DOI: 10.15439/2019F50
ISSN 2300-5963 ACSIS, Vol. 18

A Deep Learning and Multimodal Ambient Sensing Framework for Human Activity Recognition

Ali Yachir, Abdenour Amamra, Badis Djamaa, Ali Zerrouki and Ahmed khierEddine Amour
Military Polytechnic School
PO BOX 17, Bordj-El-Bahri, 16111, Algiers, Algeria
Email: {ali.yachir, amamra.abdenour, badis.djamaa}@gmail.com

Abstract— **Human Activity Recognition (HAR) is an important area of research in ambient intelligence for various contexts such as ambient-assisted living. The existing HAR approaches are mostly based either on vision, mobile or wearable sensors. In this paper, we propose a hybrid approach for HAR by combining three types of sensing technologies, namely: smartphone accelerometer, RGB cameras and ambient sensors. Acceleration and video streams are analyzed using multiclass Support Vector Machine (SVM) and Convolutional Neural Networks, respectively. Such an analysis is improved with the ambient sensing data to assign semantics to human activities using description logic rules. For integration, we design and implement a Framework to address human activity recognition pipeline from the data collection phase until activity recognition and visualization. The various use cases and performance evaluations of the proposed approach show clearly its utility and efficiency in several everyday scenarios.**

I. INTRODUCTION

The combination of the Ambient Intelligence and the Internet of Things **[1]** aims at building smart environments by integrating a variety of interconnected devices such as camera, smartphone, smart watch and actuator. Such a sensing and actuating technology, has allowed to the analysis of human daily activities to become easier and straightforward. Particularly, in smartly controlled environments such as smart home, HAR can be envisioned for several potential applications and different contexts including security, healthcare, ambient assisted living and behavior analysis. For instance, many HAR systems surveyed in **[2,3],** where the authors focus on different activities (walking, running, cooking, exercising, etc.) in different application domains.

In practice, there are diverse ways of using sensors for human activity recognition in a smart environment. Hence, the existing approaches can be divided into two main categories, namely: vision-based and sensors-based approaches. In the former approaches, the primitive actions of an activity are detected by analyzing the images transmitted by an RGB camera. Such an analysis can exploit computer vision techniques to recognize patterns. Whereas the latter approaches (sensor-based) use sensors that are either worn by a person or placed on everyday objects. Wearable sensors can be placed on clothing, in a pocket, or stuck directly to the body (wrist, hip or torso) to provide valuable information about an individual's degree of functional ability and lifestyle **[4]**. Indeed, sensors' position should be well chosen in order to ensure their usability while offering a maximum comfort to the user. In addition, sensors can be placed seamlessly on ordinary objects to detect and control the environment. They can also be of different types such as: contact detectors to give the state (close/open) of doors and cabinets, pressure mats to indicate the position of the person in the room or to detect if a person is sitting on a sofa or laying on a bed, RFID tags to give the location of objects, etc. According to a recent study **[5]**, the RGB cameras have lower popularity when compared to depth sensors and wearable devices in HAR research.

In order to implement a HAR system, the data collected and transmitted by various cameras and sensors disseminated in the smart environment can be analyzed using several techniques in either vision or sensors-based approaches. Regarding vision-based approaches, a survey of action recognition approaches based on Space-Time Interest Points (STIP) was proposed in **[6]**. Most recent approaches are based on Convolutional Neural Networks (CNNs) including Deep Convolutional Networks (ConvNets) **[7]** and TwoStream **[8]**. These deep learning methods aim to learn automatically the semantic representation of raw videos by using a deep neural network in a discriminatory manner from a large number of tagged data. For analyzing real-time videos, Recurrent Neural Networks (RNN) among which there are Long Short-Term Memory (LSTM) units have been proposed. LSTM networks have proved their effectiveness in several areas such as: images and videos subtitling **[9]** and temporal information of movements and videos streams. Regarding sensors-based approaches, a deep ConvNet was also used in **[10]** to perform HAR using smartphone sensors by exploiting the inherent characteristics of activities. In **[11]**, acceleration streamed by a smartphone are analyzed with K-Nearest Neighbors (KNN) for recognizing several types of activities (walking, climbing, sitting, standing and falling down). In **[12]**, data from inertial and pressure sensors placed on the trunk of a patient are used to recognize activities such as walking, sleeping and climbing stairs. In **[13]**, Hidden Markov Model (HMM) is used to classify complex actions such as running, walking or laying, using the accelerometer data of a wristwatch. In **[14]**, simple and complex activities such as cleaning, hand washing, and plant watering are recognized using fixed window lengths with an overlapping halved window. In **[15]**, human activity recognition is analyzed through the segmentation of the multidimensional time series of acceleration data based on a specific multiple regression model. In **[16]**, a digital low-pass filter is designed to recognize certain types of human physical activities using acceleration data. In **[17]**, the selection and placement of wearable sensors is investigated for classifying sixteen activities of daily living for six healthy subjects.

