Recognition of Human Hand Activities Based on a Single Wrist IMU Using Recurrent Neural Networks

Patricio Rivera[1], Edwin Valarezo[1, 2], Mun-Taek Choi[3], and Tae-Seong Kim[1]
[1] Dept. of Biomedical Engineering, Kyung Hee University, Republic of Korea
[2] Escuela Superior Politécnica del Litoral, ESPOL, Guayaquil, Ecuador
[3] School of Mechanical Engineering, Sungkyunkwan University, Republic of Korea
Email: {patoalejor, edgivala, tskim}@khu.ac.kr, mtchoi@skku.edu

***Abstract*—Recognition of hand activities could provide new information towards daily human activity logging and gesture interface applications. However, there is a technical challenge due to delicate hand motions and complex movement contexts. In this work, we proposed hand activity recognition (HAR) based on a single inertial measurement unit (IMU) sensor at one wrist via deep learning recurrent neural network. The proposed HAR works directly with signals from a tri-axial accelerometer, gyroscope, and magnetometer sensors within one IMU. We evaluated the performance of our HAR with a public human hand activity database for six hand activities including Open Door, Close Door, Open Fridge, Close Fridge, Clean Table and Drink from Cup. Our results show an overall recognition accuracy of 80.09% with discrete standard epochs and 74.92% with noise-added epochs. With continuous time series epochs, the accuracy of 71.75% was obtained.**

***Index Terms*—hand activity recognition, IMU, wrist sensor, deep learning, RNN**

I. Introduction

Recognition of human hand activities could generate new information towards hand gesture user interface and daily hand activity loggings applications since this could provide contextual information of activity logs or gestures. To measure hand motions, recently inertial measurements units (IMUs) are readily available on smartphones, smart bands, and smart watches. Each IMU includes a set of tri-axial accelerometer, gyroscope, and magnetometer. With novel classification algorithms, one can recognize hand activities by classifying IMU signal features.

Substantial challenges still remain for hand activity recognition using wearable sensors. First, data recorded using wearable sensors contains very complex contexts of signal variations due to the freedom of hand movements. Second, this kind of HAR works requires a database of sensor recordings with the ground truth annotation of true hand activities, but it is not easy to collect such data. Third, extracting and selecting features to be used in classification are necessary for most applications. In general, most classifiers cannot handle raw sensor directly as inputs. Traditional hand activity recognition works used different classification methods, such as decision tree, k-nearest neighbor, Naïve Bayes, and Support Vector Machines [1]. Even for time sequential data, Conditional Random Forest or Hidden Markov Model were used [2], [3]. Most of these conventional approaches require careful selection of extracted features.

Recently deep learning techniques offer a new opportunity to handle more complex data and inference. These deep learning methods have proven the potential to advance the state-of-the-art in HAR. Especially, these approaches overcome the needs of standard features extraction procedures, since they offer an advantage through their ability to learn and extract hidden representation from raw data and classify at the same time. Lately, some sophisticated models using these techniques have successfully been used for a challenging HAR tasks [4]. A dominant deep learning approach for HAR has been convolutional neural networks (CNNs). This technique utilizes convolutional kernels to extract key features from the temporal axis [5]-[7]. All the information obtained from the convolutional operations was unified to estimate the probability for human activities. Recently sequential modeling approach has been employed with favorable results via recurrent neural networks (RNNs) [4], [8], [9]. Specifically, these models are based on Long Shot Term Memory (LSTM) [10]. Combining CNN and RNN previous works were able to increment the performance of their results. RNN algorithm allows taking into account not only the current input data also previous ones. The memory unit improved by LSTM allowed a better abstraction of sequential input data like HAR using raw signals recorded from sensors. Most of these studies employed multiple IMU sensors to improve the recognition accuracy. There are only a few studies operating with a single IMU. A technical challenge remains to achieve feasible recognition with a single IMU and novel classifiers for practical applications.

In this work, we propose a deep learning RNN based framework for hand activities recognition using a single wearable IMU. We have evaluated our RNN-based HAR on the standard benchmark dataset Opportunity [11]. We focus on six representatives hand activities from the

Manuscript received June 9, 2017; revised September 2, 2017.

doi: 10.18178/ijpmbs.6.4.114-118

activities in the database. They are "Close Door," "Open Door," "Close Fridge," "Open Fridge," "Clean Table," and "Drink from Cup." In our evaluations, we test the performance with standard epoch activity data first. Second, we test with noise-added epoch activity data to assess the robustness of our proposed system. Finally, we test with continuous time series data considering a real application scenario.

