2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

Real Time Human Activity Recognition Using Acceleration and First-Person Camera data*

Christos Androutsos, Nikolaos S. Tachos, *Member, IEEE*, Evanthia E. Tripoliti, *Senior Member*, *IEEE*, Ioannis Karatzanis, Dimitris Manousos, Manolis Tsiknakis, *Member, IEEE*, Dimitrios I. Fotiadis, *Fellow, IEEE*

***Abstract*— The aim of this work is to present an automated method, working in real time, for human activity recognition based on acceleration and first-person camera data. A Long-Short-Term-Memory (LSTM) model has been built for recognizing locomotive activities (i.e. walking, sitting, standing, going upstairs, going downstairs) from acceleration data, while a ResNet model is employed for the recognition of stationary activities (i.e. eating, reading, writing, watching TV working on PC). The outcomes of the two models are fused in order for the final decision, regarding the performed activity, to be made. For the training, testing and evaluation of the proposed models, a publicly available dataset and an "in-house" dataset are utilized. The overall accuracy of the proposed algorithmic pipeline reaches 87.8%.**

I. Introduction

Human activity recognition has gained the interest of researchers due to its frequent use in applications related to surveillance, home health monitoring, human-computer interaction *etc.*. Activities of daily living (ADLs) and instrumental activities of daily living (IADLs) are the two main groups of human activities [1]. The complexity and variety of daily activities lead the researchers to explore different sources of data for activity recognition. Acceleration and visual observations are the most common source of information used for activity recognition [1, 2].

Wearable accelerometers are often used, either alone or in combination with additional sensors such as gyroscopes, to classify ADLs and recognize fall events [2]. Although they present high performance in classifying activities with high motion magnitude, their performance is not acceptable in the recognition of activities with low motion magnitude due to the similarity presented in acceleration signals [2-6].

Visual information is exploited in order for the abovementioned "weakness" to be overcome. In this case a sequence of images depicting the human body is processed in order for the body posture and/or motion information to be extracted. Furthermore, data provided by first-person cameras have been exploited recently [7-9]. The processing of video frames, through object-oriented methods, leads to the recognition of stationary activities strongly related to the detected objects [10-14]. Finally, the analysis of motion flow information allows the identification of activities based on optical flow features [2, 8].

The complementarity of the abovementioned approaches has been examined by Zhan *et al.* [2], Possas *et al.* [7] and Song *et al.* [8, 9]. Zhan *et al.* [2] introduced an automatic activity recognition system, integrating both accelerometers and a first-person view camera in three steps: (i) video and acceleration feature extraction, (ii) classification, and (iii) structure prediction. More specifically, time and frequency domain features were extracted from 3-axis accelerometer raw data, while motion features were extracted using the Lucas-Kanade optical flow method [15], which estimates the motion of objects across a series of consecutive image frames. For the classification a two-level approach is followed, the local and structured. The local classification exploits features directly extracted by raw sensor data, while the structured classification depends on the graph structure. The local classification provides a time independent prediction on the contrary to the structured classification that takes into account temporal dependencies.

In the work presented by Song *et al.* [8] temporal trajectory-like features are extracted from sensor data and the Fisher Kernel framework is applied to fuse video and temporal enhanced sensor features. They evaluated their approach on a Multimodal Egocentric Activity dataset which includes egocentric videos and sensor data of 20 fine-grained and diverse activity categories [9]. Later, Song *et al.* [9] developed a multi-stream Convolutional Neural Network (CNN) to learn the spatial and temporal features from egocentric videos and a multi-stream Long Short-Term Memory (LSTM) architecture to learn the features from multiple sensor streams (accelerometer, gyroscope, etc.). The final prediction results were achieved through a two-level fusion technique and various pooling techniques.

