Human Activity Recognition on Smartphones Using a Bidirectional LSTM Network

Fabio Hernández
Faculty of Electrical and Electronic Engineering
Pontifical Bolivarian University
Bucaramanga, Colombia
fabio.hernandezr@upb.edu.co

Luis F. Suárez
Faculty of Electrical and Electronic Engineering
Pontifical Bolivarian University
Bucaramanga, Colombia
luis.suarez.2014@upb.edu.co

Javier Villamizar
Faculty of Electrical and Electronic Engineering
Pontifical Bolivarian University
Bucaramanga, Colombia
javier.villamizar.2015@upb.edu.co

Miguel Altuve, *Senior Member, IEEE*
Faculty of Electrical and Electronic Engineering
Pontifical Bolivarian University
Bucaramanga, Colombia
miguel.altuve@upb.edu.co

***Abstract*—Thanks to the technological development on the last decades, inertial sensors have reduced their size to the point of occupying a small space inside a smartphone. The information of these sensors has been exploited in order to recognize human activities using different approaches, especially using machine learning techniques. This paper presents a human activity recognition (HAR) system that uses accelerometer and gyroscope data obtained from a smartphone as inputs to a bidirectional long short-term memory (LSTM) network. Six human activities were recognized: sitting, standing, laying, walking, walking upstairs, and walking downstairs. Different network architectures were tested using a grid search methodology. We have found that the easiest activity to recognize was laying down whereas the most difficult activity to recognize was sitting. Thanks to the property of bidirectional LSTM networks to process past and future information of a signal, walking downstairs and walking upstairs (two related activities) were recognized correctly. An overall accuracy of 92.67% was obtained using the proposed HAR system.**

***Index Terms*—Human activity recognition, pattern recognition, long short-term memory, recurrent neural network, smartphone.**

I. Introduction

In contrast to most species, human beings are characterized by their ability to perform a series of activities, even simultaneously, and by their ability to learn how to perform new tasks and to teach how to perform them to their offspring. These activities can be as simple as walking or as complex as sewing while talking. Nevertheless, the technological development and the resources availability have influenced our lifestyle, favoring a sedentary lifestyle and making us too much dependent on technology. A sedentary lifestyle and unhealthy eating habits increase the risk of cardiovascular diseases and type 2 diabetes [1], and negatively impact our quality of life and health expenditures, particularly in developing countries. Therefore, there is a growing interest in recognizing and monitoring the activities carried out by human beings. The automatic recognition of human activities can be used in different contexts, including health (fall detection system for the elderly [2], recognition of sedentary behavior [3], identification and assessment of laparoscopic skills [4]), video surveillance and security [5], sports (baseball [6] and football [7] actions recognition), and comfort (smart homes [8] and intelligent vehicles [9]).

Human activity recognition (HAR) is a promising pattern recognition research topic and a challenging classification task, particularly because of the complexity and diversity of activities, high dimensionality and quality of the data, intraclass variability (the same activity may vary from subject to subject), and interclass similarity (different activities may express similar shapes) [10]. Given a set of data from one or more sensors, HAR seeks to identify the current activity carried out by a human or group of humans. This objective can be accomplished using image processing, signal processing, and a combination of image and signal processing approaches. In this sense, human activities are broadly recognized by analyzing 2D and 3D videos and images from cameras, and/or by analyzing motion data from inertial sensors (accelerometer, gyroscope, magnetometers). However, thanks to the progress made in the last decade in the development of inertial sensors (wired and wireless) and microprocessors with the ability to acquire and process several signals simultaneously, while protecting privacy, and the possibility of embedding inertial sensors in smartphones, HAR using inertial sensors has become more popular, and easier and faster to perform. In addition, unlike specific purpose devices, such as wearable inertial sensors, smartphones with embedded inertial sensors are particularly suitable for HAR with potential applications in assisted living technologies given its multitasking ability, dimension and easy portability, without requiring additional equipment, and that it goes unnoticed as a sensing device [11].

In the last decade, given the growing interest in HAR, several datasets have been made freely available, such as the

978-1-7281-1491-0/19/$31.00 ©2019 IEEE

Opportunity dataset[1] in which several activities of daily living were recorded from 12 subjects using a multiple wireless and wired sensor systems integrated in the environment, in objects, and on the body [12]; the hand gesture dataset[2] in which 11 human's hand movements (eight gestures in daily living and three gestures in playing tennis) were collected from two subjects using body-worn accelerometer and gyroscope sensors [13]; and the Human Activity Recognition Using Smartphones dataset collected from 30 subjects using accelerometers and gyroscopes sensors from a smartphone, targeting the recognition of six different human activities [14].

