error rates drop from 5.70% to 4.35%. F-BiGRU converges faster (iteration #1300 V.S. #1500) than BiGRU, and the error rates decline from 3.20% to 2.60%. The results validate the convergence effect of Fisher criterion again. Table. 7 lists the classification results for different gestures. Notice that our proposed models perform considerably better than the baselines across the 20 gestures.

Based on the experimental evaluations in Table 8, we can observe that F-BiLSTM and F-BiGRU consistently gain improvements on SmartWatch gesture database, because we incorporate the Fisher criterion with softmax in the

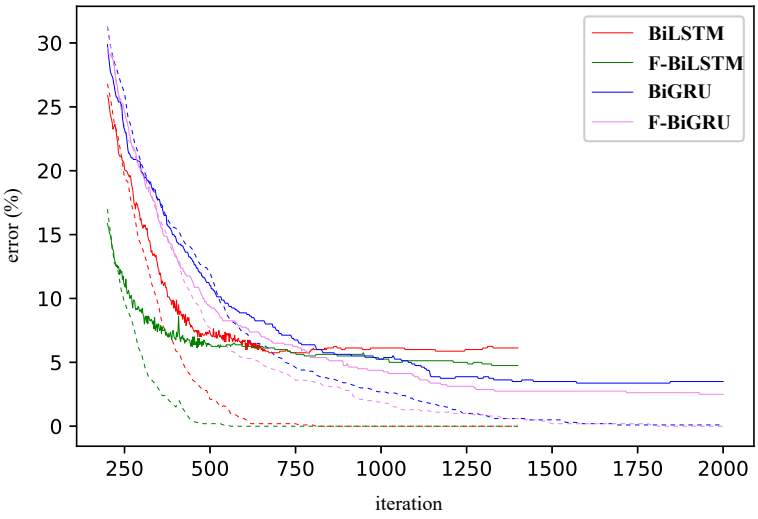


Figure 10: Training on **SmartWatch Gesture Database**. Dotted lines denote training errors, and solid lines denote testing errors.

loss function. Furthermore, with even small size training data, the proposed Fisher criterion improves the performance of BiLSTM and BiGRU models. The improvement comes from that the Fisher discriminant criterion can jointly minimize the intra-class variations and maximize the inter-class variations.

5. Conclusion

In this paper, we build a large gesture database, namely MGD, for hand gesture recognition based on mobile devices. We incoporate Fisher criterion into the BiLSTM and BiGRU networks termed as Fisher discriminant learned BiLSTM (F-BiLSTM) and Fisher discriminant learned BiGRU (F-BiGRU) to improve the mobile gesture recognition performance. With appropriate values assigned for the Fisher criterion parameters, the proposed methods achieve the state-of-the-art performance compared to existing RNN based methods and classical machine learning methods. In the future work, we will also apply our

framework to other tasks [61, 62] with sequential data.

Acknowledgement

The work was supported by the Natural Science Foundation of China under Contract 61601466, 61672079, 61473086. This work is supported by the Open Projects Program of National Laboratory of Pattern Recognition, and Supported by Shenzhen Peacock Plan.

References

[1] C. Hong, J. Yu, J. Wan, D. Tao, M. Wang, Multimodal deep autoencoder for human pose recovery, IEEE Transactions on Image Processing 24 (12) (2015) 5659–5670.

[2] C. Hong, J. Yu, D. Tao, M. Wang, Image-based three-dimensional human pose recovery by multiview locality-sensitive sparse retrieval, IEEE Transactions on Industrial Electronics 62 (6) (2015) 3742–3751.

[3] E. P. Ijjina, K. M. Chalavadi, Human action recognition in RGB-D videos using motion sequence information and deep learning, Pattern Recognition 72 (2017) 504–516.

[4] K. Altun, B. Barshan, O. Tuncel, Comparative study on classifying human activities with miniature inertial and magnetic sensors, Pattern Recognition 43 (10) (2010) 3605–3620.

[5] K. Liu, C. Chen, R. Jafari, N. Kehtarnavaz, Fusion of inertial and depth sensor data for robust hand gesture recognition, IEEE Sensors Journal 14 (6) (2014) 1898–1903.

[6] T. T. Ngo, Y. Makihara, H. Nagahara, Y. Mukaigawa, Y. Yagi, Similar gait action recognition using an inertial sensor, Pattern Recognition 48 (4) (2015) 1289–1301.

[7] M. Patacchiola, A. Cangelosi, Head pose estimation in the wild using convolutional neural networks and adaptive gradient methods, Pattern Recognition 71 (2017) 132–143.

[8] N. D. Lane, E. Miluzzo, H. Lu, D. D. Peebles, T. Choudhury, A. T. Campbell, A survey of mobile phone sensing, IEEE Communications Magazine 48 (9) (2010) 140–150. doi:10.1109/MCOM.2010.5560598.

