A Single Wearable IMU-based Human Hand Activity Recognition via Deep Autoencoder and Recurrent Neural Networks

P. Rivera Lopez[1], E. Valarezo Añazco[1,2], S. M. Lee[1], K. M. Byun[1], M. H. Cho[1], S. Y. Lee[1], and T.-S. Kim[1]*
[1]Department of Biomedical Engineering
Kyung Hee University
Yongin, Republic of Korea
[2]Faculty of Engineering in Electricity and Computation, FIEC
Escuela Superior Politécnica del Litoral, ESPOL
Guayaquil, Ecuador
email: {patoalejor, edgivala, sangmlee, kmbyun, mhcho, sylee01}@khu.ac.kr
*Corresponding author email: tskim@khu.ac.kr

***Abstract*— Human Hand Activity Recognition (HAR) using wearable sensors can be utilized in various practical applications such as lifelogging, human-computer interaction, and gesture interfaces. Especially with the latest deep learning approaches, the feasibility of HAR in practice gets more promising. In this paper, we present a HAR system based on deep Autoencoder for denoising and deep Recurrent Neural Network (RNN) for classification. The proposed HAR system achieves a mean accuracy of 79.38% for seven complex hand activities, while only of 72.65% without the autoencoder. The presented combination of autoencoder and RNN could be useful for much improved human activity recognition.**

Keywords- Human Hand Activity Recognition; Autoencoder; Deep Learning; RNN; CNN.

I. INTRODUCTION

Human Hand Activity Recognition (HAR) is an essential technology in many user-centric applications such as human-computer interactions, assisted living, smart homes, and lifelogging [1]. In general, there are two ways for HAR: using imaging sensors or inertial sensors that capture human activities [2]. Wearable devices are generally equipped with inertial sensors such as accelerometer, gyroscope, and magnetometer, which have proven useful for HAR. There have been many studies recognizing Activities of Daily Living (ADL) with these wearable devices [1]-[10]. Besides, various classifiers have been employed such as Hidden Markov Models (HMM), Support Vector Machine (SVM), and Restricted Boltzmann Machines (RBMs) [3], [4], [5].

Recently, data-driven approaches using deep learning for HAR have led to a significant recognition improvement by self-learning without the need of handcrafting features [6], [7]. Approaches based on Convolutional Neural Networks (CNN) demonstrate the advantages of using convolutional filters to capture local dependencies and scale invariance features. Previous works, such as [8] and [9] applied CNN to extract features from multi-channel sensor data and recognized locomotion activities such as walking, sitting, walking upstairs, and walking downstairs.

Recently, there is a growing interest in hand activity recognition [10], due to the widespread use and availability of wristbands and smartwatches. In the work [11], CNN was utilized to recognize multiple daily life hand activities from multiple sensors signals. Approaches in [12] and [13] used Recurrent Neural Networks (RNN) to recognize locomotion and hand gestures using multiple Inertial Measurement Units (IMU) on the wrist and body parts. The work in [14] presented improvements in a multi-sensor based HAR combining CNN and RNN.

Although these previous studies accomplished some success recognizing hand activities, because of the delicate movements of hands and sensor noise, some additional preprocessing is needed to improve the recognition rate. One latest study in [15] examined different motion artifacts in constrained and free-mode motion sensor networks and demonstrated the effect of alleviating noise motion artifacts in HAR performance.

In this work, we present a HAR system for daily hand activities consisting of a deep autoencoder for denoising and a deep RNN for classification. As reducing signal noise and improving signal representations can be dealt with a deep autoencoder [16], we have designed a supervised autoencoder for denoising and better signal representation. Then, a classifier based on RNN recognizes daily hand activities using only the signals from a single IMU on one dominant wrist. Our results show a significant improvement in recognizing complex hand activities.

The rest of this paper is organized as follows. Section II describes the proposed methodology. In Section III, the experimental results of HAR are presented. Finally, the conclusion is given.

II. METHODS

Our proposed hand activity recognition system is shown in Figure 1. The input signal is composed of thirteen feature channels collected from a single IMU sensor at the right wrist of subjects. The Autoencoder (AE) module processes this input signal and transfers to the RNN classifier for hand activity recognition.

