A Novel Recurrent Neural Networks Architecture for Behavior Analysis

Neziha Jaouedi[1], Noureddine Boujnah[2], and Mohamed Bouhlel[3]
[1]Electrical Engineering Department, Gabes university, Tunisia
[1,3]SETIT Lab, Tunisia
[2]Faculté des sciences de Gabes, Tunisia

Abstract: *Behavior analysis is an important yet challenging task on computer vision area. However, human behavior is still a necessity in differents sectors. In fact, in the increase of crimes, everyone needs video surveillance to keep their belongings safe and to automatically detect events by collecting important information for the assistance of security guards. Moreover, the surveillance of human behavior is recently used in medicine fields to quickly detect physical and mental health problems of patients. The complex and the variety presentation of human features in video sequence encourage researches to find the effective presentation. An effective presentation is the most challenging part. It must be invariant to changes of point of view, robust to noise and efficient with a low computation time. In this paper, we propose new model for human behavior analysis which combine transfer learning model and Recurrent Neural Network (RNN). Our model can extract human features from frames using the pre-trained model of Convolutional Neural Network (CNN) the Inception V3. The human features obtained are trained using RNN with Gated Recurrent Unit (GRU). The performance of our proposed architecture is evaluated by three different dataset for human action, UCF Sport, UCF101 and KTH, and achieved good classification accuracy.*

Keywords: *Deep learning, recurrent neural networks, gated recurrent unit, video classification, convolutional neural network, behavior modelling, activity recognition.*

Received December 29, 2018; accepted January 19, 2020
https://doi.org/10.34028/iajit/18/2/1

1. Introduction

The analysis of human behavior from a video sequence is a rich field in diverse experiments. The term human behavior in computer vision is used to denote the physical behavior. Behavior analysis in videos [17] has proven to be an essential tool for various applications. However, in the field of medicine there is a rapidly increasing request for systems to recognize human actions, and to quickly detect patients' physical and mental health problems. Indeed, Gao *et al*. [6] developed an human action monitoring application for healthcare. This application controls the behavior of patients or the elderly. Moreover, in the field of security, the video surveillance [19, 23] is frequently integrated in public an private places which are occupied by great personal activities like: airports, banks, etc for the prevention of theft and crime. Jin *et al*. [9] presented an application to recognize the posture, the locomotion, and the gesture level of person in surveillance systems. The goal of this application is to help human for alerting, retrieval the maximum of the data. Recently, behavior analysis becomes more and more important in the assistive robotic field. Indeed, Adama *et al*. [1] developed a robots application to learn human actions by the extraction of descriptive informations about the activities in order. These activities are integrated in the context of daily living and normal human activities in indoor positioning.

Despite the great need for human behavior analysis applications, the search for an effective technique remains an important challenge. The improvement and the evaluation of algorithms rely on the extraction and learning of the features.The extraction of significant features is the most challenging part.

With the development of artificial intelligence demonstrated by the machine and the high computing capacity, some deep learning and transfer learning methods are adopted for the learning and automatic extraction of complex features from video sequence. The problem of human behavior analysis and action recognition can be explained by taking a sequential inputs and outputting a single classification. In order to solve this problem we have proposed new approach based on the combination of Convolutional Neural Network (CNN) model and Recurrent Neural Network (RNN). The Convolutional Neural networks are a type of deep models in which the convolution with trainable filters and local neighborhood pooling operations are used alternatingly on the input frames for the extraction of complex and high-level features.

The RNNs with hidden units have demonstrated superior performance in long sequential and temporal data such as speech recognition and language modeling. Their performance is primarily due to the

hidden unit that control how the update of the current activation and the production of the current state are used by the current input and the previous memory.
The aim of this work is to develop a novel efficient architecture for human behavior analysis. To achieve our goal, we have encoded the input sequential data with InceptionV3 CNN model and using RNN with Gated Recurrent Unit (GRU) units to attain the classifications. The input of our model is videos sequences. We encode the video as vector features with Inception V3, then we fed it into the GRU RNN to obtain the action predicted in the output. In this challenging, we have evaluated our architectcure on the University of Centre Florida (UCF) Sports action, UCF101 and KTH human action dataset.