The aforementioned discussed works show that most of the proposed approaches recognize simple human activities such as laying, sitting, and standing. Moreover, these approaches focus on the data received from either cameras or other sensors without a real combination of the different modalities that can become unavailable due to their temporary or permanent disappearance, and should therefore, be replaced to ensure HAR continuity. Furthermore, contextual information such as localization, acceleration and object state provided by mobile or wearable sensors combined with machine learning methods offer a higher accuracy and diversity for recognizing complex human activities (watching TV, cooking, exercising, etc.).

We propose a hybrid approach for HAR in an ambient environment by combining three types of sensing technologies, namely: smartphone accelerometer, RGB cameras and ambient sensors. First, real time accelerations and video streams are analyzed separately using machine learning algorithms to detect and recognize simple human activities or postures. Video streams are used by default for indoor spaces, but they are replaced by smartphone accelerometer data in the case of inaccessible cameras. Switching between these two modes can considerably increase the reliability of the designed HAR system. Second, additional information is extracted from the available activated ambient sensors to assign semantics to human activity using Description Logic (DL) rules. Finally, the three types of the provided information are combined inside a HAR framework using supervised machine learning algorithms in order to recognize and visualize more complex activities.

The remainder of the paper is organized as follows. In Section 2, we describe our acceleration-based activity recognition method. In Section 3, the video-based activity recognition method is explained. In Section 4, we present the hybrid approach and the designed framework for complex activity recognition. In section 5, we conduct several validation scenarios for the recognition of everyday activities. The paper is concluded with section 6, and potential future works are announced.

II. Acceleration based activity recognition

The acceleration data along the three axes (x, y, and z) is collected from a smartphone worn on the waistband of the user's pelvis. This data collection operation is performed by an Android application with a sampling rate of 50 Hz, i.e. the data is divided into a window of 50 records per second. We distinguished six (6) classes of elementary actions or postures namely: sitting, standing, running, walking, walking upstairs and walking downstairs. We collected 500 records for each class. For a better distinction between the different classes, we chose, as an input to our machine learning model, a vector of 30 characteristics such as average, variance, and min-max with respect to x, y and z; resulting average of the acceleration; AR-coefficient; Angle Tilt; and Signal Magnitude Area (SMA) **[18]**. We opted for this choice after performing several tests by combining these different characteristics. For each combination, we calculate classification success rate. These characteristics ensure a high degree of independence between the different classes and minimize the correlation between them. When constructing the learning model, we tested six (06) learning algorithms, which were Naïve Bayes, SVM with linear kernel, SVM with rbf kernel, nonlinear SVM, k-Nearest Neighbors (kNN) and MultiLayer Perceptron (MLP) with a single hidden layer. The success rates obtained from these algorithms are shown in the **Table I**. The latter shows that SVM with linear kernel gives the best performance (93%).