The structure of this paper is organized as follows: Section 2 describes our proposed RNN-based HAR. Section 3 presents the database and our experimental results. Finally, we conclude the study in Section 4.

II. Methods

A. Our Proposed RNN-based HAR System

Considering the temporal sequences of natural human hand movements, we adopted RNN based on LSTM cells as a classification method in this work. From an IMU sensor located on the right wrist, thirteen features channels are utilized: three channels from a tri-axial accelerometer, tri-axial gyroscope, tri-axial magnetometer, and four channels from the quaternion sensor orientation. The input to the HAR is a matrix of stacked time series data corresponding to these feature channels. Fig. 1 shows our IMU- and RNN-based hand activity recognition. The system includes two recurrent layers: each one with 256 LSTM cells and a fully connected layer. Recurrent layers use a hyperbolic tangent activation function. The output layer uses a Softmax function to obtain the activities probabilities for input data. The final output is the recognized hand activity context.

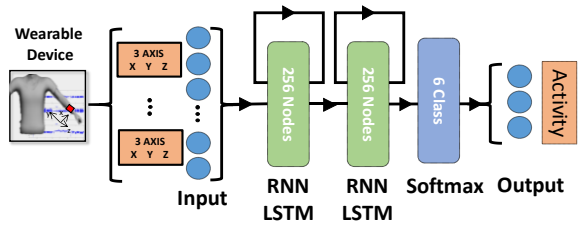

Figure 1. The proposed single IMU- and RNN-based hand activity recognition system.

B. Recurrent Neural Network

RNN offers an advantage to make a decision based on current and past inputs. RNN is a specialized neural network to process sequential data of $x^{(1)}, \ldots, x^{(\tau)}$. One of the feature that makes RNN possible to process sequential information is the parameter sharing by which the model extends the same output to different time steps and reduces the amount of data to be learned. Unlike traditional deep neural networks, the parameter sharing is important when relevant information has to be recognized, if it occurs at multiple positions. LSTM proposed by [10] avoids the vanishing gradient problem in the training process by the backpropagation update algorithm. LSTM creates internal paths that help to preserve errors for long durations. A variant of the backpropagation algorithm, Backpropagation Through Time (BPTT) is used to train recurrent networks as described in [12]. BPTT reduces the complexity of the parameters update in RNN and allows training networks faster. An overview of the internal LSTM memory unit structure can be given as follows: the input gate (i_t) determines which value is an update; the forget gate determines what information set away; the output gate (o_t) controls what information is going to be the output of the cell. Fig. 2 shows a single LSTM cell with their hidden connection (h_t), the connection between hidden nodes (c_t), and internal cell connection (f_t). Note that LTSM works with time varying signals. Equations (1)~(5) offer a mathematical description. Fig. 2 represents the cell operations graphically.

$$i_t = \sigma(\mathbf{W_{xi}x_t} + \mathbf{W_{hi}}h_{t-1}+\mathbf{b_i}) \quad (1)$$

$$f_t = \sigma(\mathbf{W_{xf}x_t} + \mathbf{W_{hf}}h_{t-1}+\mathbf{b_f}) \quad (2)$$

$$c_t = f_t c_{t-1}+i_t \tanh(\mathbf{W_{xc}x_t} + \mathbf{W_{hc}}h_{t-1}+\mathbf{b_c}) \quad (3)$$

$$o_t = \sigma(\mathbf{W_{xo}x_t} + \mathbf{W_{ho}}h_{t-1}+\mathbf{b_c}) \quad (4)$$

$$h_t = o_t \tanh(c_t) \quad (5)$$

Where **W** indicates weight and **b** indicate bias.

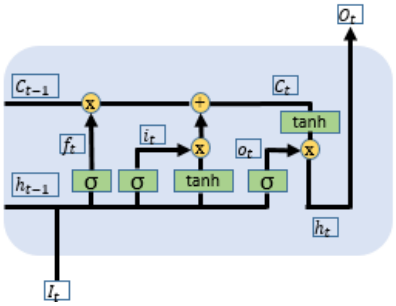

Figure 2. The structure of LSTM cell gates.

C. Training and Evaluation

Training and testing were performed on a PC with Intel Core i7-2600 CPU, 16 GB RAM, and GPU NVIDIA GeForce GT630. Algorithms were developed using a Java toolkit, DeepLearning4j [13] for building, training, and deploying neural networks. This library supports different types of neural networks and allows to compose deep neural nets from various shallow nets, such as autoencoders, convolutional nets, RNN, LSTM, Bidirectional LSTM, and recurrent base networks.