The remainder of this paper is organized as follows; the section 2 addressed to the development of the behavior analysis and action recognition approaches. The section 3 presents our conributions and explains the novel architecture. The last section discusses the experimental results and summarizes the comparison with the state of the art.

2. Related Work

Behavior analysis and action recognition can be consider as a video classification problem, where the input is a sequence of frames and the output is one human action. The performance of classification methods is strongly related to the efficient extraction of significant features vector. To slove this problem many methods have been developed.

To define person, some researchers have chosen the visual description. However, Oren *et al*. [14] applied Haar wavelet coefficients in static images to extract low-level intensity human features.The wavelet coefficients are obtained by applying a differential operator at various locations, scales, and orientations on the image grid of interest. In the same contexte, Perronnin *et al*. [15] applied the Fisher Kernel (FK) algorithm and the color information to extract features from images. FK is used to extend the popular bag-of-visual-words of images.

Some other research used only human motion to present action. Indeed, Bobick and Davis [2] developed a human action recognition application based on motion detection. They used temporal template to define human motion in over time. Firstly, they applied the binary motion-energy image to detect human motion in images sequence. Secondly, they generate the motion-history image. Moreover, Wang *et al*. [21] defined the human motion by his trajectory. They used dense trajectories and motion boundary descriptors to recognize human action in video sequence. Here the trajectories capture the local motion information and the descriptor encodes the relatives motion between pixels.

Another line of research involves the human features problem by the combination of spatial and temporal informations. Dollar *et al*. [4] proposed a recognition algorithm based on spatio-temporally data. The extraction of human features is realized by Cuboid detector. Wang *et al*. [22] presented R transform algorithm to extract low-level human features. This algorithm captures both boundary and internal content of person in sequence of video. Moreover, it based on the human body and silhouette presentation to extract features. Laptev *et al*. [13] used the spacio-temporel features for action recognition. They detect space-time features using Harris operator. These features can provide tolerance to background clutter, occlusions and scale changes in video sequence. Zhen *et al*. [24] used Motion History Image and Gabor Filters (GF) to extract human feature. The GF is applied to intensify the edge information at multiple orientations and MHI is used to project human motions in video into one image. Shao *et al*. [18] presented spatiotemporal Laplacian pyramid coding and 3-D Gabor filters, for holistic representation of human actions. The Laplacian pyramid decomposes spatio-temporal data into different levels with a certain band of frequencies. Recently, Khuangga and Widyantoro [11] introduced human identification system to track the presence of persons in a room. This system based on the human body features. They applied Histogram of Oriented Gradients descriptors, HSV color conversion, and template matching for human detection.

In the last years, there has been growing interest in deep learning approaches for complex features extraction. Especially, the CNNs and the tranfer learning models have shown good performance in many applications of action recognition. Indeed, Ji *et al.* [8] presented 3D CNN model for human features extraction in video sequence. This model is evaluated by Kungliga Tekniska Högskolan (KTH) dataset. In the same context, Zeng *et al.* [25] developed an application for human activity recognition using mobile sensors in smart phone and CNNs model. This later is used to capture local dependency and scale invariance. Recently, Kamel *et al.* [10] used depth motion image descriptor to represent the depth maps sequence and moving joints descriptor to represent body postures sequence. This both features are used with CNNs for human action recognition. Some others research employed the pre-trained models of deep CNNs to describe human action. Indeed, Donahue *et al.* [5] applied invented by Visual Geometry Group (VGGNnet) pre-trained model for human visual description. Howerver Sargano *et al.* [16] chose to apply another transfer learning model to define person in video sequence. They selected the pre-trained deep CNNs AlexNet.

The performance of action recognition technique is strongly related to an efficient features vector and classification algorithms. Video classification is an important challenging in computer vision. In addition, the support vector machine is frequently used. Indeed,

[14, 23] used the linear SVM for action classification. [12] applied the non linear SVM with Radial Basic Function (RBF) kernel for action recognition. However, Kingma and Adam [12] used SVM classifier with Gaussian kernel and Sargano et al. [16] applied SVM with K Nearest Neighbors (KNN) classification methods. Recently, RNNs have demonstrated great success in video classification. Donahue et al. [5] developed an application for human action classification based on visual features and RNNs classifier. This later is used with recurrent kernel Long Short-Term Memory (LSTM) [7]. Based on the good performance of the RNNs model, we developed our contributions. In this paper, we developed novel architecture for huamn action recognition using InceptionV3 tranfer learning model for human features extraction and the RNNs with GRU [3] units for training and classification features.