An automatic HAR system generally involves data acquisition, activity detection, activity modeling, and activity classification. According to the nature of the sensing device, HAR methods can be i) unimodal, if data come from a single modality, such as images, and ii) multimodal, if data come from different sources [15]. Space-time, stochastic, rule-based and shape-based methods have been used for unimodal data, whereas affective, behavioral, and social networking methods have been used to treat multimodal data. For a thorough review of these methods, please refer to [15]. Among HAR methods, those based on machine and deep learning techniques have succeeded in HAR from image and time series data given their ability to handle high-dimensional complex data. For instance, on one hand, using data from body-worn sensors, an accuracy of 95.6% has been achieved with hidden Markov models to classify seven activities from four accelerometers [16], and an accuracy of 96% has been achieved with convolutional neural networks to classify eighteen activities from fifteen accelerometers and gyroscopic sensors [17]. On the other hand, using data from accelerometers and gyroscopic sensors embedded in a smartphone to classify six human activities, an accuracy of 96% has been achieved with support vector machines [14], and an accuracy of 93.79% has been achieved with a bidirectional long short-term memory (LSTM) network [18].

In this work, we propose an HAR system based on a bidirectional LSTM recurrent neural network. We identify tree static postures (sitting, standing, and laying) and tree dynamic motions (walking, walking upstairs, walking downstairs) collected from 30 volunteers using inertial sensors embedded in a smartphone that was worn on the subjects' waist. A grid search methodology was used to select the number of layers and the number of nodes per layer.

II. Materials and methods

A. Dataset description

Triaxial linear acceleration and triaxial angular velocity signals were collected at a sampling rate of 50 Hz using embedded accelerometer and gyroscope inertial sensors of a Samsung Galaxy S II smartphone from 30 volunteers (ages ranging from 19 to 48 years) while wearing the smartphone on the waist [14]. Subjects were asked to perform six activities: standing (ST), sitting (SD), laying down (LD), walking (WK), walking downstairs (WD), and walking upstairs (WU). The dataset is composed of 10299 instances of 2.56 s (128 samples) of triaxial acceleration from the accelerometer (total acceleration) and estimated body acceleration, and triaxial angular velocity from the gyroscope. Therefore, each subject is represented by nine signals. The dataset was randomly partitioned into training (70%) and test (30%) sets. The dataset is freely available in the UCI machine learning repository [3]

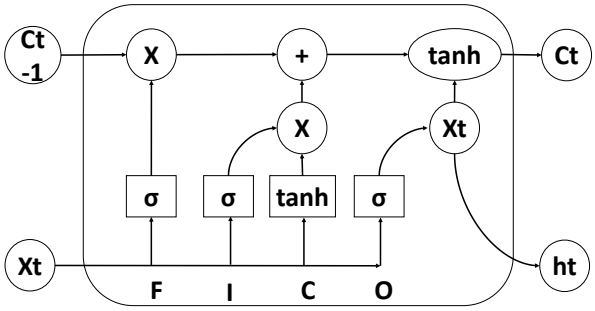


Fig. 1. Representation of an LSTM cell.

B. Bidirectional LSTM for HAR

A recurrent neural network is composed of memory cells: a kind of neuron, similar to a perceptron in an artificial neural network, where its output goes back into its input. One of the biggest problems with memory cells is that after a while, the network forgets the first inputs because memory cells have only short-term memory (vanishing gradient problem). LSTM cell was introduced to solve this issue and improve the accuracy. As shown in Fig 1, an LSTM cell is made up of four gates (forget, input, state and output) and one cell state that provide additional interactions. In the following equations, W_q represents the weights of the input and h_q represents the recurrent connection, q can be the input gate (i), output gate (o), the forget gate (f) or the memory cell (c). X represents the input vector to the LSTM unit.

