Mobile Gesture Recognition using Hierarchical Recurrent Neural Network with Bidirectional Long Short-Term Memory

Myeong-Chun Lee
Dept. of Computer Science
Yonsei University
Seoul, Korea
lmspring@sclab.yonsei.ac.kr

Sung-Bae Cho
Dept. of Computer Science
Yonsei University
Seoul, Korea
sbcho@cs.yonsei.ac.kr

***Abstract*—As the sensors embedded to a smartphone are proliferating, many application systems for context-aware services are actively investigated. This paper proposes a gesture recognition system with smartphones for better interface. It is important to maintain high accuracy even with the large number of gestures. To improve the accuracy, we adopt the recurrent neural network based on hierarchical BLSTM (Bidirectional Long Short-Term Memory). The first level BLSTMs are used to discriminate the gestures and non-gestures, and the second level BLSTMs classify the input into one of twenty gestures. Experiments with 24,850 sequence data consisting of 11,885 gesture sequences and 12,965 non-gesture sequences confirm the high performance of the proposed method over the competitive alternatives.**

Keywords-mobile interface; gesture recognition; hierarchical neural network; bidirectional recurrent neural network; long short-term memory

I. INTRODUCTION

A variety of sensors such as accelerometer, ambient light, proximity, dual cameras, GPS, dual microphones, compass, and gyroscope are embedded to a smartphone. They are not only sophisticated, but also show good performance [1]. Especially, accelerometer is one of the most commonly used sensors for the physical movements of the user carrying the phone. For this reason, several user interfaces with gesture and activity recognitions have been developed by using the accelerometer [2].

However, there are two crucial problems to develop user recognition systems with the smartphone sensors. One is to identify the non-gesture or non-activity data. The gesture or activity data includes meaningful and non-meaningful parts. Sometimes the amount of non-meaningful data can be even more than gesture data. In this case, it is time-consuming to recognize both meaningful and non-meaningful data. The other is to maintain high-accuracy even with the large number of classes. Many pattern recognition systems depend on machine learning methods to learn the complex patterns. However, the performance degrades when classifying a large number of classes. It is an important problem when providing the various services to users.

In this paper, we propose a mobile gesture recognition system where data is collected from the accelerometer sensors embedded to a smartphone. Figure 1 shows example of a gesture data set. To alleviate the aforementioned problems, a recurrent neural network based on BLSTM (Bidirectional Long Short-Term Memory) is used to classify twenty classes. To discriminate the gestures and non-gestures, hierarchical structure is exploited. In experiments, the proposed method outperforms the standard BLSTM.

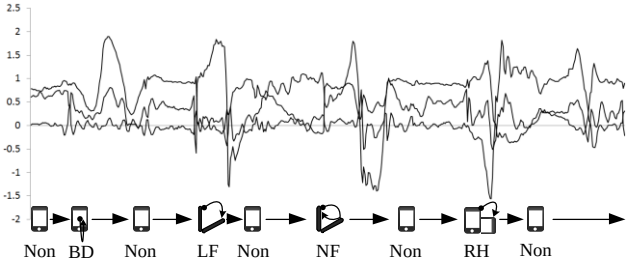


Figure 1. Example of a gesture data set

The paper is organized as follows. Section II presents the related works for gesture recognition. Section III describes the overall architecture, BRNN (Bidirectional Recurrent Neural Network), and LSTM (Long Short-Term Memory) respectively. Section IV shows the experimental results and we conclude with some remarks in Section V.

II. RELATED WORKS

The related works using accelerometer is shown in Table I.

TABLE I. GESTURE RECOGNITION WORKS USING ACCELEROMETER

Author	Classifier	No. of gestures	Collector	Overview
J. Liu *et al.* (2009) [3]	DTW	8	Wiimote	Personalized gesture using uWave
G. Niezen *et al.* (2009) [4]	HMM, ANN, DTW	8	Smartphone	comparison of classifier algorithms
J.–K. Min *et al.* (2010) [5]	DTW, NB, K-means	20	Smartphone	DTW model selection through NB
T. Marasovic *et al.* (2011) [6]	K-NN	7	Smartphone	Combination of the PCA and K-NN
A. Akl *et al.* (2011) [7]	DTW, AP	18	Wiimote	Dimensional reduction through RP