TABLE I. SUCCESS RATE OF THE TESTED ALGORITHMS

Approach	**Success rate**
Naïve Bayes	92%
SVM with linear kernel	**93**%
SVM with rbf kernel	90%
Nonlinear SVM	92%
k-Nearest Neighbors (kNN)	87%
MultiLayer Perceptron (MLP)	87%

III. Vision based activity recognition

A. *Dataset construction*

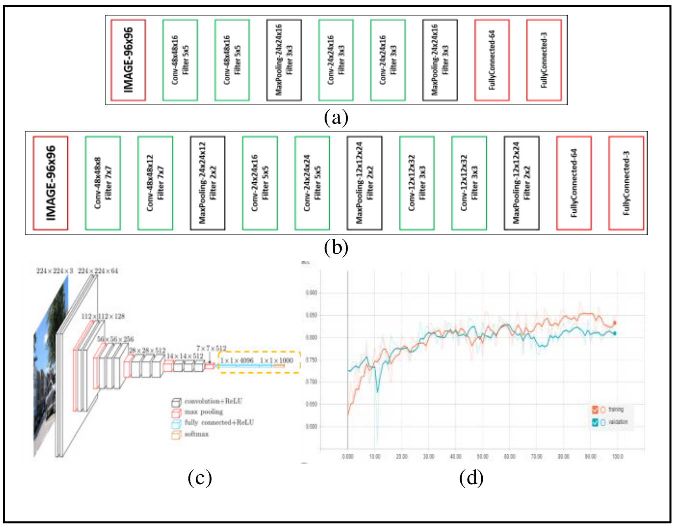
The dataset is constructed using two different sources: Multiple Pose Human Body Database (LSP / MPII-MPHB) **[19]** and other data that we gathered from Google Image search engine. First, the LSP / MPII-MPHB contains 26675 images and 29732 human bodies that are divided into six (06) action categories: curving, knee, laying, occlusion, sitting and standing. For each image, we detect the persons using the Single Shot MultiBox Detector (SSD) method **[20]**, which is a unified Framework for detecting objects with a single neural network. We focus only on three main postures: standing, sitting and laying. Then, we used Google Image search facility to retrieve all possible images for these three main postures.

After building the dataset, we obtain a considerable number of images for each action or posture category. The images represent the inputs of the learning model while the actions represent the classes respective to the outputs. Thus, in order to standardize and reduce the size of the training data and to speed up their processing, we perform a pre-processing on all the input images. First, each image is stored with a reduced size of 224×224 pixels and three channels representing the values of the three colors: red, green and blue. Then, the images are normalized and scaled at loading time by a centering and reduction transformation in the interval [-1,1]. Finally, we split the images into three folders: sitting, standing and laying down to assign them to the different classes. In order to feed our learning model, we transformed the set of images into a four-dimensional (height 224 pixels, width 224 pixels and the three-color channels) tensor and the class labels into a one-dimensional vector.

B. *Learning model*

We proposed two architectures of convolutional neural networks, where each is composed of several layers of different types. We also tested a learning method based on the transfer of parameters. These relatively deep architectures have a structure inspired by that of the VGGNet network **[20].** The difference between them lies in the depth and the number of hyperparameters. Both architectures are composed of large blocks; where the initial blocks are constituted of two convolutional layers followed by a pooling layer, and the last ones are composed of only dense layers. The depth of the convolutional layers increases from one block to another, although the spatial size of the filters decreases. The choice of convolutional layers comes from the fact that they are the most adapted to image recognition tasks as they consider the multidimensional aspect of images. Moreover, each neuron in these layers is connected to just a small set of neurons in the preceding layer, the number of learnable parameters is therefore smaller due to the parameter sharing property of convolution. The *z* outputs of each convolutional layer are filtered by the Rectified Linear Unit (ReLU) activation function.

Fig. 1. Our activity recognition architectures, (a), Architecture 1, (b) Architecture 2, (c) Architecture 3, (d) Results of our activity recognition CNN



A maxpooling layer follows each pair of convolutional layers. These layers lead to the reduction of the dimensions of the feature maps by applying the *max* function on a window of neighboring pixels at a given region of the image. Therefore, the maxpooling reduces the intraclass variance by discarding the unnecessary information. At the last level of the network, a dense layer is considered in order to gather all the features detected throughout the network. The output layer is also a dense layer whose number of neurons is equal to the number of classes (in our case, the number of classes is set to 3). The z results are passed through a *softmax* function in order to be squashed into the interval [0,1] leading to a probability distribution over the classes. The reason why the stacked architecture is preferable is that the first layers detect low-level features (such as edges and simple shapes) whereas the deeper ones detect high-level features (such as complex shapes and objects).