[9] E. Choi, W. Bang, S. Cho, J. Yang, D. Kim, S. Kim, Beatbox music phone: gesture-based interactive mobile phone using a tri-axis accelerometer, IEEE International Conference on Industrial Technology (2005) 97–102doi:10.1109/ICIT.2005.1600617.

[10] V. Mantyla, J. Mantyjarvi, T. Seppanen, E. Tuulari, Hand gesture recognition of a mobile device user, International Conference on Multimedia and Expo 1 (2000) 281–284. doi:10.1109/ICME.2000.869596.

[11] J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, V. Vasudevan, uwave: Accelerometer-based personalized gesture recognition and its applications, ieee international conference on pervasive computing and communications 5 (6) (2009) 1–9. doi:10.1109/PERCOM.2009.4912759.

[12] B. Zhang, Y. Yang, C. Chen, L. Yang, J. Han, L. Shao, Action recognition using 3d histograms of texture and a multi-class boosting classifier, IEEE Transactions on Image Processing 26 (10) (2017) 4648–4660.

[13] C. Catal, S. Tufekci, E. Pirmit, G. Kocabag, On the use of ensemble of classifiers for accelerometer-based activity recognition, Applied Soft Computing 37 (2015) 1018–1022. doi:10.1016/j.asoc.2015.01.025.

[14] H. Junker, O. Amft, P. Lukowicz, G. Tr̈oster, Gesture spotting with body-worn inertial sensors to detect user activities, Pattern Recognition 41 (2008) 2010–2014.

[15] A. Akl, S. Valaee, Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing, International

Conference on Acoustics, Speech, and Signal Processing (2010) 2270–2273doi:10.1109/ICASSP.2010.5495895.

[16] M. Sundermeyer, R. Schluter, H. Ney, LSTM neural networks for language modeling, Conference of the International speech Communication Association.

[17] G. Mesnil, X. He, L. Deng, Y. Bengio, Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding, Conference of the International Speech Communication Association.

[18] O. Vinyals, A. Toshev, S. Bengio, D. Erhan, Show and tell: A neural image caption generator, CVPR (2015) 3156–3164doi:10.1109/CVPR.2015.7298935.

[19] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhutdinov, R. Zemel, Y. Bengio, Show, attend and tell: Neural image caption generation with visual attention, Computer Science (2015) 2048–2057.

[20] L. Yang, C. Li, J. Han, C. Chen, Q. Ye, B. Zhang, Image reconstruction via manifold constrained convolutional sparse coding for image sets, Journal of Selected Topics Signal Processing 11 (7) (2017) 1072–1081.

[21] J. Y. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, G. Toderici, Beyond short snippets: deep networks for video classification, CVPR (2015) 4694–4702doi:10.1109/CVPR.2015.7299101.

[22] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, S. Savarese, Social LSTM: Human trajectory prediction in crowded spaces, 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016) 961–971doi:10.1109/CVPR.2016.110.

[23] J. Yu, C. Hong, Y. Rui, D. Tao, Multi-task autoencoder model for recovering human poses, IEEE Transactions on Industrial Electronics PP (99) (2017) 1–1.

[24] Y. Du, W. Wang, L. Wang, Hierarchical recurrent neural network for skeleton based action recognition, CVPR (2015) 1110–1118doi:10.1109/CVPR.2015.7298714.

[25] V. Veeriah, N. Zhuang, G. Qi, Differential recurrent neural networks for action recognition, ICCV (2015) 4041–4049doi:10.1109/ICCV.2015.460.

[26] J. Wang, Z. Liu, Y. Wu, J. Yuan, Learning actionlet ensemble for 3d human action recognition, IEEE Transactions on Pattern Analysis and Machine Intelligence 36 (5) (2014) 914–927. doi:10.1109/TPAMI.2013.198.

[27] J. Liu, A. Shahroudy, D. Xu, G. Wang, Spatio-temporal LSTM with trust gates for 3d human action recognition, ECCV (2016) 816–833doi:10.1007/978-3-319-46487-9_50.

[28] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, X. Xie, Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks, in: Association for the Advancement of Artificial Intelligence, 2016, pp. 3697–3704.

[29] L. G. Hafemann, R. Sabourin, L. S. Oliveira, Learning features for offline handwritten signature verification using deep convolutional neural networks, Pattern Recognition 70 (2017) 163–176.

[30] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural Computation 9 (8) (1997) 1735–1780. doi:10.1162/neco.1997.9.8.1735.

[31] K. Cho, B. van Merriënboer, Ç. Gülçehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder–decoder for statistical machine translation, in: Empirical Methods in Natural Language Processing, 2014, pp. 1724–1734.