A. Hand Activity Database

In this study, we utilized the Opportunity public database [17], which contains continuous time-series data of various human hand activities. The database includes the recordings from four subjects: each subject performed an unscripted session of hand movements and ADL without constraints.

 ISBN: 978-1-61208-657-6

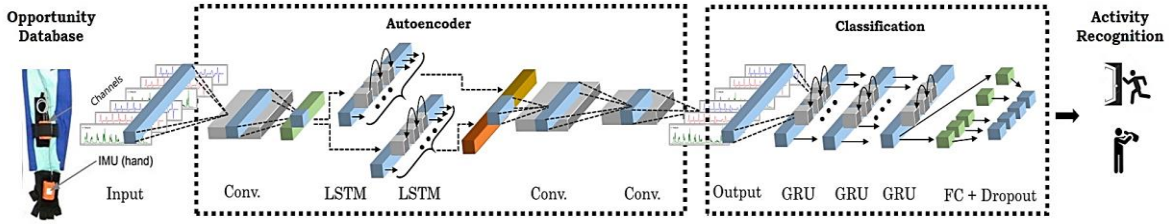

Figure 1. Proposed HAR system for hand activity recognition. From the left, signals coming from a single IMU go through our autoencoder module. Autoencoder reconstructs the data and transfer to RNN classifier. Classifier predicts activity class probabilities.

Each session was performed five times with different numbers of repetition for the activities. Additional hand activities were collected in an extra control (Drill) sessions, where each subject performed twenty scripted sequences of hand activities. We followed the Opportunity multi-modal gesture challenge guidelines in [17] to split the data into train and test datasets. We focus on data collected from a sensor placed on the right wrist of a custom jacket, which was worn by the subjects. This sensor included a commercial RS458-networked XSense IMU composed of a three-axis accelerometer, a three-axis gyroscope, a three–axis magnetometer, and four-channel quaternion orientation information.

From the total of hand gesture classes in the database, we selected thirteen activities of our interest. The activities that involve similar executions are grouped as the same class. Resultant seven classes of hand activities are Close Door (Close Door 1 and Close Door 2), Open Door (Open Door 1 and Open Door 2), Close Fridge, Open Fridge, Open Drawer (Open Drawer 1, Open Drawer 2, and Open Drawer 3), Close Drawer (Close Drawer 1, Close Drawer 2, and Close Drawer 3), and Drink from Cup.

Using a sliding window approach, the IMU signals were segmented with a window size of four seconds and an overlap of 50%. The data were normalized to a range of [-1, +1] with zero mean, which we denote them as epochs. Each epoch is tagged with a specific class label. We named these datasets of epochs as the IMU-train and IMU-test datasets respectively.

To train our supervised autoencoder, we modeled the previous datasets using an Autoregressive Moving Average (ARMA) model and named them as the ARMA-train and ARMA-test datasets. Training the AE used these ARMA datasets as the ideal targets of the reconstructed and denoised signals. Finally, the AE reconstructed outputs are named as the AE-train and AE-test datasets. The classifier uses these datasets for performance analysis of recognition.

B. Proposed Autoencoder

In this section, the proposed AE and RNN classifier are presented.

B1. Autoencoder Model

The encoder $f(x)$ in our AE architecture is a combination of a CNN layer and a Bidirectional RNN (BRNN). A convolution layer extracts features from the input signal through a one-dimensional filter. These features capture local correlations hidden in the data and form an augmented representation in a set of multiple feature maps [14]. We use the hyperbolic tangent function as a non-linear activation function for the output of the convolution. The RNN layers process sequential data, taking advantage of parameter sharing, making possible each unit in the output be a function of the previous units. BRNN takes the output from CNN and uses it in two parallel layers: forward and backward loops used for exploding context from the past and future of a specific time step. The BRNN units are based on Long Short Term Memory (LSTM) cells, which use a concept of gates that define the behavior of the memory cell. The input x_t is fed into different gates such as the forget gate f_t, input gate i_t, and output gate o_t with the previous cell output h_{t-1} to compute the current output. In the following equations, we describe the LSTM unit where σ represents a non-linear function and $[W, b]$ are the weight matrices and bias vector associated with each gate.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (4)$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t * \tanh(C_t) \quad (6)$$

From the encoder hidden representation h, the decoder $g(h)$ reconstructs the signals by two stacked convolutional layers. The last decoder convolution layer has its feature map size constrained to the same size of the input channels.