In what follows, we present the three architectures that we proposed and tested in the light of this contribution.

C. *Architecture 1*

The first block of the network consists of two convolutional layers having each a depth of 16 feature maps and a filter size of 5 × 5, with stride 1 horizontally and vertically, and zero padding on all four edges. These layers are followed by a maxpooling layer with a filter of size 2 × 2 and a step of 2, meaning that we reduce the dimensions of the input by a half. The second block is identical to the first one except that the depth of each convolutional layer is 32 instead of 16 and the size of the convolution filter is 3 × 3.

The third block, is a dense layer of 64 neurons connected to all the outputs from the previous maxpooling layer. The output layer is of size 3, where each neuron corresponds to a class. **Figure 1 (a)** illustrates Architecture 1. The green rectangles represent the convolutional layers, the blacks represent the pooling layers, and the reds represent the dense layers.

D. *Architecture 2*

This architecture is similar to the first one but it is composed of four main blocks. The first block encompasses two convolutional layers with a filter of 7 × 7 and a depth of 8 and 12, respectively. The second contains similar layers whose filter size is 5 × 5 and depths is 16 and 24, respectively. The two layers of the last block have a 3 × 3 filter and a depth of 32 each. Similarly, a pooling layer follows each pair of convolutional layers in all the blocks. Identically to the previous architecture, this one contains a dense layer of 64 neurons in the last block (see **Figure 1 (b)**).

E. *Architecture 3*

The first two architectures, that we proposed, are prone to overfitting because the training dataset is small compared to the size of the network. In order to avoid such a drawback, we used a Transfer Learning approach and we augmented our dataset by applying several image transformations. We initialized the new model with the weights of VGG16 network trained on our dataset. During the training, we maintained a fixed number of layers (the first convolutional layers) and we optimized the parameters of the latter ones. Our purpose is to reduce optimization space dimension, whilst reusing the first convolutional blocks. The latter are most likely to remain similar regardless of the problem at hand, and the model must be able to fit specific top-level layers. We adapted the dense layer with our classes of actions. **Figure 1 (c)** shows the modifications made to the VGG16 architecture for transfer learning. Our best results for activity recognition were obtained with the third architecture (**Figure 1 (c)**).

F. *Cost function and the optimizer*

The neural networks are general functions estimators. Nevertheless, their estimation is never perfect as there is always a discrepancy between the output of the network and the ground truth values. Therefore, we define a cost function to measure the error and we adapt an optimization algorithm to minimize its value. Given the transformation of this problem into a classification task, the most appropriate interclass error measure is the cross-entropy function **[21].** The learning is achieved by minimizing the norm of the cross-entropy with gradient descent algorithm. Such an algorithm consists in changing the weights of the network by a factor of a fixed learning rate. We chose the *Adadelta* **[22]** optimizer as it does not depend.

G. *Training platform and implementation*

One of the major problems in deep learning is the significant requirement for computation resources during the training. In order to alleviate this problem, we trained our algorithm on a machine equipped with a Nvidia GTX 860 GPU and 12Gb of RAM. For the purpose of robustness and versatility, we implemented our CNN it in Python3 using TensorFlow library. This library runs a low-level C++ routine, which are able to perform massively parallel computations using all the available processing power and benefiting from the existing hardware vectorization of the GPU. Training and validation progress can be seen in **Figure 1 (d)**

IV. HYBRID APPROACH FOR HUMAN ACTIVITY RECOGNITION

The two previously presented models, based on accelerometer and camera, are able to recognize elementary actions or postures of the person (e.g., standing, sitting and laying) while an activity is defined as a task of daily life that the person performs over a given time interval. Thus, other contextual information can complement the information of the posture to deduce the effective activity. The localization of the subject, for example, is an essential attribute, due to the fact that most activities are carried out in a specific room. For instance, the "cooking" activity takes place in the kitchen while "Watching TV" is more likely to appear in the living room. In addition, the posture of the person is a influential factor to recognize an activity. We cannot imagine, for example, a person sleeping in a "standing" posture. It should also be noted that the actions related to a particular activity can activate or deactivate a number of ambient sensors. In order to represent all these causal relationships, we used DL rules to infer five main activities namely: watching TV, sleeping, preparing a meal, communicating (talking on the phone) and working with a laptop or PC. For example, the activity "Sleeping" can be inferred using the two DL rules shown in **Table II**.

