Fuzzy Systems and Data Mining VII
A.J. Tallón-Ballesteros (Ed.)

doi:10.3233/FAIA210196

A Deep Learning Approach to Recognize Human Activity Using Inertial Sensors and Motion Capture Systems

M. Jaén-Vargas [a,1], K. Reyes Leiva [a, b], F. Fernandes [c], S.B. Gonçalves [e], M. Tavares Silva [e], D.S. Lopes [c, d], J. Serrano Olmedo [a, f]
[a] *Bioinstrumentation and Nanomedicine Laboratory (LBN), Center for Biomedical Technology (CTB), Universidad Politécnica de Madrid, Madrid, Spain*
[b] *Engineering Faculty, Universidad Tecnológica Centroamericana UNITEC, San Pedro Sula, Honduras*
[c] *INESC-ID, Lisbon, Portugal*
[d] *Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal*
[e] *IDMEC, Instituto Superior Técnico, University of Lisbon, Lisbon, Portugal*
[f] *Centro de Investigación Biomédica en Red en Bioingeniería, Biomateriales y Nanomedicina, (CIBER-BBN), Madrid, España*

Abstract. Human Activity Recognition (HAR) plays an important role in behavior analysis, video surveillance, gestures recognition, gait analysis, and posture recognition. Given the recent progress of Artificial Intelligence (AI) applied to HAR, the inputs that are the data from wearable sensors can be treated as time-series from which movement events can be classified with high accuracy. In this study, a dataset of raw sensor data served as input to four different deep learning networks (DNN, CNN, LSTM, and CNN-LSTM). Differences in accuracy and learning time were then compared and evaluated for each model. An analysis of HAR was made based on an attempt to classify three activities: walking, sit-to-stand, and squatting. We also compared the performance of two different sensor data types: 3-axis linear acceleration measured from two inertial measurement units (IMUs) versus 3D acceleration of two retro-reflective markers from the high-end optoelectronic motion capture system (MOCAP). The dataset created from observations of ten subjects was preprocessed with labelling and sliding windows and then used as input to the four frameworks. The results indicate that, for HAR prediction, linear accelerations estimated using IMUs are as reliable as those measured using the MOCAP system. Also, the use of the hybrid CNN-LSTM framework for both methods resulted in higher accuracy (99%).

Keywords. human activity recognition; motion capture; inertial measurement unit; artificial intelligence, deep learning.

1. Introduction

Currently, devices that aim to identify human activities play an important role in understanding how people perform their daily activities. The growing applications of Human Activity Recognition (HAR) include behavior analysis, video surveillance,

[1] Corresponding Author: Universidad Politécnica de Madrid, Parque Científico y Tecnológico de la UPM 28223, Pozuelo de Alarcón, Madrid, Spain; E-mail: milagros.jaen@ctb.upm.es

gestures recognition, gait analysis, and posture recognition [1]. Consequently, there are two types of HAR: that based on data extracted from video and that based on motion sensors (e.g., wearable sensors, smartphones, radio frequency (RF) sensors (Wi-Fi, RFID), LED light sensors, cameras [2]. For HAR based on motion sensors, the raw data obtained translates into signal patterns that represent the movement of the sensed area of the body.

The application of AI for HAR has revolutionized the way researchers segment and identify information extracted from wearable motion sensors. Machine Learning techniques first were used to extract features, and later to apply classifiers and obtain predictions (i.e., Support Vector Machines, k-Nearest Neighbor, and Artificial Neural Networks) [3]. However, the use of Deep Learning is recommended for feature extraction due to the power that these networks offer [4]. In Deep Learning, the raw data is managed as *time series sequences*. Hence, sliding windows must be created to organize this raw data [5]. The literature suggests using special networks to work with this type of data: these include Deep Neural Networks (DNN), Convolutional Neural Networks (CNN), and Recurrent Neural Networks (RNN). Lately, a hybrid model has been used that is based on the combination of two deep neural networks: CNN-LSTM (Long Short-Term Memory). This hybrid model allows for the achievement of high recognition scores in image processing and video-based HAR [6].

Overall, by using deep learning frameworks, it is possible to obtain high accuracy predictions and avoid the manual feature extraction process that occurs when applying Machine Learning [7]. In other words, this process overcomes the disadvantages of basic Machine learning algorithms, so that data scientists need not perform feature engineering manually. Given that, this study aimed to evaluate the performance of four deep learning networks (DNN, CNN, LSTM, CNN-LSTM) for HAR and, specifically, their ability to predict three different activities: walking, sit-to-stand, and squatting. The experimental data acquired includes 3D acceleration values from two different systems: a low-cost inertial measurement system composed of two IMUs, and an optoelectronic MOCAP system.