We trained our RNN-based HAR using early-stopping as a regularization method to avoid overfitting. We established a set up of 200 training steps for the algorithm with an early stop of 20 steps, which means that if no improvement is shown in 20 training steps, the algorithm would be stopped. Moreover, input data was fed using mini-batches of 32. We follow with the evaluating to the proposed RNN-based HAR in three different settings. In the first state, standard discrete epochs only reflect activities of interest. In the second state, the system tested corrupted epochs with some external noise. Final evaluation simulates continuous time series as a real application with streaming values segmented and feeds to the HAR.

Evaluation of the recognition results was obtained regarding the accuracy and the mean F1-score:

$$Acc = \frac{1}{|c|}\sum_{c}\frac{True\ Positive_c + True\ Negative_c}{Condition\ Positive_c + Condition\ Negative_c}$$

$$F_m = \frac{2}{|c|}\sum_{c}\frac{precision_c * recall_c}{precision_c + recall_c}$$

Where c is the current class and $|c|$ is the total number of classes. This performance metric is independent of the class distribution as suggested in [11], which considers all activities relevant.

III. Experiment and Results

A. Hand Activity Database

Opportunity database contains information from different sensors that measured daily activities in a kitchen. Records were obtained using 72 sensors of 10 modalities placed in the environment, objects, and subjects. The database includes complex naturalistic activities from four subjects in six different sessions. The first five sessions correspond to daily living activities (ADL) where each subject performed household activities without any particular protocol. The last record corresponds to a drill session that collects a large number of activity instances. The hand activity recognition challenge includes an 18-class classification problem for right arm activities. They measure their performance using the F-measure or weighted F-measure. Activities labeled in the database include Open Door 1, Open Door 2, Close Door 1, Close Door 2, Open Fridge, Close Fridge, Open Dishwasher, Close Dishwasher, Open Drawer 1, Close Drawer 1, Open Drawer 2, Close Drawer 2, Open Drawer 3, Close Drawer 3, Clean Table, Drink from Cup, and Toggle Switch.

In the Opportunity dataset, five Xsense IMUs positioned in a custom jacket were used. Each IMU contained a tri-axial acceleration sensor, a gyroscope, and a magnetic field sensor. This unit communicates with the receiver through a serial bus that was connected by USB or Bluetooth. The database was created using a dedicated laptop carried by the subjects in a backpack. The CRN toolbox [14] was used to manage data acquisition, which is based on the Portable Operating System Interfaces that permits quick construction of complex systems.

In this work, we used the same guidelines presented in the task of multimodal activity recognition for the Opportunity challenge [11]. Out of the 18 hand activities, we selected eight including Close Door 1, Close Door 2, Open Door 1, Open Door 2 Close Fridge, Open Fridge, Clean Table, and Drink from Cup. We unified similar activities into one class, and end up with Close Door and Open Door classes. From the IMU signals in Opportunity, with a sliding window of 2-second, we segmented the data into epochs with a fixed length and overlap of 50% Our proposed IMU- and RNN-based HAR was tested with the discrete epoch datasets.

We trained and tested our RNN-based HAR first by creating standard and noise-added epochs from time series data. Each epoch was made with a 2-second segment, corresponding to activities in all ADL and drill session from Subject 1 and ADL1, ADL2, ADL3, and drill sessions from Subjects 2 and 3. The test set was made using ADL 4 and ADL5 from Subjects 2 and 3. The total number of standard discrete epochs in the first validation was 2,393 for training and 272 for testing for six classes. The noise-added activity epochs were derived including the ADL and Drill sessions from Subject 4 for training. Subject 4 data contain artificially added rotational noise, which was a part of the task to test robustness to noise in the Opportunity challenge. A total number of training epochs including noise were 3,613, and the test epochs remain the same. We also created a continuous time series data set from the drill session of Subject 2. Continuous test dataset contains a total of 675 epochs.

B. Recognition Results with Hand Activity Epoch Data.

Table I shows the averaged confusion matrix from a five-fold test, reporting the overall recognition accuracy of 80.09% ± 1.57% and the mean and standard deviation of an F1 score of 0.789 ± 0.02. The LSTM cells proved to be useful to distinguish between similar events like Open/Close Door, or Open/Close Fridge. These activities are defined by the same basic hand motions but differ in their sequence of action. The sensor orientation in quaternions also helped to improve the recognition accuracy of the system.