B2. ARMA Modeling of IMU Activity Signals

Before training the supervised AE, the ideal target dataset is obtained by modeling the original IMU datasets via ARMA. The Akaike information criterion was used to select an appropriate order for the autoregressive and moving average models. For each channel, an optimized model was carefully chosen from a pool of different combination of orders: in most cases, the autoregressive model order of 3 and moving

average of 4 were selected. The ARMA-train and ARMA-test datasets represent a denoised and improved representation of the signals in the IMU-train and IMU-test datasets. In Figure 2, one set of epoch instances from the IMU-test and ARMA-test datasets is shown.

B3. Training and Testing Autoencoder

The input to the AE was carried by mini-batches composed of epochs in the IMU-train dataset and target ARMA-train dataset. The AE used the Mean Square Error (MSE) as a loss function. The training algorithm iterated up to 100 training steps with a learning rate of 1e-4. Gradient decedent recursively updated the network parameters using Adam optimizer algorithm. Weights initialization used a random Gaussian distribution with a mean of zero and standard deviation of 0.5. To validate the AE performance, we quantified the similarity between the AE-test and ARMA-test datasets. This similarity is based on the overall Root Mean Square Error (RMSE) and Pearson Correlation Coefficient (R) for each corresponding channel from both datasets.

C. RNN Classifier

The classifier module is composed of three RNN layers based on Gate Recurrent Unit (GRU) memory cells. The GRU cell possesses a reset gate r and an update gate z, unlike the LSTM variant it does not have an internal memory c_t and an output gate o_t. The GRU cell combines the input gate i_t and forget gate f_t in the update gate, and directly apply the reset gate to the previously hidden state. We describe the GRU gates in the following equations:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (7)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (8)$$

$$\tilde{h}_t = tanh(W[r_t * h_{t-1}, x_t]) \quad (9)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (10)$$

The output from last RNN layer is connected to a dense layer to obtain the class probabilities. Despite the compelling representation from RGRU, there is still a possibility of overfitting. We address this using a dropout technique for optimization with a value of 0.4 before the dense layer. The final layer produces the class probabilities from a Softmax function. Initialization of the weights uses a random Gaussian distribution with of mean zero and standard deviation of 0.5. The network is trained over 50 training steps with a learning rate of 3e-4 with an optimization based on Adam algorithm. We compute the weighted F1-score and accuracy of classification for the given test datasets.

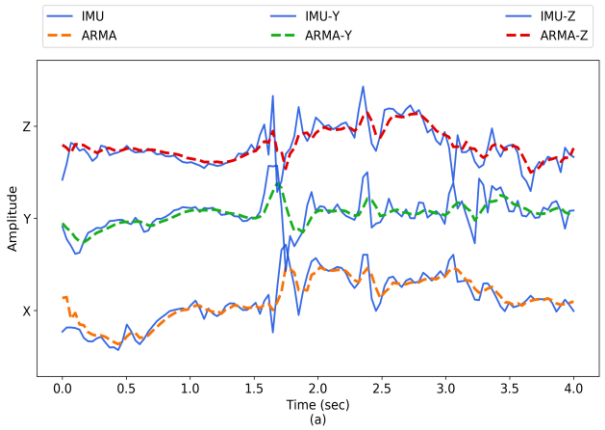

Figure 2. Time series from the 3-axis accelerometer in the "Open Door" activity: IMU (solid) and ARMA (dotted).

III. EXPERIMENTAL RESULTS

A. Validation of Autoencoder

We computed the RMSE and R coefficient between the ARMA-test and AE-test datasets to evaluate the performance of AE. Table 1 shows a summary of these values. The signals in Figure 3 illustrate an exemplary epoch of "Open Door" activity from both datasets.

TABLE 1. THE COMPUTED RMSE AND R-VALUES BETWEEN ARMA-MODELED AND AE OUTPUT DATASETS.