Possas *et al.* [7] proposed a Reinforcement Learning framework that makes use of policy learning in order to

*Research supported by the See Far project (https://www.see-far.eu/) which has received funding from the European Union's Horizon 2020 research and innovation program under the grant agreement No 826429. This article reflects only the author's view. The Commission is not responsible for any use that may be made of the information it contains.

C. Androutsos, E.E. Tripoliti, N.S. Tachos are with the Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas (FORTH), GR 45110, Ioannina, Greece (e-mail: xristosandroutsos@hotmail.com, etripoliti@gmail.com, ntachos@gmail.com).

D. Manousos and I. Karatzanis are with the Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH), GR-700 13 Heraklion, Crete, Greece (e-mail: mandim@ics.forth.gr, karatzan@ics.forth.gr).

M. Tsiknakis is with the Institute of Computer Science, Foundation for Research and Technology Hellas (FORTH) and the Department of Electric and Computer Engineering, Hellenic Mediterranean University, GR-710 04 Heraklion, Crete, Greece (e-mail: tsiknaki@ics.forth.gr).

Dimitrios I. Fotiadis is with the Department of Biomedical Research, Institute of Molecular Biology and Biotechnology, Foundation for Research and Technology Hellas (FORTH), GR 45110, Ioannina, Greece and with the Unit of Medical Technology and Intelligent Information Systems, University of Ioannina, GR 45110, Ioannina, Greece (phone: +302651009006, fax: +302651005588, e-mail: fotiadis@uoi.gr).

balance two different activity predictors using data from motion and vision sensors. For predictions of sensor data, an LSTM network was employed, while for vision predictor a mix of CNNs and Recurrent Neural Networks (RNNs), called Long-term Recurrent Convolutional Networks (LRCN), is used. The CNN acts as a feature extractor, while the LSTM captures the temporal structure of the data.

In the current study, an automated method for human activity recognition is proposed. The method provides inference on the performed activity by fusing information from an Inertial Measurement Unit (IMU) and a first-person camera. An LSTM and a ResNet model are developed for the recognition of locomotive and stationary activities, respectively. The models were trained on a publicly available dataset and were evaluated in a dataset recorded in real-settings by integrating both IMU and a first-person view camera on 3D printed glasses. The subjects were asked to simply follow their normal ADLs.

II. Materials and Methods

A. The dataset

Two datasets are exploited. One publicly available dataset for the training and testing of the developed models and an "in-house" dataset for their validation. More specifically, the DataEgo publicly available dataset [7] contains a natural set of activities developed in a wide range of scenarios. Images from the camera are synchronized with readings from the accelerometer and gyroscope captured at 15fps and 15Hz, respectively. In total, the dataset contains approximately 4 hours of continuous activity, corresponding to 20 different activities. The recordings were performed in different conditions and by different subjects. The activities were captured using the Vuzix M300 Smart Glasses. The accelerometer and gyroscope were synchronized at 15Hz.

The dataset for the training of the high motion magnitude activities recognition model derives from the acquisition of data time series from the 9-axis IMU. More specifically, the acceleration and gyroscope vectors. Raw data are captured, while the subjects perform five daily physical activities such as "Downstairs", "Running", "Sitting", "Upstairs", "Walking". Ten sets of "Running", "Sitting" and "Walking" activities are performed. The duration of each activity is 2min. Furthermore, ten sets "Upstairs" and "Downstairs" are performed, where the duration of each activity is 1 min. The sample rate and the range for the 3-axis accelerometer are 50 Hz with and +/- 16 g, respectively, while the sample rate and the range for the 3-axis gyroscope are 50 Hz with and +/- 2000 dps. The total number of samples are 244.562 time series and the total duration of the sets are 1 hour and 20 minutes.

The dataset for the training of the low motion magnitude activities recognition model is the DataEgo. More specifically, only the videos from the indoor activities are utilized. This corresponds to a set of 6 activities such as "Reading", "Writing", "Working on PC", "Eating", "Watching TV", and an "Unknown" (includes every other activity). In total, 42 videos are used. Each video includes 5 min of recording and it contains a sequence of 4-6 activities. The camera is synchronized at 15 fps and the video resolution is 640X360. The total number of frames and the total time are 31.260 and 3 hours and 50 minutes, respectively.