2. Materials and Methods

2.1. Experimental Acquisition:

The performance of the different networks was tested by evaluating the linear acceleration data of three daily movements: walking, sit-to-stand, and squatting. The experimental acquisitions were done in the Lisbon Biomechanics Laboratory at Instituto Superior Técnico using two motion capture systems: a low-cost inertial measurement system based on two MetaMotionR sensors (MBIENTLAB INC, San Francisco, CA., USA) and a high-end optoelectronic system composed of 14 infrared ProReflex 1000 cameras (Qualisys©, Göteborg, Sweden). The study was approved by the ethics committee of Instituto Superior Técnico in January 2020 (Ref. nr. 1/2020 (CE-IST)). All volunteers expressed their agreement to participate by signing an informed consent after a detailed explanation of the study objectives and experimental protocol.

The inertial measurement acquisition protocol consisted in the use of two wearable sensors placed on the left wrist and left ankle. Simultaneously, a 2D full body protocol, composed of 24 retro-reflective makers placed on anatomical landmarks as suggested by ISB [8] [9], was implemented to acquire the volunteer's motion with the optoelectronic system. Two more markers were placed over the IMUs to directly compare the neural network predictions from the two systems (see Fig. 1). Ten healthy volunteers (M:6, F:4;

Figure 1. Representation of the experimental protocol: a) Ankle sensor; b) Wrist Sensor

Age: 30.0±6.3 years; Height: 1.70±0.11 m.; Weight: 69.7±15.3 kg.) were selected based on the inclusion/exclusion criteria defined in the ethics committee proposal. The acquisitions with the two systems were done simultaneously using a sampling frequency of 100 Hz. The MOCAP data was processed using the Qualisys Tracking Manager software and the IMU data with the MetaBase App. For the gait analysis, volunteers were instructed to walk continuously for 60 seconds; this procedure was repeated three times. For the remaining movements, volunteers performed three series of ten repetitions each. All the trials were performed after a period of adaptation to the laboratory environment.

2.2. Data Preprocessing:

All acquisition files were processed using Python programming language (Google Colab). A unified dataset was created, including all the frames in a file. Next, each activity was labeled, following the labelling technique [10], *adding* the information corresponding to the activity to which the sample belongs. The raw data was processed using sliding windows [11]. To create sliding windows it was necessary to divide the input signal into segments of a certain time duration. For this study, a window of 100 samples in length was used. Six columns (x1, y1, z1, x2, y2, z2) and the final column named label, corresponding to the variable y, contained the three classes of expected activity (walking, sit-to-stand, squatting)."

IMU data consisted of 3D vector accelerations in csv format. As the dataset was obtained using two MBIENTLAB INC sensors, the acquisition control was managed via Low Energy Bluetooth through the MetaBase App and, thus, in some cases the start time for recording the samples differed. The raw data was manually synchronized to adjust for this discrepancy.

Because the MOCAP system does not directly provide acceleration values, the acquired trajectories were processed using in-house routines developed in MATLAB software (MathWorks©, Natick, USA). The trajectories were smoothed using a 2nd order Butterworth filter with a cutoff frequency of 6 Hz and were posteriorly splined using cubic polynomials. Finally, the linear accelerations were computed as the 2nd order derivative of the spline functions.

2.3. Choosing the artificial intelligence architecture:

Several studies have suggested that time series data should be processed with Deep Learning networks [5,12,13] because these types of networks directly learn the mapping (between the inputs of time-series and class outputs) with the feature engineering technique [14]. First, a DNN architecture was implemented with fully connected layers. Second, CNN was used to treat the data as a times series and build them into 1 dimension temporal convolution (1D-CNN) to capture dependences among input data [13]. This type of architecture is recommended to learn detailed feature representations and patterns from images [15]. Third, a LSTM is the type of RNN that helps in training the model over lengthy sequences and in the retention of the memory from previous time steps of

input fed to the model [16]. Hence, this last architecture is the one most often recommended to treat time series data. However, four networks with different architectures were tested, including a new approach combining CNN and LSTM. The CNN-LSTM is a hybrid model that uses CNN layers for feature extraction on input data combined with LSTMs to support sequence prediction [6]. Thus, the CNN-LSTM model reads subsequences of the main sequence as blocks: CNN extract features from each block, and then allows the LSTM to interpret the features extracted from each block. For this, a TimeDistributed wrapper was used that allows reuse of the CNN model, once per each subsequence. The CNN output serves as input to the LSTM, which provides the final prediction.