Channels	Axis	RMSE	R
Accelerometer	*X*	0.0441	0.9387
	Y	0.0383	0.9805
	Z	0.0359	0.9953
Gyroscope	*X*	0.0262	0.9652
	Y	0.0251	0.9820
	Z	0.0234	0.9872
Magnetometer	*X*	0.0130	0.9799
	Y	0.0111	0.9892
	Z	0.0122	0.9927
Quaternion	*Q1*	0.0328	0.9873
	Q2	0.0361	0.9914
	Q3	0.0340	0.9870
	Q4	0.0345	0.9976

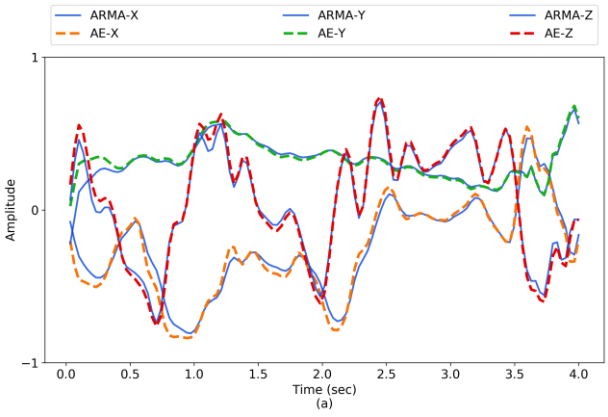

Figure 3. Time series from the 3-axis gyroscope in the "Open Door" activity: ARMA (solid) and AE (dotted).

A1. *Classification Performance*

The summary of the recall values achieved by the classifier on the IMU-test, ARMA-test, and AE-test datasets are shown in Table 2. Using the raw sensor signals in the IMU-test dataset, the classifier achieved a mean F1-score of 72.87% and accuracy of 72.65%. The recognition performance is not quite satisfactory for these complex hand activities. Using the ARMA-test dataset (i.e., modeled ideal dataset), recognition increased to a mean F1-score of 82.40% and accuracy of 82.14%. For activities such as "Open Fridge" and "Open Drawer," their recall values increased up to 78.33% and 81.55% respectively from around 60%. Finally, using the AE-test dataset (i.e., the output of AE), the classifier achieved a mean F1-score of 79.64% and accuracy of 79.38%, reflecting a 6.75% improvement over the raw signals from the IMU-test dataset and similar to the performance of the ARMA-test dataset.

TABLE 2. SUMMARY OF CLASSIFICATION PERFORMANCE FOR THE IMU-TEST, ARMA-TEST, AND AE-TEST DATASETS WITH THE PROPOSED CLASSIFIER.

Hand Activity Recognition	**Performance on Test Datasets (%)**			
	Activity	Raw IMU	ARMA	AE
Gestures	*OD*	83.89	88.33	91.67
	CD	85.80	88.27	86.42
	OF	61.00	78.33	73.00
	CF	63.89	70.83	68.75
	DC	84.08	89.20	87.44
	ODW	60.71	81.55	69.64
	CDW	57.58	66.67	68.18
Mean	*F1-score*	72.87	82.40	79.64
	Accuracy	72.65	82.14	79.38

*OD: Open Door, CD: Close Door, OF: Open Fridge, CF: Close Fridge, DC: Drink from Cup, ODW: Open Drawer, CDW: Close Drawer

A2. *Comparison of Related Works*

In this work, we have implemented a HAR system of deep denoising AE and RNN classifier, through which the improved representation of activity signals are utilized to recognize seven daily hand activities using only a single IMU sensor.

There are rare works of HAR systems utilizing denosing AE. The HAR work in [15] used an unsupervised Variational Autoencoder (VAE) in combination of CNN with LSTM. It shown that using 75 sensor channels that presented significative motion artifacts from Opportunity the denoised signals could improve the accuracy from 72.96% to 90.81%. Also there have been HAR works utilzing multiple sensors (i.e., >70 sensor channels) to improve the performance. These studies reported F1-score of 75.4% [12], a recall value of 83.5% [13] , and F1-score of 86.6% [14] without the use of AE. In contrast with those studies, our architecture receive an input data compose of 13 feature channels extracted from only one IMU sensor, which is more practical for an end-user application.

IV. CONCLUSION

In this work, we have presented a HAR system for daily human hand activities combining a denoising autoencoder and RNN for classification. Our results prove that AE helps the deep classifier and eventually HAR by reducing noises and representing signals better. The promising results demonstrate the effectiveness of this approach, which could be used for other HAR systems.

ACKNOWLEDGMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP; Ministry of Science, ICT & Future Planning) (No. NRF-2017M3A9E2062707), and by International Collaborative Research and Development Program funded by the Ministry of Trade, Industry and Energy (MOTIE, Korea) (N0002252).

REFERENCES

[1] A. Bulling, U. Blanke, and B. Schiele, "A Tutorial on Human Activity Recognition Using Body-worn Inertial Sensors," *ACM Comput. Surv.*, vol. 46, no. 3, p. 33:1--33:33, Jan. 2014.

[2] A. M. Khan, Y.-K. Lee, S. Y. Lee, and T.-S. Kim, "A Triaxial Accelerometer-based Physical-activity Recognition via Augmented-signal Features and a Hierarchical Recognizer," *Trans. Info. Tech. Biomed.*, vol. 14, no. 5, pp. 1166–1172, Sep. 2010.

[3] E. Garcia-Ceja, R. F. Brena, J. C. Carrasco-Jimenez, and L. Garrido, "Long-Term Activity Recognition from Wristwatch Accelerometer Data," *Sensors*, vol. 14, no. 12, pp. 22500–22524, 2014.

[4] S. Bhattacharya and N. D. Lane, "From smart to deep: Robust activity recognition on smartwatches using deep learning," in *2016 IEEE International Conference on Pervasive Computing and Communication Workshops (PerCom Workshops)*, 2016, pp. 1–6.

[5] H. P. Gupta, H. S. Chudgar, S. Mukherjee, T. Dutta, and K. Sharma, "A Continuous Hand Gestures Recognition Technique for Human-Machine Interaction Using Accelerometer and Gyroscope Sensors," *IEEE Sens. J.*, vol. 16, no. 16, pp. 6425–6432, Aug. 2016.

[6] M. Zeng *et al.*, "Convolutional Neural Networks for human activity recognition using mobile sensors," in *6th International Conference on Mobile Computing, Applications and Services*, 2014, pp. 197–205.

[7] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A Survey," *Pattern Recognit. Lett.*, 2018.

[8] Y. Zheng, Q. Liu, E. Chen, Y. Ge, and J. L. Zhao, "Exploiting multi-channels deep convolutional neural networks for multivariate time series classification," *Front. Comput. Sci.*, vol. 10, no. 1, pp. 96–112, Feb. 2016.

[9] H. Gjoreski, J. Bizjak, M. Gjoreski, and M. Gams, "Comparing Deep and Classical Machine Learning Methods for Human Activity Recognition using Wrist Accelerometer," 2016.

[10] Choudhary, Santosh, and N. Choudhary, "Towards Developing an Effective Hand Gesture Recognition

System for Human Computer Interaction: A Literature Survey.," *Glob. J. Comput. Sci. Technol.*, 2016.

[11] 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 *Proceedings of the 24th International Conference on Artificial Intelligence*, 2015, pp. 3995–4001.

[12] N. Y. Hammerla, S. Halloran, and T. Plötz, "Deep, Convolutional, and Recurrent Models for Human Activity Recognition Using Wearables," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 2016, pp. 1533–1540.

[13] A. Murad and J.-Y. Pyun, "Deep Recurrent Neural Networks for Human Activity Recognition," *Sensors*, vol. 17, no. 11, 2017.

[14] F. J. Ordóñez and D. Roggen, "Deep Convolutional and LSTM Recurrent Neural Networks for Multimodal Wearable Activity Recognition," *Sensors*, vol. 16, no. 1, p. 115, 2016.

[15] S. Mohammed and I. Tashev, "Unsupervised deep representation learning to remove motion artifacts in free-mode body sensor networks," in *2017 IEEE 14th International Conference on Wearable and Implantable Body Sensor Networks (BSN)*, 2017, pp. 183–188.

[16] S. Rifai, P. Vincent, X. Muller, X. Glorot, and Y. Bengio, "Contractive Auto-encoders: Explicit Invariance During Feature Extraction," in *Proceedings of the 28th International Conference on International Conference on Machine Learning*, 2011, pp. 833–840.

[17] R. Chavarriaga *et al.*, "The Opportunity challenge: A benchmark database for on-body sensor-based activity recognition," *Pattern Recognit. Lett.*, vol. 34, no. 15, pp. 2033–2042, 2013.

 ISBN: 978-1-61208-657-6