B. Hardware components

For the acquisition of the "in-house" dataset a 3D printed glasses frame, where a 9-axis IMU sensor (MetaMotion R) and a first-person camera (FPC) (Intel RealSense D435i) are mounted, is utilized. IMU and FPC communicate with an EDGE computing processing unit (LattePanda A864S) through a BLE and USB protocol, respectively. The processing unit is powered by a portable power bank of 30000 mAh in order for the whole hardware components to compose a wearable system.

C. The proposed methodology

The basic steps of the proposed methodology are depicted in Fig. 1: i) High motion magnitude activities recognition, and ii) Low motion magnitude activities recognition.

The IMU sensor is first activated providing input to the high motion magnitude activity model. In case "Sitting" or "Standing" activity is predicted, then the low motion magnitude activity model is activated and one of the activities of "Eating", "Reading", "Writing", "Working on PC", "Unknown" or "Watching TV" is recognized. If the outcome of the low motion magnitude activity model is "Unknown", then the activity recognition process is controlled by the high motion magnitude activity model in order one of the activities of "Walking", "Running", "Upstairs", "Sitting" or "Downstairs" to be detected.

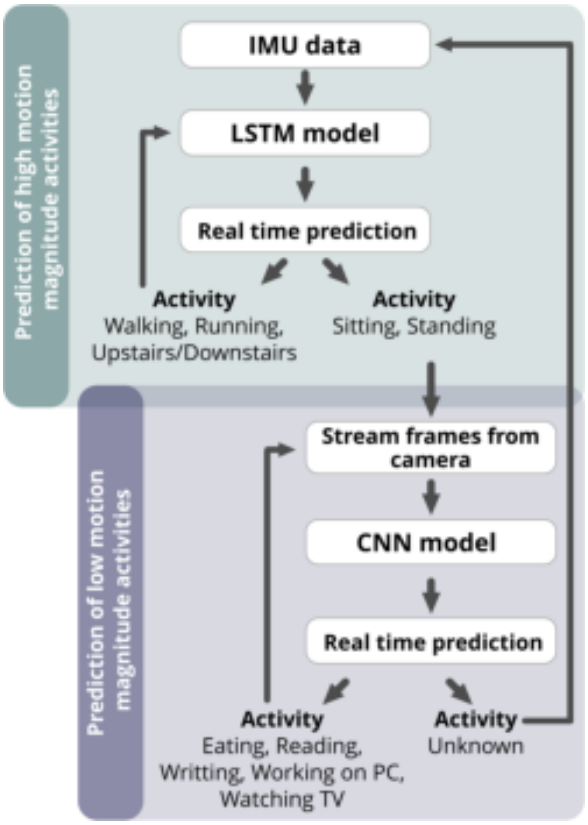

Figure 1: Proposed activity recognition model.

High motion magnitude activities model

For high motion magnitude activities recognition an LSTM model is developed [16]. The LSTM architecture consists of 2 layers with 128 hidden units each. The activation function that has been used is the Rectified Linear Unit (ReLU) [16]. It

takes as input batches of time series with a batch size of 500. Every batch consists of 200 time points and 6 columns, where three columns correspond to the three axes of the accelerometer and the rest to the gyroscope. The number of epochs used for the training phase is 300. The model provides a vector with the probabilities of the classes as output. The LSTM model has been optimized using the Adam algorithm [17] with learning rate 0.0025.