[32] M. Schuster, K. K. Paliwal, Bidirectional recurrent neural networks, IEEE Transactions on Signal Processing 45 (11) (1997) 2673–2681.

[33] E. Kiperwasser, Y. Goldberg, Simple and accurate dependency parsing using bidirectional LSTM feature representations, Transactions of the Association for Computational Linguistics 4 (0) (2016) 313–327.

[34] D. Bahdanau, K. Cho, Y. Bengio, Neural machine translation by jointly learning to align and translate, in: International Conference on Learning Representations, 2015.

[35] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, European Conference on Computer Vision (2016) 499–515.

[36] W. Liu, Y. Wen, Z. Yu, M. Yang, Large-margin softmax loss for convolutional neural networks, International Conference on Machine Learning (2016) 507–516.

[37] J. G. Rekimoto, Gesturepad, Unobtrusive wearable interaction devices, Fifth International Symposium on Wearable Computers (2001) 21–27doi:10.1109/ISWC.2001.962092.

[38] I. J. Jang, W. Park, Signal processing of the accelerometer for gesture awareness on handheld devices, Robot and Human Interactive Communication (2003) 139–144doi:10.1109/ROMAN.2003.1251823.

[39] S. Kallio, J. Kela, J. Mantyjarvi, Online gesture recognition system for mobile interaction, Systems, Man and Cybernetics 3 (2003) 2070–2076. doi:10.1109/ICSMC.2003.1244189.

[40] A. Bulling, U. Blanke, B. Schiele, A tutorial on human activity recognition using body-worn inertial sensors, ACM Computing Surveys 46 (3) (2014) 33. doi:10.1145/2499621.

[41] X. Zhang, X. Chen, W. H. Wang, J. H. Yang, V. Lantz, K. Q. Wang, Hand gesture recognition and virtual game control based on 3d accelerometer and EMG sensors, 2009, pp. 401–406. doi:10.1145/1502650.1502708.

[42] H. Gjoreski, J. Bizjak, M. Gjoreski, M. Gams, Comparing deep and classical machine learning methods for human activity recognition using wrist accelerometer, in: Proceedings of the 25th International Joint Conference on Artificial Intelligence, New York, 2016, pp. 1–7.

[43] A. Tang, K. Lu, Y. Wang, J. Huang, H. Li, A real-time hand posture recognition system using deep neural networks, ACM Transactions on Intelligent Systems and Technology (TIST) 6 (2) (2015) 21.

[44] S. Agrawal, I. Constandache, S. Gaonkar, R. R. Choudhury, K. Caves, F. Deruyter, Using mobile phones to write in air, in: International Conference on Mobile Systems, Applications, and Services, 2011, pp. 15–28. doi:10.1145/1999995.1999998.

[45] G. Lefebvre, S. Berlemont, F. Mamalet, C. Garcia, BLSTM-RNN based 3D gesture classificationdoi:10.1007/978-3-642-40728-4_48.

[46] F. G. Hofmann, P. Heyer, G. Hommel, Velocity profile based recognition of dynamic gestures with discrete hidden markov models, Lecture Notes in Computer Science (1998) 81–95doi:10.1007/BFb0052991.

[47] J. Kela, P. Korpipaa, J. Mantyjarvi, S. Kallio, G. Savino, L. Jozzo, D. Marca, Accelerometer-based gesture control for a design environment, Personal and Ubiquitous Computing 10 (5) (2006) 285–299. doi:10.1007/s00779-005-0033-8.

[48] T. Pylvanainen, Accelerometer based gesture recognition using continuous hmms, iberian conference on pattern recognition and image analysis (2005) 639–646doi:10.1007/11492429_77.

[49] D. Mace, W. Gao, A. Coskun, Accelerometer-based hand gesture recognition using feature weighted naïve bayesian classifiers and dynamic time warping, in: Proceedings of the Companion Publication of the 2013 International Conference on Intelligent User Interfaces Companion, 2013, pp. 83–84. doi:10.1145/2451176.2451211.

[50] J. Wu, G. Pan, D. Zhang, G. Qi, S. Li, Gesture recognition with a 3-d accelerometer, Ubiquitous Intelligence and Computing (2009) 25–38doi:10.1007/978-3-642-02830-4_4.

[51] W.-H. Hsu, Y.-Y. Chiang, W.-Y. Lin, W.-C. Tai, J.-S. Wu, Integrating LCS and SVM for 3d handwriting recognition on handheld devices using accelerometers, in: Proceedings of the 3rd International Conference on Communications and Information Technology, 2009, pp. 195–197.

[52] V. P. Tea Marasovic, Accelerometer-based gesture classification using principal component analysis, in: SoftCOM 2011, 19th International Conference on Software, Telecommunications and Computer Networks, 2011, pp. 1 – 5.

[53] Z. He, Accelerometer based gesture recognition using fusion features and SVM, JSW 6 (2011) 1042–1049. doi:10.4304/jsw.6.6.1042-1049.