The success of the gesture recognition is subject to which classification method is used, how many gestures are used, and what kind of collector is used for collecting the data. As can be seen in Table I, most of the research use static algorithms such as MLP (Multi-Layer Perceptron), k-means clustering and combination of DTW (Dynamic Time Warping) and other methods. However it is more suitable to

 ISBN: 978-1-61208-236-3

use the dynamic classification algorithms directly for the time-series data. For this reason, we adopt a dynamic classification algorithm which is a kind of recurrent neural network. BLSTM shows better performance than other algorithms for recognizing the time-series patterns in several domains. F. Eyben *et al.* proposed an audiovisual approach and LSTM for recognizing conversational speech. From the experiments, they showed that the LSTM outperformed SVM (Support Vector Machine) [8]. T. Thireou *et al.* applied BLSTM to the sequence-based prediction of protein localization, and showed that the proposed method is better than FFNN (Feed Forward Neural Network) and BRNN[9].

III. ARCHITECTURE AND METHOD

A. The overall system architecture

This paper aims at enhancing the accuracy by using hierarchical structure. The entire system configuration is shown in Figure 2. The accelerometer data collected from a smartphone are segmented by using sliding window and average variation. The preprocessed data is hierarchically classified after training. We adopt the recurrent neural network based on BLSTM[10], which is a hybridization of BRNN[11] and LSTM[12]. First, training data are used to classify the gestures and non-gestures. Second, classified gesture data are used for classifying the twenty gesture classes.

B. Bidirectional recurrent neural network

The basic idea of BRNN is to present each training sequence forwards and backwards to the two separate recurrent hidden layers, both of which are connected to the same output layer. This provides the network with past and future context for every point in the input sequence. Figure 3 shows a structure of unfolded bidirectional recurrent neural network over two time steps. BRNN shows better performance than other approaches such as regression and classification experiments [11].

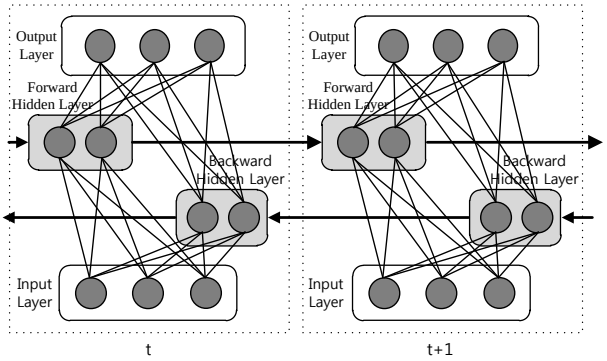


Figure 3. The structure of bidirectional recurrent neural network

C. Long Short-Term Memory

LSTM is an extension of the recurrent neural network. It uses the three gates that can store and access the data collected from rest of the network. Gates are activated from logistic sigmoid activation function. Figure 4 shows a memory block of LSTM. The Hyperbolic tangent activation function is used for squashing functions. The basic calculation of each gate is the same as the standard artificial neural network [12].