Low motion magnitude activities model

For low motion magnitude activities recognition, a pre-trained ResNet50 model is used [18]. It is a CNN trained on a million images from the ImageNet database and it can classify images into 1000 object categories. The architecture of the ResNet50 includes convolution layers, max pooling layers and a fully connected layer. The ResNet50 consists of 48 convolution layers along with one MaxPool and one Average Pool layer. The pre-trained model involves 5 stages each one having a convolution and an identify block. Each convolution and each identify block has 3 convolution layers. The ResNet50 has over 23 million trainable parameters. The pre-trained model acts as a feature extractor for the images.

The ResNet50 as described above, without the fully connected layer, was given as input to a new model. The new model consists of 2 layers. The first layer contains 512 hidden units and the activation function is the ReLU, while the second layer consists of 6 units, each corresponding to a target class, with the softmax activation function [19]. Average Pooling [19] and Flattening methods [19] were utilized to transform the frames into an acceptable form for the first layer. Additionally, dropout was applied between the two hidden layers. A dropout rate of 0.5 was used, as a weight constraint on those layers. The new model receives a sequence of frames with a batch size of 32 and it outputs a vector with the probabilities of the classes. Only the layers of the new model were trained, while the layers of the pre-trained model were stacked. The number of epochs during the training phase was 50. The optimization algorithm that has been used is the Stochastic Gradient Descent (SGD) [19] with learning rate 0.0001.

Data augmentation has been performed to the training data (frames) with the Keras ImageDataGenerator class [20]. This class accepts a batch of images and it applies a series of random transformations to each image in the batch (rotation range, zoom range, width shift range, height shift range, shear range, horizontal flip). The augmentation has been performed to the training and not to the validation data. The rotation range was implemented with value 30, zoom range with value 0.15, width and height shift range with value 0.2 and shear range with value 0.15.

III. Results

The two models were trained and tested in the datasets described in Section II.A, while they were evaluated on a validation set recorded under real time conditions. For the training, 80% of the dataset is utilized, while the rest 20% is used for the testing.

More specifically, using the wearable system described in Section II.B, data time series corresponding to a set of 5 high motor intensity physical activities and a set of 6 low motor

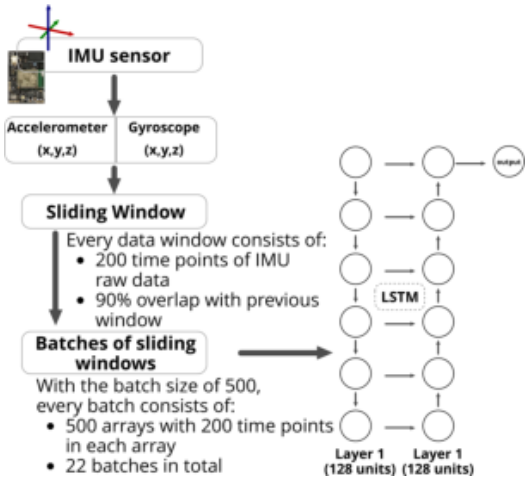

Figure 2: Hign motion magnitude activity recognition model.

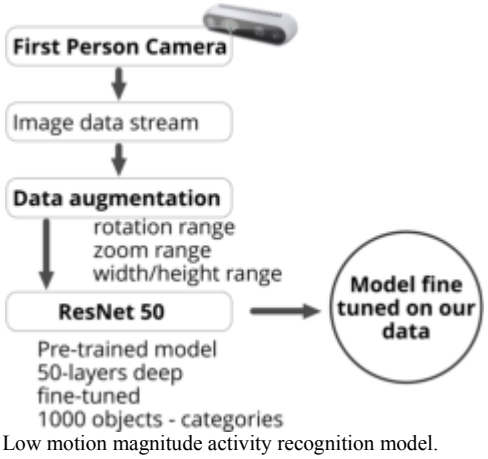

Figure 3: Low motion magnitude activity recognition model.