TABLE II. EXAMPLE OF DL RULES FOR "SLEEPING" ACTIVITY

Sleeping ≡ Posture (laying_down) ∩ Location (bedroom) ∩ AmbientSensors.(Bed).HasStatus(On) ∩ AmbientSensors.(Light).HasStatus(Off)
Sleeping ≡ Posture (laying_down) ∩ Location (living_room) ∩ AmbientSensors.(Sofa).HasStatus(On) ∩ AmbientSensors.(Light).HasStatus(Off) ∩ AmbientSensors.(TV).HasStatus(Off)

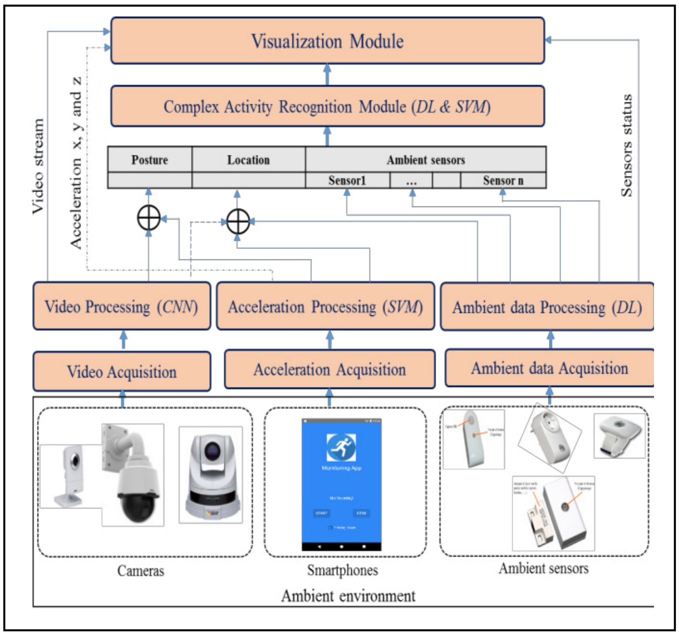
Since we can have several rules for the same activity, we transformed these rules to a description vector containing posture, location and the state of the other sensors as shown in **Table III**. Hence, all the data contained in description vectors summarize a temporal window. Such vectors are used as an input for the SVM algorithm with linear kernel and their corresponding activities, in DL rules, represent the output. The supervised learning model is constructed by generating a set of input vectors with the identifier of the corresponding class or activity. These vectors contain the different possible combinations of values regarding the considered characteristics. For example, in "Watching TV" activity, the generated posture can be "sitting" or "laying down". If a feature vector does not match any of the existing classes, the person's activity is considered as "Unknown".

TABLE III. STRUCTURE OF THE INPUT LEARNING VECTOR

Posture	Location	Ambient sensors			
		Sensor1	...		Sensor n

We developed a framework for human activity recognition in ambient environment. As shown in **Figure 2**, the general architecture of the proposed framework allows data acquisition and processing as well as activity recognition and visualization. First, heterogeneous data are collected using different sensing modalities, namely: cameras for real-time video streams, the various sensors of the Smartphone (accelerometer, gyroscope, etc.) as well as the others ambient environment sensors (pressure, temperature, humidity, light, movement, etc.). Then, collected data are processed separately using CNN, SVM and DL respectively for video, acceleration and ambient sensors. We should notice that the posture is recognized, by default, using video stream, but automatic switching to acceleration is performed in case of cameras unavailability to ensure service continuity. The location can be also deduced from cameras and smartphones' position or provided by other sensors. Finally, a description vector is constructed from all the previous processed data and used as input of the machine learning model to recognize more complex activities. The recognized activity along with the collected data from camera, smartphone and ambient sensors are transmitted to the visualization module for display and monitoring in real time.