TABLE I. Confusion Matrix for RNN-Based Hand Activities Classification with the Standard Epoch Data

Activities	Model Classification					
	Open Door	***Close Door***	***Open Fridge***	***Close Fridge***	***Clean Table***	***Drink from Cup***
Open Door	**85.37%**	9.76%	0.00%	2.44%	0.00%	2.44%
Close Door	13.16%	**84.21%**	0.00%	0.00%	0.00%	2.63%
Open Fridge	0.00%	0.00%	**78.72%**	17.02%	0.00%	4.26%
Close Fridge	0.00%	0.00%	11.11%	**72.22%**	5.56%	11.11%
Clean Table	3.57%	0.00%	0.00%	7.14%	**82.14%**	7.14%
Drink from Cup	17.82%	4.95%	0.00%	1.98%	0.00%	**75.25%**

After evaluating the HAR with the standard epochs, we applied the trained system again to the noise-added epochs. Table 2 shows the confusion matrix with the recognition accuracy of 74.92% ± 0.69% and an F1 score of 0.737 ± 0.049. Fig. 3 shows a comparison between of the precision recognition values with the standard and noise-added epochs. Even under the noise, our proposed HAR performs with little-diminished accuracy. The results show the feasibility for real life logging and gesture interface applications. However, noise-added epochs created confusion for the hand activities like Open Door and Close Fridge. Furthermore, misclassification among all other activities (or classes) got more prominent.

TABLE II. CONFUSION MATRIX FOR RNN-BASED HAND ACTIVITIES CLASSIFICATION WITH THE NOISE-ADDED EPOCH DATA

Activities	Model Classification					
	Open Door	*Close Door*	*Open Fridge*	*Close Fridge*	*Clean Table*	*Drink from Cup*
Open Door	**75.47%**	20.75%	0.00%	1.89%	0.00%	1.89%
Close Door	19.57%	**72.57%**	0.00%	0.00%	0.00%	10.87%
Open Fridge	0.00%	0.00%	**81.82%**	18.18%	0.00%	0.00%
Close Fridge	13.04%	0.00%	13.04%	**69.57%**	0.00%	4.35%
Clean Table	5.88%	2.94%	8.82%	2.94%	**73.53%**	5.88%
Drink from Cup	9.32%	4.24%	2.54%	5.08%	0.00%	**78.81%**

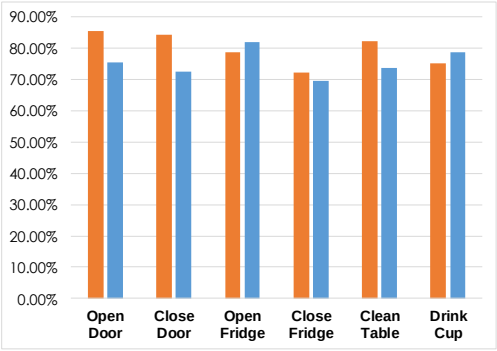

Figure 3. Recognition accuracies without noise (left in orange), and with noise (right in blue).

C. Recognition Results with Continuous Activity Data.

With the previous trained RNN-based HAR, we tested the continuous time series data. Fig. 4 shows a portion of the model output displaying the difference within the ground- truth for various activities performed. Our proposed RNN-based HAR achieves the overall accuracy of 71.75%. We noticed that most of the confusion occurs between transient states between activities. One option to reduce confusion from these transient states is to use an averaging filter of states: update the current state based on the previous and next recognized states. This could make smoother transitions in the recognized activity logs.

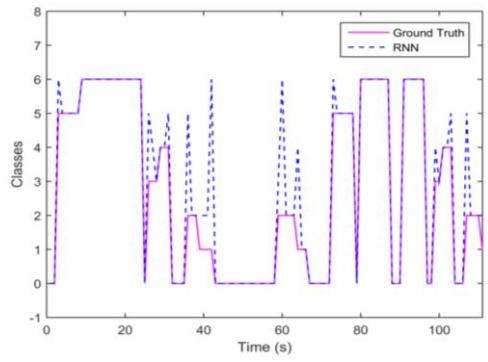

Figure 4. HAR results with continuoustime series data. Ground-truth activities are shown in magenta and the recognized in blue. class 0 represents null class, class 1 open door, class 2 close door, class 3 open fridge, class 4 close fridge, class 5 clean table, and class 6 drink from cup.