3. Proposed Action Recognition Method

3.1. Human Features

The human feature is the first step of our proposed action recognition method. In this part, we adapted the transfer learning model to extract human descriptor. Transfer learning consists of applying knowledge obtained solving one problem to improve learning in performing a different but related problem. The objective of this transfer of learning strategies is to develop machine learning as effective as human learning. The transfer learning is frequently used by deep learning algorithms to improve learning and classification object. In this paper, we applied Inception V3 [20] model to extract human descriptors. Inception V3 (Figure 1) consists of symmetrical and asymmetrical basic components including convolutional layers, maximum and average pooling layers, concatenation layers, dropout layers and fully connected layers. The batch normalization layer is widely used in this model. This model is trained in ImageNet dataset.

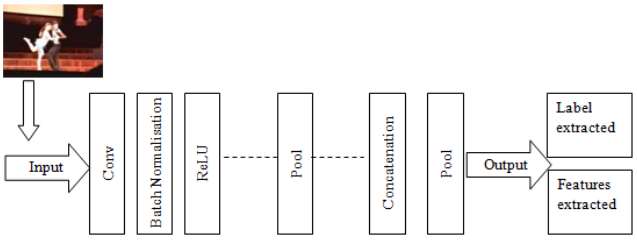

Figure 1. The transfer learning model used for human features extraction.

The human features in each frame are extracted from the last average pooling layer. Inception V3 seemed the best transfer learning model compromise between resource consumption, learning time and performance. Moreover, the concatenation of convolutional and pooling layers make the features more specifics. This model detect perfectly the humane details; face and body. This extraction make our classification model more efficient with low computing time.

3.2. Action Classification

Action classification is the second part of our proposed method for human action recognition. Our classification model presented in Figure 2. In this part we proposed the both using of transfer learning and recurrent neural network. This later is caracteriased by the good performance to sloving the long term problem by the hidden units. Two types of RNN units were developed LSTM and GRU. Firstly, the LSTM can be considered as a deep neural network architecture that can not only process static data (such as images), but also entire sequences of data (such as video). The sequential processing of a video is done by its three gates: input, output and forget gates. These gates are used to know what decision have been made in the past to make an optimal decision at time t. Secondly, the GRU has a simplified LSTM architecture with fewer parameters. Instead of having three gating like LSTM, the GRU unit has only two gating, reset gate and update gate. The reset gate determines how much information from previous memory to forget. The update gate decides how much information from previous memory can be passed along to the future. It acts similar to the forget and input gate of an LSTM cell.

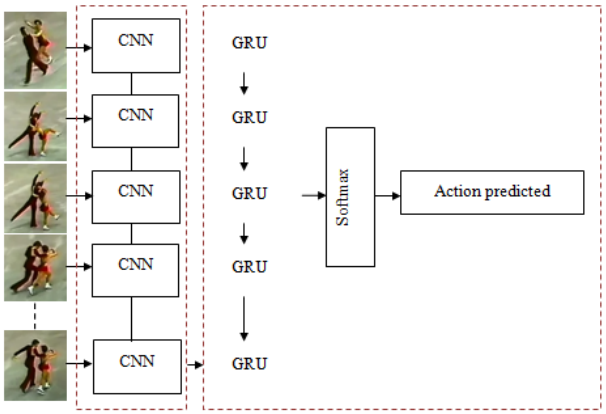

Figure 2. The proposed model for human action recognition.

Our proposed model for human action recognition relies on the extraction of relevant features from Inception v3. The model presented in three parts:

a) Presents the input of our system, here our input is a video sequence in which each video is presented by sequence of frames.
b) Presents the firt step of our system. It is the human features extraction from each frame. Indeed each frame is linked to Inception V3 model then to present the features vector of all the sequence of video, we concatenate the output of each Inception V3 of frames.

c) Presents the second step of our system. It is the action classification. The features vector extracted in the previous step present the input of this step. These features are trained by GRU RNN model. In the training phase, our model can be trained in an end-to-end manner by gradient descent using Adam update rule [12] and cross-entropy loss function through time to adjust the GRU parameters. This later are adjusted by the calculation of gradient. During the training phase, we minimize a loss function, which is measured by the difference between the true action label and the predicted action obtained. In the test phase, we repeat (a) and (b) then we project directly our features in the training GRU RNN model to obtain the probability of actions with Softmax function.