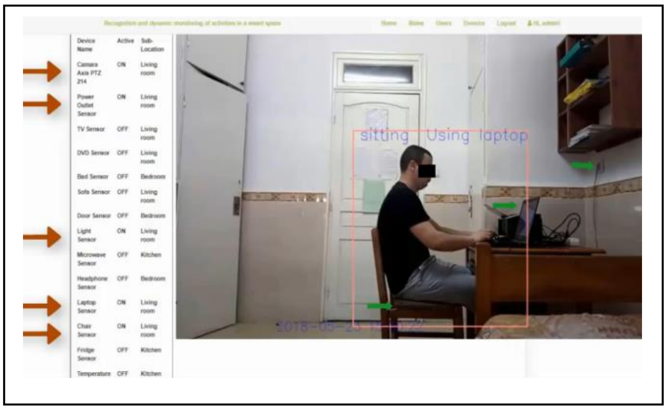
Fig. 2. Global architecture of the designed HAR Framework



V. IMPLEMENTATION AND USE CASES SCENARIOS

The proposed Framework has been applied in several situations for human activity recognition. For example, **Figure 3** shows a person using a personal computer. This activity is recognized with the combination of several information: the posture is "sitting", the location is "office", the pressure sensor on the chair is activated "ON" and the laptop is "ON".

Fig. 3. Example of recognized activity "Working with laptop"



Recognition and dynamic monitoring of activities in a smart space
Home
Roles
Users
Devices
Logout
Hi, admin
sitting
Using laptop

Device Name	Active	Sub-Location
Camara Axis PTZ 214	ON	Living room
Power Outlet Sensor	ON	Living room
TV Sensor	OFF	Living room
DVD Sensor	OFF	Living room
Bed Sensor	OFF	Bedroom
Sofa Sensor	OFF	Living room
Door Sensor	OFF	Bedroom
Light Sensor	ON	Living room
Microwave Sensor	OFF	Kitchen
Headphone Sensor	OFF	Bedroom
Laptop Sensor	ON	Living room
Chair Sensor	ON	Living room
Fridge Sensor	OFF	Kitchen
Temperature	OFF	Kitchen

Regarding the performance evaluation, we achieved a success rate of 97% after several classification tests of the proposed HAR approach of the designed framework. As shown in the **Table IV**, our approach gives a better result considering five classes of activities and several multi-source data (ambient sensors, cameras and acceleration). The result of 98.2% obtained in **[12]** is due to the fact that the authors used the acceleration by considering only three classes of activities which further minimizes the ambiguity between classes.

TABLE IV. COMPARISON OF SUCCESS RATES OF ACTIVITY RECOGNITION

Approach	Devices	Classifier	Success rate
[12]	Inertial and barometric sensor	KNN	98.2%
[13]	Accelerometer of a wristwatch	HMM, CRF	90.4%
[15]	Accelerometer on Chest, Thigh and Ankle	Multiple regression model	90.3%
[16]	Accelerometer on Hand and pocket	digital low-pass filter	91.15%
[17]	Pressure sensor	KNN	89.08%
Our approach	Cameras, smartphone, ambient sensors	CNN, DL and SVM with linear kernel	97%

VI. CONCLUSION AND FUTURE WORK

We proposed a hybrid solution for human activity recognition using smartphone inertial sensors (accelerometers), RGB cameras and ambient sensors (pressure, localization, etc.). Acceleration data and videos were analyzed using machine learning algorithms, SVM and CNN in order to detect the current potential posture of the person. Such an analysis is augmented with ambient sensor data to assign semantics to the human activity based on description logic rules. A HAR framework is also designed to build the whole pipeline from data collection until activity recognition and visualization. We are currently working to use our approach in the context of ambient-assisted living to assist elderly or dependent persons to improve their quality-of-life. Future works will include further experimentation and combination of other techniques such as automatic image captioning using deep learning.

REFERENCES

[1] P. Suresh, J. V. Daniel, V. Parthasarathy, and R. H. Aswathy, "*A state of the art review on the Internet of Things (IoT) history, technology and fields of deployment,*" in Proc. IEEE Int. Conf. Sci., Eng. Manage. Res. (ICSEMR'14), Chennai, India, Nov. 2014, pp. 1-8.