IV. CONCLUSION

In this paper, a proposed IMU- and RNN-based HAR for recognizing six different hand activities of daily living were developed using signals from a single IMU sensor positioned on the right wrist. The performance of this RNN HAR approach holds a potential for hand activity recognition using only a single IMU sensor.

ACKNOWLEDGMENTS

This work was supported by International Collaborative Research and Development Program (funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) (N0002252). This material is based upon work supported by the Ministry of Trade, Industry & Energy (MOTIE, Korea) under Industrial Technology Innovation Program (No. 10063300).

REFERENCES

[1] L. V. Nguyen-Dinh, D. Roggen, A. Calatroni, and G. Tröster, "Improving online gesture recognition with template matching methods in accelerometer data," in *Proc. 12th International Conference on Intelligent Systems Design and Applications(ISDA)*, Kochi, 2012, pp. 831-836.

[2] E. Garcia-Ceja, R. F. Brena, J. C. Carrasco-Jimenez, and L. Garrido, "Long-term activity recognition from wristwatch accelerometer data," *Sensors (Basel, Switzerland)*, vol. 14, no. 12, pp. 22500-22524, 2014.

[3] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, pp. 1-33, 2014.

[4] F. J. Ordóñez and D. Roggen, "Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition," *Sensors (Switzerland)*, vol. 16, no. 1, p. 115, 2016.

[5] J. B. Yang, M. N. Nguyen, P. P. San, X. L. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Proc. of the Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015, pp. 3995-4001.

[6] S. Ha, J.-M. Yun, and S. Choi, "Multi-modal convolutional neural networks for activity recognition," in *Proc. International Conference on Systems, Man, and Cybernetics (SMC)*, 2015, pp. 3017-3022.

[7] M. Zeng, *et al.*, "Convolutional neural networks for human activity recognition using mobile sensors," in *Proc. 6th International Conference on Mobile Computing, Applications and Services*, Austin, 2014.

[8] N. Y. Hammerla, S. Halloran, and T. Ploetz, "Deep, convolutional, and recurrent models for human activity recognition using wearables," in *Proc. of International Joint Conference on Artificial Intelligence*, 2016.

[9] M. Edel and K. Enrico, "Binarized-BLSTM-RNN based human activity recognition," in *Proc. of International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, Alcala de Henares, 2016.

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

[11] R. Chavarriaga, *et al.*, "The opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognition Letters*, vol. 34, no. 15, pp. 2033-2042, November 2013.

[12] F. A. Gers and N. N. Schraudolph, "Learning precise timing with LSTM recurrent networks," *JMLR*, vol. 3, pp. 115-143, 2002.

[13] D. D. Team. Deeplearning4j: Open-source distributed deep learning for the JVM. *Apache Software Foundation License*. [Online]. Available: http://deeplearning4j.org.

[14] D. Bannach, K. Kunze, P. Lukowicz, and O. Amft, "Distributed modular toolbox for multi-modal context recognition," in *Proc. International Conference on Architecture of Computing Systems*, 2006, pp. 99-113.

Patricio A. Rivera received his B.E. degree in Electronics, Automation and Control Engineering from University of the Armed-Forces-ESPE, Ecuador. He is currently working toward his Ph.D. degree in the Department of Biomedical Engineering at Kyung Hee University, Republic of Korea. His research interest includes artificial intelligence, signal processing, and machine learning.

Tae-Seong Kim received the B.S. degree in Biomedical Engineering from the University of Southern California (USC) in 1991, M.S. degrees in Biomedical and Electrical Engineering from USC in 1993 and 1998 respectively, and Ph.D. in Biomedical Engineering from USC in 1999. After his postdoctoral work in Cognitive Sciences at the University of California at Irvine in 2000, he joined the Alfred E. Mann Institute for Biomedical Engineering and Dept. of Biomedical Engineering at USC as Research Scientist and Research Assistant Professor. In 2004, he moved to Kyung Hee University in Korea where he is currently Professor in the Department of Biomedical Engineering. His research interests have spanned various areas of biomedical imaging, bioelectromagnetism, neural engineering, assistive biomedical lifecare technologies. Dr. Kim has been developing advanced signal and image processing methods, pattern classification, machine learning methods, novel medical imaging modalities, and rehabilitation technologies. Dr. Kim has published more than 300 papers and seven international book chapters. He holds ten international and domestic patents and has received nine best paper awards.