4. Experimental Results

4.1. Dataset and Implementation Details

In order to validate our method, we have used three differents dataset: KTH[1], UCFSport[2] and UCF101[3]. These datasets has been selected because they are widely used in the activity recognition and intelligent environment literature.Firstly, KTH contains 6 types of video actions with resolution 160x120. This actions presented by 25 person with different clothes in two environments in door and out door. Secondly, UCFSport contains 10 types of sport actions with resolution 720x480. It has 150 videos. Thirdly, UCF101 contains 13320 videos with 101 actions. The resolution of the frames is 320x240.

To evaluate the performance of our proposed method for human action recognition, the results are implemented and evaluated in python with the support of the Keras framework using TensorFlow. The python is installed on a laptop computer with Intel Core i7-8550U 8th Generation Processor 4.0 GHz and 8 GB memory. For the classification process, the dataset was split into a training set (75% of the dataset) and a validation set (25% of the dataset) of each dataset. To do the training, we used in each time one dataset. Firstly we used 6 actions of KTH dataset, Secondly, we taken 10 actions of UCFSport and finaly, we used 101 action of UCF101. The 75% of these actions took as the input to training our proposed model and the 25% took as input to predict action. For example in the classification process of UCF Sport, we taken 100 sequences of 10 action in the training and 50 videos in the test. Each sequence of training took as the input of our model. The sequence of video presented by his frames. Each frames is related to Inception V3 CNN model for human features extraction. The feature of frame at time t is concatenated to the feature of frame t+1 and we repeat this concatenation until the end of the video sequence. The output of Inception V3 CNN model is one vector. This vector present the concatenation of features vectors of each frame. These features took as input of our GRU RNN model. This model was using 200 GRU unit, Softmax activation function, categorical cross-entropy as the loss function and Adam as the optimizer. To validate our model, we extracted features of 50 videos of the test and we project its in the trained model GRU RNN model. The experiment was compiled for 25 epochs, with a batch size of 128 and accuracy metric function. We repeated the classification process for the two others datasets KTH and UCF101. The classification recall, precision and accuracy are presented in the Tables 1, 2, and Figure 3. The Table 1 presents the classification rates of KTH datatset. All the KTH actions present high rates recall and precision. For example handclapping and handwaving actions have the same values of recall and presicion. Here we can conclude that our proposed model is a good classifier for KTH dataset.

Table 1. Classification recall, precision and accuracy of KTH human actions.

Actions	Recall	Precision
Boxing	0.95	0.93
Handclapping	0.93	0.94
Handwaving	0.94	0.95
Jogging	0.88	0.85
Running	0.91	0.97
Walking	0.89	0.88
Accuracy	0.92	

Table 2. Classification recall, precision and accuracy of UCF sports human actions.

Actions	Recall	Precision
Diving-side	0.90	0.79
Golf swing	0.71	0.97
Kicking-front	0.88	0.66
Lifting	0.70	0.90
Riding-horse	0.79	0.91
Run-side	0.93	0.82
Skateboarding-front	0.84	0.88
Swing-bench	0.87	0.77
Swing-side angle	0.94	0.94
Walk-front	0.92	0.99
Accuracy	0.85	

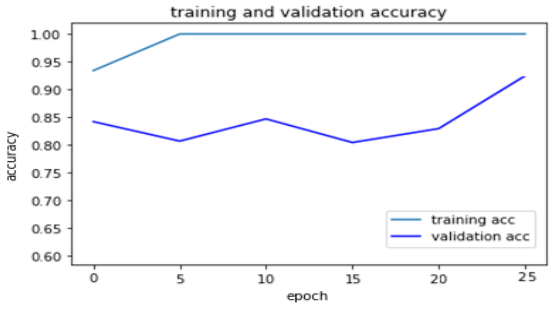

Figure 3. The training and validation accuracy of UCF101 human actions.