[2] O. Lara, M. Labrador, "*A survey on human activity recognition using wearable sensors*", IEEE Commun. Surv. Tutor. 1 (2012) 1–18.

[3] H.F. Nweke, Y.W. Teh, M.A. Al-garadi, and U.R. Alo, "*Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges*" Expert Systems With Applications 105 (2018).

[4] M.M. Hassan, M. Zia Uddin, A. Mohamed, and A. Almogren: "*A robust human activity recognition system using smartphone sensors and deep learning*" in Future Generation Computer Systems 81 (2018) 307–313.

[5] O.C. Ann, L.B. Theng, "*Human activity recognition: A review*" in 2014 IEEE International Conference on Control System, Computing and Engineering.

[6] D.D. Dawn and S.H. Shaikh, "*A comprehensive survey of human action recognition with spatio-temporal interest point (stip) detector*," The Visual Computer, vol. 32, no. 3, pp. 289-306, 2016.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, "*Largescale video classification with convolutional neural networks*," in Proceedings of the IEEE conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732.

[8] K. Simonyan and A. Zisserman, "*Two-stream convolutional networks for action recognition in videos,*" in Advances in neural information processing systems, 2014, pp. 568-576.

[9] S. Yousfi, "*Embedded Arabic text detection and recognition in videos*," Ph.D. dissertation, Lyon University, 2016.

[10] C.A. Ronao, S-B. Cho "*Human activity recognition with smartphone sensors using deep learning neural networks*" Expert Systems With Applications 59 (2016).

[11] Y. Kwon, K. Kang, and C. Bae, "*Unsupervised learning for human activity recognition using smartphone sensors,*" Expert Systems with Applications, vol. 41, no. 14, pp. 6067-6074, 2014.

[12] F. Massé, R. R. Gonzenbach, A. Arami, A. Paraschiv-Ionescu, A. R. Luft, and K. Aminian, "*Improving activity recognition using a wearable barometric pressure sensor in mobility-impaired stroke patients,*" Journal of neuroengineering and rehabilitation, vol. 12, no. 1, p. 72, 2015.

[13] 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. 22 500-22 524, 2014.

[14] S. Dernbach, B. Das, N. C. Krishnan, B. L. Thomas, and D. J. Cook, "*Simple and complex activity recognition through smart phones*," in Intelligent Environments (IE), 2012, pp. 214-221.

[15] F. Chamroukhi, S. Mohammed, D. Trabelsi, L. Oukhellou, and Y. Amirat, "*Joint segmentation of multivariate time series with hidden process regression for human activity recognition*," Neurocomputing, vol. 120, pp. 633-644, 2013.

[16] A. Bayat, M. Pomplun, and D. A. Tran, "*A study on human activity recognition using accelerometer data from smartphones*," Procedia Computer Science, vol. 34, pp. 450-457, 2014.

[17] A. Moncada-Torres, K. Leuenberger, R. Gonzenbach, A. Luft, and R. Gassert, "*Activity classification based on inertial and barometric pressure sensors at different anatomical locations*," Physiological measurement, vol. 35, no. 7, 2014.

[18] A. M. Khan, Y.-K. Lee, and T.-S. Kim, "*Accelerometer signal-based human activity recognition using augmented autoregressive model coefficients and artifcial neural nets*," in Engineering in Medicine and Biology Society, 2008. EMBS 2008. 30th Annual International Conference of the IEEE. IEEE, 2008, pp 5172-5175

[19] Y. Cai and X. Tan, "*Weakly supervised human body detection under arbitrary poses*," in Image Processing (ICIP), 2016, pp. 599-603.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "*Ssd : Single shot multibox detector*," in European conference on computer vision. Springer, 2016, pp. 21-37.

[21] I. Goodfellow, Y. Bengio, and A. Courville, "*Deep Learning*". MIT Press, 2016.

[22] S. Ruder, "*An overview of gradient descent optimization algorithms*," CoRR, vol. abs/1609.04747, 2016.