[54] S. Shin, W. Sung, Dynamic hand gesture recognition for wearable devices with low complexity recurrent neural networks, International Symposium on Circuits and Systems (2016) 2274–2277doi:10.1109/ISCAS.2016.7539037.

[55] F. J. Ordonez, D. Roggen, Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition, Sensors 16 (1) (2016) 115.

[56] G. Costante, L. Porzi, O. Lanz, P. Valigi, E. Ricci, Personalizing a smartwatch-based gesture interface with transfer learning, 2014 22nd European Signal Processing Conference (EUSIPCO) (2014) 2530–2534.

[57] D. B. K. Cho, B. van Merrienboer, Y. Bengio, On the properties of neural machine translation: Encoder-decoder approaches, arXiv preprint 1409 (1259).

[58] J. Chung, C. Gulcehre, K. Cho, Y. Bengio, Empirical evaluation of gated recurrent neural networks on sequence modeling, Eprint arXiv.

[59] C. Xie, S. Luan, H. Wang, B. Zhang, Gesture recognition benchmark based on mobile phone, CCBRdoi:10.1007/978-3-319-46654-5_48.

[60] L. V. D. Maaten, E. o. Postma, H. J. V. D. Herik, Dimensionality reduction: A comparative review, IEEE Transactions on Pattern Analysis and Machine Intelligence 10.

[61] B. Zhang, Z. Li, X. Cao, Q. Ye, C. Chen, L. Shen, A. Perina, R. Ji, Output constraint transfer for kernelized correlation filter in tracking, IEEE Transactions on Systems, Man, and Cybernetics: Systems 47 (4) (2017) 693–703.

[62] B. Zhang, Z. Li, A. Perina, A. Del Bue, V. Murino, J. Liu, Adaptive local movement modeling for robust object tracking, IEEE Transactions on Circuits Systems for Video Technology 27 (7) (2017) 1515–1526.

Biography

Ce Li. received the B.E. degree in Computer Science from Tianjin University, Tianjin, China, in 2008, the M.S. and Ph.D. degrees in Computer Science from the School of Electronic, Electrical and Communication Engineering at the University of Chinese Academy of Sciences, Beijing, China, in 2012 and 2015, respectively. She is currently a research assistant with China University of Mining & Technology, Beijing, China. Her current interests include computer vision, video analysis, and machine learning. She was supported by the Natural Science Foundation of China for Youth.

Chunyu Xie. received the B.S. degree and is a master in automation from Beihang University. His current research interests include signal and image processing, pattern recognition and computer vision.

Baochang Zhang. received the B.S., M.S. and Ph.D. degrees in Computer Science from Harbin Institue of the Technology, Harbin, China, in 1999, 2001,

and 2006, respectively. From 2006 to 2008, he was a research fellow with the Chinese University of Hong Kong, Hong Kong, and with Griffith University, Brisban, Australia. Currently, he is an associate professor with the Science and Technology on Aircraft Control Laboratory, School of Automation Science and Electrical Engineering, Beihang University, Beijing, China. He was supported by the Program for New Century Excellent Talents in University of Ministry of Education of China. His current research interests include pattern recognition, machine learning, face recognition, and wavelets.

Chen Chen. received the B.E. degree in automation from Beijing Forestry University, Beijing, China, in 2009, the M.S. degree in electrical engineering from Mississippi State University, Starkville, MS, USA, in 2012, and the Ph.D. degree from the University of Texas at Dallas, Richardson, TX, USA, in 2016. He is currently a Postdoctoral Fellow with the Center for Research in Computer Vision, University of Central Florida, Orlando, FL, USA. His current research interests include compressed sensing, signal and image processing, pattern recognition, and computer vision. He has published over 40 papers in refereed journals and conferences in the above areas.

Jungong Han. is currently a Senior Lecturer with the Department of Computer Science and Digital Technologies at Northumbria University, Newcastle, UK. Previously, he was a Senior Scientist (2012-2015) with Civolution Technology (a combining synergy of Philips Content Identification and Thomson STS), a Research Staff (2010-2012) with the Centre for Mathematics and Computer Science (CWI), and a Senior Researcher (2005-2010) with the Technical University of Eindhoven (TU/e) in Netherlands. Dr. Hans research interests include Multimedia Content Identification, Multi-Sensor Data Fusion, Computer Vision and Multimedia Security. He is an Associate Editor of Elsevier Neurocomputing (IF 2.4) and an Editorial Board Member of Springer Multimedia Tools and Applications (IF 1.4). He has been (lead) Guest Editor for five international journals, such as IEEE-T-SMCB, IEEE-T-NNLS. Dr. Han is the recipient of the UK Mobility Award Grant from the UK Royal Society in 2016.