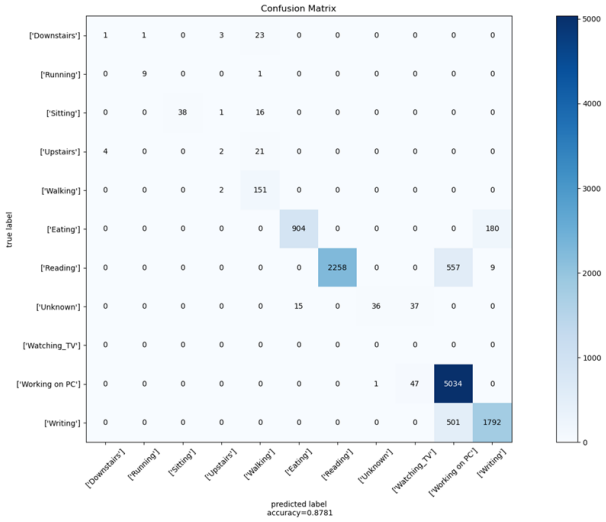

Figure 4: Low motion magnitude activity recognition model.

intensity indoor activities are recorded from 5 healthy volunteers. Each individual performed the above-mentioned set of activities in a 20 minutes session resulting in a total time recording of 1 hour and 40 minutes.

The accuracy of the proposed activity recognition model, in the validation dataset, is 87.8%. The confusion matrix is presented in Fig. 4. The accuracy of each model separately is 73.63% for the high motion magnitude activities recognition

model and 88.15% for the low motion magnitude activities recognition model.

IV. Discussion

The proposed activity recognition model fuses the outcomes of the high and low motion magnitude daily activities recognition models with an overall prediction accuracy 87.8% in real settings. The proposed algorithmic pipeline is validated utilizing a wearable system emulating the behavior of the See Far solution (https://www.see-far.eu/). The See Far solution includes smart glasses, based on augmented and machine learning technologies, providing functionalities for subjects with specific vision impairments based on the personalized profile of each subject. The proposed model is part of the See Far personalized profile.

Similar approaches have been presented in the literature by Zhan *et al.* [2], Possas *et al.* [7] and Song *et al.* [8, 9]. A comparison of the proposed approach with those works is presented in Table I in terms of the dataset utilized, the number of activities recognized and the accuracy achieved. The proposed pipeline exhibits better performance compared to Song *et al.* and Possas *et al.* work and similar performance to Zhan *et al.*

TABLE I. Comparison with the literature.

Authors	Dataset	No. of activities	Accuracy %
Zhan *et al. [2]*	In house dataset	12	90.38
Song *et al. [8]*	Multimodal Egocentric Activity dataset	20	80.50
Song *et al. [9]*			83.70
Possas *et al. [7]*	Multimodal Egocentric Activity dataset	20	84.84
	DataEgo	20	80.00
Proposed approach	In house dataset	11	87.80

V. Conclusion

In this work, an algorithmic pipeline has been proposed for human activity recognition. The latter is part of the personalized profile of the See Far solution and specifically it is integrated into the See Far augmented reality smart glasses which provide, in real time, services and recommendations for subjects suffering from specific vision impairments. The developed models utilize data streams from an IMU sensor and image frames from an FPC and are based on an LSTM architecture and a refined CNN model. The performance of the model exhibits 87.8% accuracy and provides the inference in real time on an edge wearable processing unit. In the near future, we will investigate the robustness of the model in more challenging activities specifically in the workplace of persons with vision impairment without compromising the real time performance.

References

[1] G. Vavoulas, C. Chatzaki, T. Malliotakis, M. Pediaditis, M. Tsiknakis, "The MobiAct Dataset: Recognition of Activities of Daily Living using Smartphones", *Proceedings of the International Conference on Information and Communication Technologies for Ageing Well and e-Health*, Rome, Italy, pp. 143-151, 2016.

[2] K. Zhan, S. Faux, F. Ramos, "Multi-scale Conditional Random Fields for First-Person Activity Recognition", Proceedings of the 2014 IEEE International Conference on Pervasive Computing and Communications (PerCom), Budapest, Hungary, 2014, pp. 51-59.