The Table 2 addressed to the classification of UCFSport actions. This table presents a good classification of some actions like swing-side angle

[1]http://www.nada.kth.se/cvap/actions/
[2]http://crcv.ucf.edu/data/UCF_Sports_Action.php
[3]http://crcv.ucf.edu/data/UCF101.php

where we found the same values of recall and precision.

The Figure 3 presents the variation accuracy rate in each epoch of the traning and validation process. Our proposed model can achieve the 100% of accuracy in the traning and 92% in the validation process. According to the experimentals results presented in the Tables 1, 2 and Figure 3 we can conclude that our system presents accuracy value greater than 85% for three datasets.

4.2. Comparisons with the State of the Art

To compare the performance of our methode with the methods mentioned in the state of the art, we used three datasets. For KTH dataset Table 3, some existing methods, such as [4] used the cuboids which contain spatio-temporal features and the distance euclidean for the classification of behavior. However, other methods, such as [8] used the 3D CNN and [13] employed the spatio-temporal features and multichannel non linear SVMs for human action recognition. For UCF Sport Table 4, the approaches mentioned in the related works are CNN [19] and Local Trinary Patterns [23]. These methodes requires less accuracy than our method. The same for UCF 101 Table 5 the most approaches used for action recognition based on recurrent CNN [5, 6]. In this case, we can conclude that the recognition accuracy of the proposed method is better than most of existing state-of-the art results.

Table 3. Performance comparison of different Methods on the KTH dataset.

Methods	Accuracy
Sparse Spatio-Temporal Features [10]	80%
3D CNN [15]	90.2%
multichannel non-linear SVMs [12]	91.8%
Our method	**92%**

Table 4. Performance comparison of different Methods on the UCFSport dataset.

Methods	Accuracy
CNN [2]	78.46%
Local Trinary Patterns [9]	79.2%
Our method	**85%**

Table 5. Performance comparison of different Methods on the UCF101 dataset.

Methods	Accuracy
Recurrent CNN [20]	82.34%
Recurrent 3D CNN [1]	86.8%
Our method	**92%**

5. Conclusions and Future Work

In this work we have shown the important and the dependence between the features extraction method and the classification algorithm. A new architecture was presented based on the human features and GRU RNN for human action recognition and behavior analysis. The human features technique based on all characteristic of each frame on video sequence using a model of deep learning networks. The results of this paper provide a strong importance of features to obtain high classification rate and recognize human action for some dataset like KTH and UCF101. For future work, we will improve our results in the other dataset using other methods of features extraction and actions classification.

Acknowledgment

This work was supported and financing by the Ministry of Higher Education and Scientific Research of Tunisia.

.References

[1] Adama D., Lotfi A., Langensiepen C., Lee K., and Trindade P., "Human Activity Learning for Assistive Robotics Using A Classifier Ensemble," *Soft Computing*, vol. 22, no. 21, pp. 7027-7039, 2018.

[2] Bobick A. and Davis J., "The Recognition of Human Movement Using Temporal Templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257-267, 2001.

[3] Cho K., Merrienboer B., Gülçehre Ç., Bougares F., Schwenk H., and Bengio Y., "Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation," *in Proceedings of the Conference on Empirical Methods in Natural Language*, Doha, pp. 1724-1734, 2014.

[4] Dollar P., Rabaud V., Cottrell G., and Belongie S., "Behavior Recognition via Sparse Spatio-Temporal Features," *in Proceedings of IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, Beijing, pp. 65-72, 2005.

[5] Donahue J., Hendricks L., Rohrbach M., Venugopalan S., Guadarrama S., Saenko K., and Darrell T., "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 677-691, 2017.

[6] Gao Y., Xiang X., Xiong N., Huang B., Lee H., Alrifai R., Jiang X., and Fang Z., "Human Action Monitoring for Healthcare based on Deep Learning," *IEEE Access*, vol. 6, pp. 52277- 52285, 2018.

[7] Hochreiter S. and Schmidhuber J., "Long Short-Term Memory," *Neural Computer*, vol. 9, no. 8, pp. 1735-1780, 1997.

[8] Ji S., Xu W., Yang M., and Yu K., "3D Convolutional Neural Networks for Human Action Recognition," *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221-231, 2012.

[9] Jin C., Do T., Liu M., and Kim H., "Real-Time Action Detection in Video Surveillance using a Sub-Action Descriptor with Multi-Convolutional Neural Networks," *Journal of Institute of Control*, vol. 24, no. 3, pp. 298-308, 2018.