The forget gate is responsible for taking information from the past that is not important and should be forgotten, the recurrent input h_{t-1} and current input x_t multiplied by their weights are the input of a sigmoid (σ) function, the (σ) output (h_{f_t}) is a number between 0 and 1 that is multiplied with cell state (c_{t-1}), if ($h_{f_t} = 1$) the LSTM keep this new information, otherwise, if ($h_{f_t} = 0$) the LSTM forget completely this new information:

$$\mathrm{h_{f_t}} = \sigma(\mathrm{W_f h_{t-1}} + \mathrm{W_f X_t}) \tag{1}$$

The input gate consists of a sigmoid function and its output h_{i_t} represents the value that is going to update the LSTM cell:

$$\mathrm{h_{i_t}} = \sigma(\mathrm{W_i h_{t-1}} + \mathrm{W_i X_t}) \tag{2}$$

[1]http://www.opportunity-project.eu/challengedatasetredirect.html

[2]https://github.com/andreas-bulling/ActRecTut

[3]https://archive.ics.uci.edu/ml/datasets/Human+Activity+Recognition+Using+Smartphones

The state gate creates a vector with different values, these values are combined with h_{i_t} and do a status update (eq. 5):

$$\mathrm{h_{c_t}} = \tanh(\mathrm{h_c h_{t-1}} + \mathrm{W_c X_t}) \tag{3}$$

The output gate represents the information from the cell state that should come out immediately combined with past information:

$$\mathrm{h_{o_t}} = \sigma(\mathrm{W_o h_{t-1}} + \mathrm{W_o X_t}) \tag{4}$$

The update states consist in forget what is to be forgotten and add what is to be added:

$$\mathrm{C_t} = \mathrm{h_{f_t}} * \mathrm{C_{t-1}} + \mathrm{h_{i_t}} * \mathrm{h_{c_t}} \tag{5}$$

And the LSTM output combines short and long terms together with

$$\mathrm{h_t} = \mathrm{h_{o_t}} * \tanh(\mathrm{c_t}) \tag{6}$$

An LSTM network takes into account past information, however, if present and future information is available (offline classifier), it would only take advantage of past information; future information would not be used. In contrast, a bidirectional LSTM (BLSTM) offers the possibility to take into account future and past information. On this basis, we propose the use of a bidirectional LSTM to take advantage of future and past information of the signals recorded during the human activities.

C. Configuration of the network and performance evaluation

The number of Bidirectional LSTM layers (L) and the number nodes per layer (N) were estimated using a grid search approach. The optimization looks to maximize the classification performance in term of accuracy. The search spaces were set to [1, 3] layers for L and [100, 300] neurons for N. The computing resources limited the range of the search spaces.

The input of the network is of size 9×128 (features $\times$ samples) and the output is a probability for each of the six classes. We used the weighted sum of squares of the errors as loss function, and the Adam optimization algorithm. The learning rate was set to 0.001.

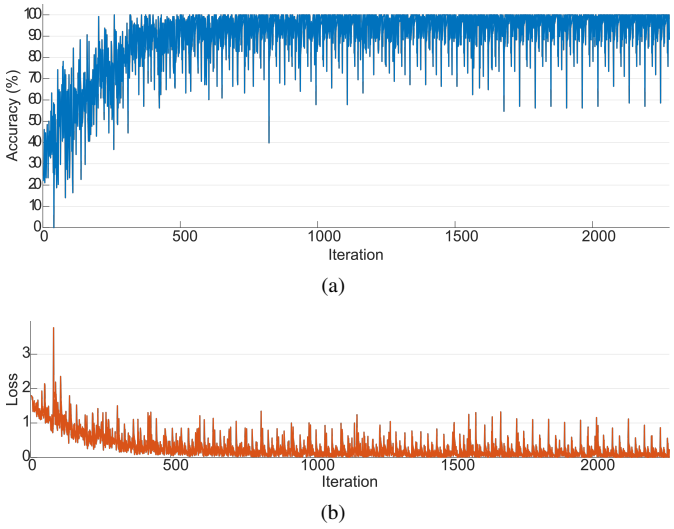
From a confusion matrix, precision (PRE), recall (REC) and accuracy (ACC) were used to evaluate the classification performance.

D. Hardware and software setup

The experiments were carried out in a personal computer with the specifications detailed in Table I.

TABLE I
HARDWARE AND SOFTWARE SPECIFICATIONS.

Feature	**Description**
OS	Windows 10 Home Single Language 64 bits
CPU	Intel(R) Core(TM) i7-6700HQ 2.60GHz (8 CPUs)
RAM	DDR4-1300 8 GB
GPU	NVIDIA GeForce GTX 950M
Computing envir.	MATLAB ver R2017b



(a)

(b)

Fig. 2. Example of accuracy (a) and loss (b) of the HAR system based on bidirectional LSTM.

III. RESULTS

Figure 2 shows an example of the training progress (accuracy and loss) of the proposed HAR system. We can see that, during training, the performance of the network stabilizes from 500 iterations.