Two metrics were chosen to measure the classification performance of each algorithm: accuracy and F1 score. Accuracy is the ratio of the number of correct predictions to the total number of input samples [17]. The F1 score combines two measures defined in terms of the total number of correctly recognized samples, which are known in the information retrieval community as precision and recall. Precision is defined as $\frac{TP}{TP+FP}$, and recall corresponds to $\frac{TP}{TP+FN}$, where TP, FP are the number of true and false positives, respectively, and FN corresponds to the number of false negatives [18]. Class imbalance is countered by weighting classes according to their sample proportion:

$$F1 = \sum_i 2 * w_i \frac{precision_i.recall_i}{precision_i+recall_i}$$

where i is the class index and $w_{i=} n_i/N$ is the proportion of samples of class i, with n_i being the number of samples of the i-th class and N being the total number of samples.

3. Results

3.1. Deep Learning Networks for IMU and MOCAP:

The IMU-system database contained 304,135 samples and the MOCAP 315,334 samples. To train the algorithm, data from 8 volunteers was used and split in a relation of 80% for training and 20% for testing. Once trained, the algorithm was validated using data from an additional 2 volunteers. Early-stopping was added in the training phase to avoid overfitting.

Although the number of epochs and batch normalization were not the same for each network type, the remaining configuration used the same parameter settings for the two types of data. First, the basic DNN architecture was settled with the following parameters: 4 Dense fully connected layers with a dimension of 32 neurons, each built alongside a flatten layer. This model was trained with 50 epochs using a batch size of 64. The number of network parameters was 11,939. Also, the loss function used was sparse categorical crossentropy, which is commonly used for categorical problems. The activation function was ReLU, and the optimizer function was Adam. Using this configuration, an accuracy of 0.999 was obtained for both acquisition systems (Figure 2, DNN IMU vs. DNN MOCAP). An F1 score of 91.856% was obtained for IMU and 87.198% for MOCAP. Second, a CNN architecture was created using a sequential model of one Convolutional 1D layer with 32 filters, using a kernel of size 3 and ReLU activation. Besides a dropout of 0.5, a Max pooling 1D of size 2 and a flatten layer were used. In the end, a dense 8 units layer with ReLU activation function and another dense layer using the number of classes (in this case 3 classes) using as activation SoftMax were used for obtaining the prediction. This network was trained for 20 epochs using a

batch size of 32. Also, as in DNN, the loss function used was sparse categorical crossentropy. Using this configuration, an accuracy of 0.913 for IMU and 0.989 for MOCAP was obtained in the prediction as shown in Figure 2 (CNN IMU vs CNN MOCAP). F1 score values of 91.352% and 86.303%, respectively, were obtained. Third, an LSTM architecture was created using a sequential model of one LSTM layer with 100

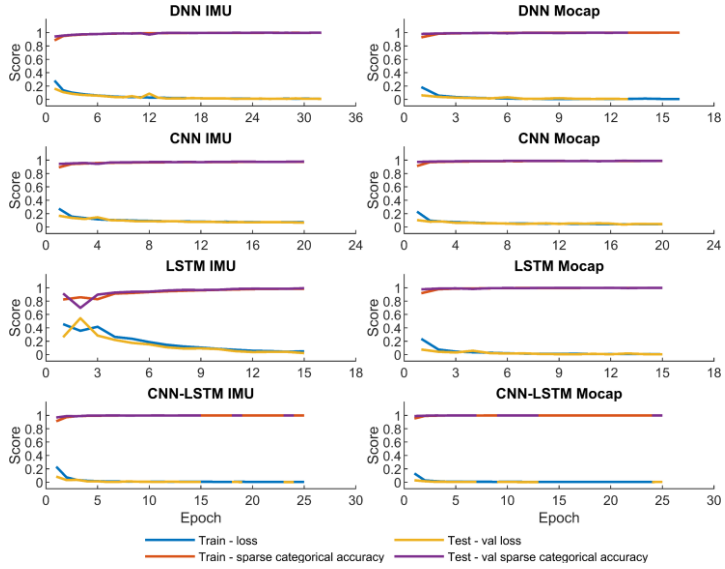

Figure 2. Accuracy and Loss metrics for the four Deep Learning networks of IMU and MOCAP data.

neurons. In addition, it featured a dropout of 0.5 and two dense layers: the first one with 100 and activation function ReLU, and the last one with the number of classes and a SoftMax activation function. The number of network parameters was higher, totaling 53,203. This network was trained for 15 epochs, using a batch size of 64. Also, as in the DNN and CNN architecture, the loss function used was sparse categorical crossentropy. Using this configuration resulted in a prediction accuracy of 0.993 (Figure 2: LSTM IMU vs LSTM MOCAP), and F1 scores of 91.856% and 86.303%, respectively.