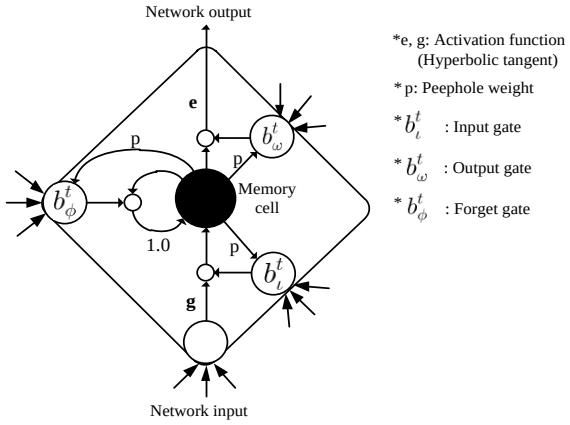


Figure 4. A LSTM memory cell

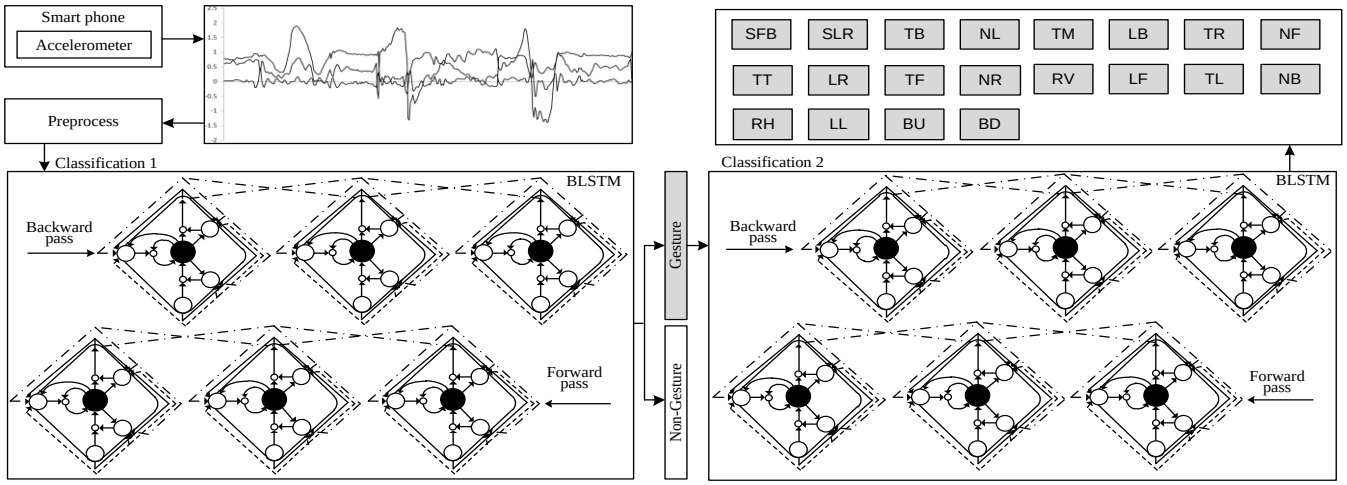


Figure 2. The overall system architecture

The input gate determines whether the input values put the memory cell or not.

$$\alpha_\iota^t = \sum_{i=1}^{I} w_{i\iota} x_i^t + \sum_{h=1}^{H} w_{h\iota} \beta_h^{t-1} + \sum_{c=1}^{C} w_{c\iota} s_c^{t-1} \quad (1)$$
$$\beta_\iota^t = f(\alpha_\iota^t)$$

where, α_ι^t is a state of the input gate at time *t*. It is calculated from input values, the output of other networks, and state of the memory cell. *I*, *H*, and *C* mean the number of input node, hidden node, and cell, respectively. *w* is the weight of connected nodes. f is the logistic sigmoid function to activate the input gate.

The output gate determines whether the information is output or not. The calculation of the output gate is similar with the input gate.

$$\alpha_\Phi^t = \sum_{i=1}^{I} w_{i\Phi} x_i^t + \sum_{h=1}^{H} w_{h\Phi} \beta_h^{t-1} + \sum_{c=1}^{C} w_{c\Phi} s_c^{t-1} \quad (2)$$
$$\beta_\Phi^t = f(\alpha_\Phi^t)$$

where, α_Φ^t is a state of the forget gate at time *t* and β_Φ^t is a state after applying the activation function.

Equation (3) is a calculation that is generated by forget gate, the state of cell, the state of input gate, and state of α_c^t after applying hyperbolic tangent activation function. Note that the fixed weight value 1.0 is used for preserving the information in the memory cell.

$$\alpha_c^t = \sum_{i=1}^{I} w_{ic} x_i^t + \sum_{h=1}^{H} w_{hc} \beta_h^{t-1} \quad (3)$$
$$S_c^t = \beta_\Phi^t S_c^{t-1} + \beta_\iota^t g(\alpha_c^t)$$

Forget gate provides the information to reset the memory cell.

$$\alpha_\omega^t = \sum_{i=1}^{I} w_{i\omega} \alpha x_i^t + \sum_{h=1}^{H} w_{h\omega} \beta_h^{t-1} + \sum_{c=1}^{C} w_{c\omega} s_c^{t-1} \quad (4)$$
$$\beta_\omega^t = f(\alpha_\omega^t)$$

where, ω is the output gate and β_ω^t is the state of an output gate after applying the activation function at time *t*.