[3] D. Anguita, A. Ghio, L. Oneto, X. Parra, and J. Reyes-Ortiz, "Human Activity Recognition on Smartphones Using a Multiclass HardwareFriendly Support Vector Machine", ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2012, ch. 30, pp. 216–223.

[4] J. Ward, P. Lukowicz, G. Trster, and T. Starner, "Activity recognition of assembly tasks using body-worn microphones and accelerometers", IEEE transactions on pattern analysis and machine intelligence, pp. 1553–1567, 2006.

[5] M. Sousa, A. Techmer, A. Steinhage, C. Lauterbach, and P. Lukowicz, "Human tracking and identification using a sensitive floor and wearable accelerometers", Proceedings of the 11th IEEE International Conference on Pervasive Computing and Communications (PERCOM), 2013, pp. 166–171.

[6] J. R. Kwapisz, G. M. Weiss, S. A. Moore, "Activity Recognition using Cell Phone Accelerometers", *SIGKDD Explor. Newsl.*, 2010, pp. 74–82.

[7] R. Possas, S. P. Caceres, F. Ramos, "Egocentric Activity Recognition on a Budget", *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, 2018, pp. 5967-5976.

[8] S. Song, V. Chandrasekhar, B. Mandal, L. Li, J. Lim, G.S. Babu, P.P. San, N. Cheung, "Multimodal Multi-Stream Deep Learning for Egocentric Activity Recognition", *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Las Vegas, NV, USA, 2016, pp. 378-385.

[9] S. Song, N. Cheung, V. Chandrasekhar, B. Mandal and J. Liri, "Egocentric activity recognition with multimodal fisher vector", *Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Shanghai, China, 2016, pp. 2717-2721.

[10] P. J. Hsieh, Y. L. Lin, Y. H. Chen, and W. Hsu. "Egocentric activity recognition by leveraging multiple mid-level representations", *Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME)*, 2016, pp. 1–6.

[11] Y. Li, A. Fathi, J. M. Rehg, "Learning to predict gaze in egocentric video", *Proceedings of the 2013 IEEE International Conference on Computer Vision*, 2013, pp. 3216–3223.

[12] M. Ma, H. Fan and K. M. Kitani, "Going Deeper into First-Person Activity Recognition", *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016, pp. 1894-1903.

[13] K. Matsuo, K. Yamada, S. Ueno and S. Naito, "An Attention-Based Activity Recognition for Egocentric Video", *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, Columbus, OH, USA, 2014, pp. 565-570.

[14] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views", *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition*, Providence, RI, USA, 2012, pp. 2847-2854.

[15] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision (darpa)", *Proceedings of the 1981 DARPA Image Understanding Workshop*, April 1981, pp. 121– 130.

[16] D. Kent and S. Fathi, "Performance of three slim variants of the long short-term memory (LSTM) layer", *2019 IEEE 62nd International Midwest Symposium on Circuits and Systems (MWSCAS), IEEE*, Dallas, TX, USA, 2019, pp. 1-4.

[17] Z. Zhang, "Improved adam optimizer for deep neural networks", *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). IEEE*, Banff, AB, Canada, 2018, pp. 1-2.

[18] D. Theckedath and R. R. Sedamkar, "Detecting Affect States Using VGG16, ResNet50 and SE-ResNet50 Networks", *SN Computer Science 1.2* (2020), pp. 1-7.

[19] S. I. Saedi and H. Khosravi, "A deep neural network approach towards real-time on-branch fruit recognition for precision horticulture", *Expert Systems with Applications 159* (2020): 113594, pp. 4-10.

[20] A. Mikołajczyk and G. Michał, "Data augmentation for improving deep learning in image classification problem", *2018 international interdisciplinary PhD workshop (IIPhDW). IEEE*, Poland, 2018, pp. 117-122.