Tables II and III show the accuracy for different combination of L and N in the training and test sets, respectively. We can observe that the classification performance varies for different values of L and N. The best classification performance (ACC = 92.67%) was obtained for 3 layers and 175 nodes per layer.

TABLE II
ACCURACY ON THE TRAINING SET FOR DIFFERENT L AND N. ACCURACY IS EXPRESSED AS A PERCENTAGE. THE HIGHEST ACCURACY IS HIGHLIGHTED IN BOLD.

L	N						
	100	125	150	175	200	250	300
1	**95.95**	94.25	95.84	94.30	95.92	92.79	94.52
2	95.18	94.82	95.70	95.32	95.02	95.62	95.16
3	74.09	94.86	95.13	95.66	94.76	91.97	92.75

TABLE III
ACCURACY ON THE TEST SET FOR DIFFERENT L AND N. ACCURACY IS EXPRESSED AS A PERCENTAGE. THE HIGHEST ACCURACY IS HIGHLIGHTED IN BOLD.

L	N						
	100	125	150	175	200	250	300
1	92.06	90.50	91.35	90.77	90.26	89.21	90.40
2	89.41	90.46	90.50	91.55	90.70	90.87	89.72
3	69.70	88.06	90.80	**92.67**	89.45	87.21	88.12

Tables IV and V show the confusion matrices in the training and test sets for the best network architecture ($L = 3$ layers and $N = 175$ nodes), respectively. Two static postures, standing and sitting, were the most difficult activities to recognize. The training of this model was carried out in 30 minutes.

TABLE IV
CONFUSION MATRIX ON THE TRAINING SET FOR THE BEST CLASSIFICATION PERFORMANCE.

	W	WU	WD	ST	SD	LD	**REC** (%)
W	1226	0	0	0	0	0	**100**
WU	0	1073	0	0	0	0	**100**
WD	0	0	986	0	0	0	**100**
ST	2	0	0	1136	148	0	**88.3**
SD	5	0	0	164	1205	0	**87.7**
LD	0	0	0	0	0	1407	**100**
PRE (%)	**99.4**	**100**	**100**	**87.4**	**89.1**	**100**	**95.66**

TABLE V
CONFUSION MATRIX ON THE TEST SET FOR THE BEST CLASSIFICATION PERFORMANCE.

	W	WU	WD	ST	SD	LD	**REC** (%)
W	488	3	5	0	0	0	**98.4**
WU	1	448	22	0	0	0	**95.1**
WD	0	0	420	0	0	0	**100**
ST	0	25	0	375	91	0	**76.4**
SD	5	2	0	62	463	0	**87.0**
LD	0	0	0	0	0	537	**100**
PRE (%)	**98.8**	**93.7**	**94.0**	**85.8**	**83.6**	**100**	**92.67**

IV. DISCUSSION

According to Tables II and III, although the best classification performance in training was achieved for one layer with 100 nodes, the detection performance on the test set drops by 3.89% on the test set. Using three layers and 175 nodes, the detection performance on the test set was maximized.

Using the proposed approach, all activities were correctly identified with a high classification performance except sitting and standing, two static postures with similar signal characteristics. Two activities that also show similar signals characteristics are walking downstairs and walking upstairs, however the property of the birectional LSTM to use past and future observation in the signal allows to correctly differentiating these two dynamic motions.

Results from this work are comparable with the best-related works using the same database. For instance, using a convolutional neural network, Ronao and Cho [19] obtained an accuracy of 94.79% using raw data and 95.75% using the Fourier transform of the data, and Jiang and Yin [20] reported an accuracy of 91.38% using the 2D wavelet transform and 95.18% using the discrete Fourier transform. Similarly, using hidden Markov models, Ann and Cho [21] obtained an accuracy of 91.76%, using a recurrent neural network, Inoue, Inoue, and Nishida [22] obtained 83.43% of accuracy, and Yu and Qin [18] reported an accuracy of 93.79% using Bidirectional LSTM Networks.

V. CONCLUSION

This paper has focused on the automatic classification of six human activities (standing, sitting, laying down, walking, walking downstairs, and walking upstairs) using signals from accelerometers and gyroscopic sensors in a smartphone by applying a deep learning technique for time series data that take into account past and future information. The Bidirectional LSTM network led to different classification performances on each class: the best classification performance was achieved to recognize the laying down activity (100%) while the worst classification performance was achieved to recognize the standing activity (REC=76.4% and PRE=85.8%). Different network architectures were tested using a grid search approach, guaranteeing the best performance possible for training and test. Future works are directed to the use of signal transformation (Fourier transform, integrals and derivatives) that can better characterize the different signals behaviors using deep learning approaches.