Equation (5) is a definition of the cell output. To activate the cells, hyperbolic tangent is used and multiply the state of output gate.

$$\beta_c^t = \beta_\omega^t e(s_c^t) \quad (5)$$

For training the LSTM recurrent neural network, we use the Back Propagation Through Time (BPTT).

IV. EXPERIMENT

A. Data preparation

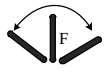
For the experiment, Samsung Omnia smartphone with MS Windows Mobile 6.1 was used as the platform. The acceleration data are sampled at 50Hz. 30 people of 10~60 years old participate in the experiment. The collected data is divided into generations and date. Total amount of the data consists of 11,885 gesture sequences and 12,965 non-gesture sequences. The number of files used in the experiment is 1075.

Table II shows the detailed description of twenty gestures. Rotating and tilting hold their physical states after the movement. Tapping represents the hand or finger stroke on a smartphone surface. In the case of shaking, subjects shake the devices two or more times in a specific direction. Snapping has an angular acceleration while bouncing moves straightly to a direction and reflected back where both are the kind of pendulum movement. We set learning rate and momentum as 0.0001 and 0.9 respectively. BPTT which is one of the dynamic learning algorithms is used for training the gesture data.

B. The results

For the first experiment, the data are divided into a ratio of seven to three as training and test data, respectively. 17,470 sequences are used for training, and 7,380 sequences are used for test. Each sequence is distributed randomly. The results through all experiments in this work are compared with the standard BLSTM. The accuracy rate for the first experiment is shown in Figure 5. The average accuracy of Hierarchical BLSTM is 91.15%, whereas the standard BLSTM is 89.20%. As can be seen in Figure 5, the hierarchical BLSTM generally outperforms the standard BLSTM.

TABLE II. DESCRIPTION OF THE GESTURE DATA

Symbol	**NL**	**NR**	**NF**	**NB**	**BU**	**BD**	**RH**	**RV**	**SLR**	**SLF**
Meaning	Snapping				Bouncing		Rotating		Shaking	
Direction	Left	Right	Forward	Backward	Up	Down	Horizontal	Vertical	Left-Right	F-Backward
Movement										
Symbol	**TL**	**TR**	**TF**	**TB**	**TT**	**TM**	**LL**	**LR**	**LF**	**LB**
Meaning	Tapping						Tilting			
Direction	Left	Right	Forward	Backward	Top	Bottom	Left	Right	Forward	Backward
Movement										

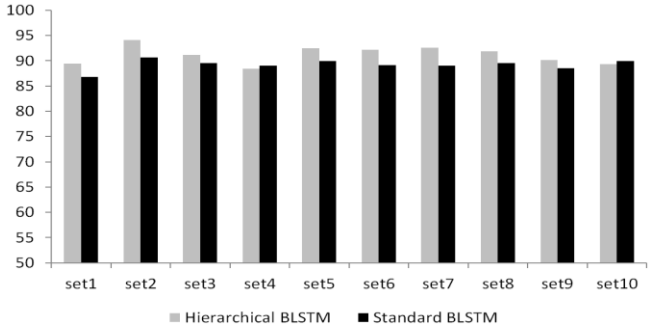


Figure 5. The Results for randomly distributed data

For the second experiment, we group the data as generations. 18,490 sequences are used for training and about 2,100 sequences are used for testing at the each generation. Table III also shows the hierarchical BLSTM outperforms the standard BLSTM in most cases.

TABLE III. GENERATION RESULTS

		Set1	Set2	Set3	Set4	Set5	Avg.
10~20	Hierarchical BLSTM	90.13	90.5	88.33	90.1	86.1	89.032
	Standard BLSTM	88.5	89.9	85.8	86.3	84.9	87.08
20~40	Hierarchical BLSTM	94.81	97.15	94.4	96.9	96.23	95.898
	Standard BLSTM	95.32	96.69	93.73	96.9	95.79	95.686
40~60	Hierarchical BLSTM	82.54	86.5	88.43	89.21	85.8	86.496
	Standard BLSTM	79.1	86.7	85.2	87.6	85.8	84.88

To get the fair comparison, we conducted ten-fold cross validation test. The raw data are randomly partitioned into ten subsamples. Of the ten subsamples, a single subsample is retained for the validation data for testing and remaining nine subsamples are used for training data. This process is repeated ten times. The results of the hierarchical BLSTM except the set3 and set8 are more accurate than the standard BLSTM.