Finally, when creating the CNN-LSTM network, we highlight the necessity of treating the data, which is in windows of 100-time steps, into subsequences using 4 steps and length 25 to feed a sequential model followed by a 1D CNN wrapped into two Time Distributed layers. Also, as in the 3 models mentioned above, the loss function used was sparse categorical crossentropy. This configuration resulted in prediction accuracies of 0.993 and 0.993 (Figure 2: CNN-LSTM IMU vs CNN-LSTM MOCAP) and F1 scores of 99.97% and 88.63%, respectively.

4. Discussion

Although the first three classification networks returned acceptable accuracies (between 91 and 99%), the F1 score, due to its robustness in the face of class imbalance, is the sole parameter that proves that a network is able to classify well, [19].

According to the accuracy metric, all values, except for CNN, are around 99% (Table 1). This is most likely due to the early-stopping settings, although it is expected

that the accuracy percentage would also reach 99% if it were allowed to train for more epochs.

Table 1. Accuracy and F1 Score results

Performance Measurements Parameters	Acquisition System	DNN	CNN	LSTM	CNN-LSTM
Accuracy	IMU	0.999	0.913	0.993	0.999
	MOCAP	0.999	0.989	0.999	0.999
F1 Score	IMU	91.856	91.352	91.856	99.97
	MOCAP	87.198	86.303	87.830	88.629

The *F1 score* metric for IMU performance surpasses 90% for all four networks; however, for MOCAP all networks (DNN, CNN, LSTM, CNN-LSTM) do not perform as expected because the samples were not sufficiently balanced (Table 1). For example, there were fewer walking frames than sitting and squatting frames. Prediction was most powerful with the last network (CNN-LSTM) for both systems, achieving a 99.97% prediction score for IMU and 88.63% for MOCAP.

Comparing our results to the literature, Deep & Zheng [20] used a UCI_HAR dataset implementing 4 different LSTM and the hybrid framework CNN-LSTM. Their CNN-LSTM model achieved the highest accuracy value with respect to the resting frameworks (93.40%); in contrast, we achieved 99% accuracy for all 4 networks tested. On the other hand, due to the imbalanced condition of our dataset, the calculated F1 score proves that the CNN-LSTM model is the best at recognition: hybrid models present higher scores [21]. Overall, using low-cost inertial sensor data with a complex Deep Learning hybrid model (CNN-LSTM) results in a good accuracy that is equal in reliability to a higher-cost and more complex MOCAP data system more specialized in motion capture.

5. Conclusion

Using Deep Learning methods, it was possible to recognize human activities such as walking, sit-to-stand, and squatting using only two IMUs and two markers located in the same place on the human body. This was accomplished by using Deep Learning Networks and processing the raw data as time series through sliding windows to extract relevant features from the segmented sequences. Accurate results were achieved with sensor data from IMUs, and the results validated with high quality MOCAP data. Using deep learning networks (DNN, CNN, LSTM), accuracies of above 91% was achieved. It was proven that, using the hybrid CNN-LSTM model, predictions can be achieved with the same accuracy (99%) as with the MOCAP system. Further research on the implementation of automatic labelling of human activities is required.

Acknowledgements

We acknowledge the support provided by the BeHealSy Program of EIT Health that promoted the collaboration between the Universidad Politécnica de Madrid and the University of Lisbon. Additionally, this work was supported by Fundação para a Ciência e a Tecnologia (FCT) with references UIDB/50021/2020 (INESC-ID) and UIDB/50022/2020 (Lisbon Biomechanics Laboratory and IDMEC under LAETA). Also, the first author thanks the IFARHU-SENACYT Panama program for supporting her PhD

scholarship and the second author acknowledges scholarship support from the Fundación Carolina FC and the Universidad Tecnológica Centroamericana UNITEC.

References

[1] Wang J, Chen Y, Hao S, Peng X, Hu L. Deep learning for sensor-based activity recognition: A survey. Pattern Recognit Lett [Internet]. 2019;119que bue:3–11. Available from: https://doi.org/10.1016/j.patrec.2018.02.010

[2] Wan S, Qi L, Xu X, Tong C, Gu Z. Deep Learning Models for Real-time Human Activity Recognition with Smartphones. Mob Networks Appl. 2020;25(2):743–55.