REFERENCES

[1] F. B. Hu, "Sedentary lifestyle and risk of obesity and type 2 diabetes," *Lipids*, vol. 38, no. 2, pp. 103–108, 2003.

[2] G. Mastorakis and D. Makris, "Fall detection system using kinect's infrared sensor," *Journal of Real-Time Image Processing*, vol. 9, no. 4, pp. 635–646, 2014.

[3] E. Kańtoch, "Recognition of sedentary behavior by machine learning analysis of wearable sensors during activities of daily living for telemedical assessment of cardiovascular risk," *Sensors*, vol. 18, no. 10, p. 3219, 2018.

[4] R. Laverde, C. Rueda, L. Amado, D. Rojas, and M. Altuve, "Artificial neural network for laparoscopic skills classification using motion signals from apple watch," in *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2018, pp. 5434–5437.

[5] W. Lin, M.-T. Sun, R. Poovendran, and Z. Zhang, "Activity recognition using a combination of category components and local models for video surveillance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 18, no. 8, pp. 1128–1139, 2008.

[6] A. Piergiovanni and M. S. Ryoo, "Fine-grained activity recognition in baseball videos," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, 2018, pp. 1853–1861.

[7] T. Tsunoda, Y. Komori, M. Matsugu, and T. Harada, "Football action recognition using hierarchical lstm," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 155–163.

[8] L. Chen, C. D. Nugent, and H. Wang, "A knowledge-driven approach to activity recognition in smart homes," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 961–974, 2012.

[9] E. Ohn-Bar, A. Tawari, S. Martin, and M. M. Trivedi, "On surveillance for safety critical events: In-vehicle video networks for predictive driver assistance systems," *Computer Vision and Image Understanding*, vol. 134, pp. 130–140, 2015.

[10] S. Zhang, Z. Wei, J. Nie, L. Huang, S. Wang, and Z. Li, "A review on human activity recognition using vision-based method," *Journal of healthcare engineering*, vol. 2017, 2017.

[11] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "Human activity recognition on smartphones using a multiclass hardware-friendly support vector machine," in *International workshop on ambient assisted living*. Springer, 2012, pp. 216–223.

[12] D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, D. Bannach, G. Pirkl, A. Ferscha *et al.*, "Collecting complex activity datasets in highly rich networked sensor environments," in *Networked Sensing Systems (INSS), 2010 Seventh International Conference on*. IEEE, 2010, pp. 233–240.

[13] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[14] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. L. Reyes-Ortiz, "A public domain dataset for human activity recognition using smartphones," in *Proceedings of the 21th International European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2013, pp. 437–442.

[15] M. Vrigkas, C. Nikou, and I. A. Kakadiaris, "A review of human activity recognition methods," *Frontiers in Robotics and AI*, vol. 2, p. 28, 2015.

[16] A. Mannini and A. M. Sabatini, "Machine learning methods for classifying human physical activity from on-body accelerometers," *Sensors*, vol. 10, no. 2, pp. 1154–1175, 2010.
[17] J. Yang, M. N. Nguyen, P. P. San, X. Li, and S. Krishnaswamy, "Deep convolutional neural networks on multichannel time series for human activity recognition," in *Ijcai*, vol. 15, 2015, pp. 3995–4001.
[18] S. Yu and L. Qin, "Human activity recognition with smartphone inertial sensors using bidir-lstm networks," in *2018 3rd International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*. IEEE, 2018, pp. 219–224.
[19] C. A. Ronao and S.-B. Cho, "Human activity recognition with smartphone sensors using deep learning neural networks," *Expert Systems with Applications*, vol. 59, pp. 235–244, 2016.
[20] W. Jiang and Z. Yin, "Human activity recognition using wearable sensors by deep convolutional neural networks," in *Proceedings of the 23rd ACM international conference on Multimedia*. ACM, 2015, pp. 1307–1310.
[21] C. A. Ronao and S.-B. Cho, "Human activity recognition using smartphone sensors with two-stage continuous hidden markov models," in *Natural computation (ICNC), 2014 10th international conference on*. IEEE, 2014, pp. 681–686.
[22] M. Inoue, S. Inoue, and T. Nishida, "Deep recurrent neural network for mobile human activity recognition with high throughput," *Artificial Life and Robotics*, vol. 23, no. 2, pp. 173–185, 2018.