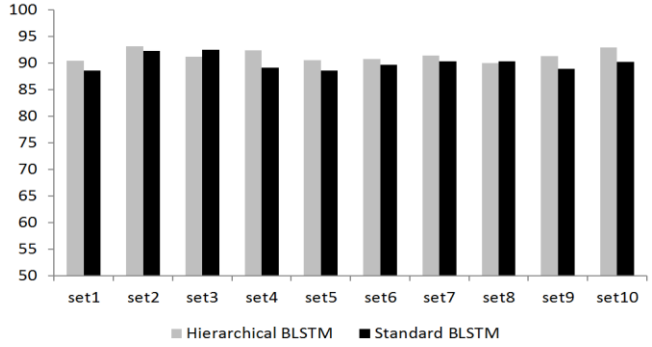


Figure 6. 10-fold cross validation result

V. CONCLUSION AND FUTURE WORK

In this paper, we collect the accelerometer data from a smartphone and classify the data by using hierarchical BLSTM. Since the gesture data contain a lot of non-gesture data, we classify the non-gesture sequences before classifying the meaningful sequences. More than 20,000 sequences were used for reliable experiment and total classes of the data were twenty one including non-gesture data. The performance of the standard BLSTM was compared with the hierarchical BLSTM and our approach outperformed the standard BLSTM. For the future work, it can be possible to achieve higher accuracy if the data are grouped with the similar meaning because some gestures have similar characteristics.

ACKNOWLEDGMENT

This research was supported by the Original Technology Research Program for Brain Science through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology (2010-0018948).

REFERENCES

[1] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles, T. Choudhury, and A. T. Cambell, "A survey of mobile phone sensing," *IEEE Communications Magazine*, vol. 48, no. 9 pp. 140-150, 2010.

[2] N. D. Lane, E. Miluzzo, H. Lu, D. Peebles. T. Choudhury, and A. T. Campbell"A. survey of mobile phone sensing," *Communications Magazine*, vol. 48, no. 9, pp. 140-150, 2010.

[3] J. Liu, Z. Wang, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uWave: accelerometer-based personalized gesture recognition and its applications," *IEEE Int. Conf. on Pervasive Computing and Communications*, pp. 1-9. 2009.

[4] G. Niezen and G. P. Hancke, "Evaluating and optimising accelerometer-based gesture recognition techniques for mobile devices," *AFRICON*, pp.1-6, 2009.

[5] J.-K. Min, B.-W. Choe, and S.-B. Cho, "A selective template matching algorithm for short and intuitive gesture UI of accelerometer-builtin mobile phones," *Conq. on Nature and Biologically Inspired Computing*, pp. 660-665, 2010.

[6] T. Marasovic and V. Papic, "Accelerometer-Based Gesture Classification Using Principal Component Analysis," *Int. Conf. on Software, Telecommunications and Computer Networks*, pp. 1-5, 2011.

[7] A. Akl, C. Feng, and S. Valaee, "A novel accelerometer-based gesture recognition system," *IEEE Trans. on Signal Processing*, vol. 59, no. 12, pp. 6197-6205, 2011.

[8] F. Eyben, S. Petridis, B. Schuller, G. Tzimiropilos, S. Zafeiriou, and M. Pantic. "Audiovisual classification of vocal outbursts in human conversation using long-short-term memory networks," *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, pp. 5844-5847, 2011.

[9] T. Thireou and M. Reczko, "Bidirectional long short-term memory networks for predicting the subcellular localization of eukaryotic proteins," *IEEE/ACM Trans. on Computational Biology and Bioformatics*, vol. 4, no. 3, pp. 441-446, 2007.

[10] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 5, pp. 855–868, 2009.

[11] M. Schuster, "Bidirectional recurrent neural network," *IEEE Trans. On Signal Processing*, vol. 45, no. 11, pp.2673-2681, 1997.

[12] S. Hochreiter and J. Schmidhuer, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735-1780, 1997.

 ISBN: 978-1-61208-236-3