[3] Kale H, Mandke P, Mahajan H, Deshpande V. Human Posture Recognition using Artificial Neural Networks. Proc 8th Int Adv Comput Conf IACC 2018. 2018;272–8.

[4] Wu Q, Wang F. Concatenate convolutional neural networks for non-intrusive load monitoring across complex background. Energies. 2019;12(8).

[5] Ortiz Laguna J, Olaya AG, Borrajo D. A dynamic sliding window approach for activity recognition. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics). 2011;6787 LNCS:219–30.

[6] Yao S, Hu S, Zhao Y, Zhang A, Abdelzaher T. DeepSense: A unified deep learning framework for time-series mobile sensing data processing. 26th Int World Wide Web Conf WWW 2017. 2017;351–60.

[7] Brownlee J. Deep Learning for Time Series Forecasting. Machine Learning Mastery. 2018.

[8] Wu G, Siegler S, Allard P, Kirtley C, Leardini A, Rosenbaum D, Whittle M, D'Lima DD, Cristofolini L, Witte H, Schmid O, Stokes I. ISB recommendation on definitions of joint coordinate system of various joints for the reporting of human joint motion—part I: ankle, hip, and spine. J Biomech [Internet]. 2002;35(4):543–8. Available from: https://www.sciencedirect.com/science/article/pii/S0021929001002226

[9] Wu G, van der Helm FCT, (DirkJan) Veeger HEJ, Makhsous M, Van Roy P, Anglin C, Nagels J, Karduna AR, McQuade K, Wang X, Werner FW, Buchholz B. ISB recommendation on definitions of joint coordinate systems of various joints for the reporting of human joint motion—Part II: shoulder, elbow, wrist and hand. J Biomech [Internet]. 2005;38(5):981–92. Available from: https://www.sciencedirect.com/science/article/pii/S002192900400301X

[10] Woodward K, Kanjo E, Oikonomou A, Chamberlain A. LabelSens: enabling real-time sensor data labelling at the point of collection using an artificial intelligence-based approach. Pers Ubiquitous Comput. 2020;24(5):709–22.

[11] Niemann F, Reining C, Rueda FM, Nair NR, Steffens JA, Fink GA, Hompel M Ten. Lara: Creating a dataset for human activity recognition in logistics using semantic attributes. Sensors (Switzerland). 2020;20(15):1–42.

[12] Xu C, Chai D, He J, Zhang X, Duan S. InnoHAR: A deep neural network for complex human activity recognition. IEEE Access. 2019;7:9893–902.

[13] Murad A, Pyun JY. Deep recurrent neural networks for human activity recognition. Sensors (Switzerland). 2017;17(11).

[14] Chao-Lung Yang Z-XC and C-YY. Sensor Classification Using Convolutional Neural Network by Encoding Multivariate Time Series as Two-Dimensional Colored Images. Sensors (Switzerland). 2019;(1).

[15] Gollapudi S, Gollapudi S. Deep Learning for Computer Vision. In: Learn Computer Vision Using OpenCV. 2019. p. 51–69.

[16] Goyal P, Pandey S, Jain K. Deep learning for natural language processing: Creating neural networks with Python [Internet]. 2018. 290 p. Available from: https://proquest-safaribooksonline-com.cyber.usask.ca/9781484236857

[17] Mishra A. Metrics to evalute your machine learnign algorithm [Internet]. 2018. Available from: https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234

[18] Ordóñez FJ, Roggen D. Deep convolutional and LSTM recurrent neural networks for multimodal wearable activity recognition. Sensors (Switzerland). 2016;16(1).

[19] Dehghani A, Sarbishei O, Glatard T, Shihab E. A quantitative comparison of overlapping and non-overlapping sliding windows for human activity recognition using inertial sensors. Sensors (Switzerland). 2019;19(22):10–2.

[20] Deep S, Zheng X. Hybrid Model Featuring CNN and LSTM Architecture for Human Activity Recognition on Smartphone Sensor Data. Proc - 2019 20th Int Conf Parallel Distrib Comput Appl Technol PDCAT 2019. 2019;259–64.

[21] Abbaspour S, Fotouhi F, Sedaghatbaf A, Fotouhi H, Vahabi M, Linden M. A comparative analysis of hybrid deep learning models for human activity recognition. Sensors (Switzerland). 2020;20(19):1–14.