[10] Kamel A., Sheng B., Yang P., Li P., Shen R., and Feng D., "Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 49, no. 9, pp. 1806-1819, 2018.

[11] Khuangga M. and Widyantoro D., "Human Identification Using Human Body Features Extraction," *in Proceedings of International Conference on Advanced Computer Science and Information Systems*, Yogyakarta, pp. 397-402 2018.

[12] Kingma D. and Adam J., "A Method for Stochastic Optimization," *in Proceedings of the 3rd International Conference for Learning Representations*, San Diego, 2015.

[13] Laptev I., Marszalek M., Schmid C., and Rozenfeld B., "Learning Realistic Human Actions from Movies," *in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, 2008.

[14] Oren M., Papageorgiou C., Sinha P., Osuna E., and Poggio T., "Pedestrian Detection Using Wavelet Templates," *in Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, San Juan, pp. 193-199, 1997.

[15] Perronnin F., Sánchez J., and Mensink T., "Improving the Fisher Kernel for Large-Scale Image Classification," *in Proceedings of European Conference on Computer Vision*, Heraklion, pp. 143-156, 2010.

[16] Sargano A., Wang X., Angelov P., and Habib Z., "Human Action Recognition Using Transfer Learning with Deep Representations," *in Proceedings of International Joint Conference on Neural Networks*, Anchorage, pp. 463-469, 2017.

[17] Shaout A. and Crispin B., "Streaming Video Classification Using Machine Learning," *The International Arab Journal of Information Technology*, vol. 17, no. 4A, pp. 677-682, 2020.

[18] Shao L., Zhen X., Tao D., and Li X., "Spatio-Temporal Laplacian Pyramid Coding for Action Recognition," *IEEE Transactions on Cybernetics*, vol. 44, no. 6, pp. 817-827, 2013.

[19] Silva V., Vidal F., and Romariz A., "Human Action Recognition Based on a Two-stream Convolutional Network Classifier," *in Proceedings of International Conference on Machine Learning and Applications*, Cancun, pp. 774-778, 2017.

[20] Szegedy C., Vanhoucke V., Ioffe S., Shlens J., and Wojna Z., "Rethinking the Inception Architecture for Computer Vision," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Las Vegas, pp. 2818-2826, 2016.

[21] Wang H., Kläser A., Schmid C., and Liu C., "Dense Trajectories and Motion Boundary Descriptors for Action Recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60-79, 2013.

[22] Wang Y., Huang K., and Tan T., "Human Activity Recognition Based on R Transform," *in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, Minneapolis, pp. 1-8, 2007.

[23] Yeffet L. and Wolf L., "Local Trinary Patterns for Human Action Recognition," *in Proceedings of 12th International Conference on Computer Vision*, Kyoto, pp. 492-497, 2009.

[24] Zhen X., Shao L., Tao D., and Li X., "Embedding Motion and Structure Features for Action Recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 7, pp. 1182-1190, 2013.

[25] Zeng M., Nguyen L., Yu B., Mengshoel O., Zhu J., Wu P., and Zhang J., "Convolutional Neural Networks for Human Activity Recognition using Mobile Sensors," *in Proceedings of the 6th International Conference on Mobile Computing, Applications and Services*, Austin, pp. 197-205, 2014.

Neziha Jaouedi is a doctor at National Engineering School of Gabes Tunisia. She is attached to the research Lab SETIT. Her research interests include computer vision, Human-Machine Interaction, image and video processing and pattern recognition.

Noureddine Boujnah is actually an assistant professor at Gabes University, he carried out his postdoctoral research activities at Lodz technical university-Poland.

Mohamed Bouhlel is a full professor at Sfax University Tunisia. He is the Head of the Research Lab SETIT since 2003. He was the director of the higher Institute of Electronics and Communications of Sfax Tunisia (ISECS) 2008-201. He received the golden medal with the special appreciation of the jury in 1999 on the occasion of the first International Meeting of Invention, Innovation and Technology (Dubai, UAE). He was the vice president and founder member of the Tunisian Association of the specialists in Electronics and the Tunisian Association of the experts in Imagery. He is the president and founder of the Tunisian association in Human- Machine Interaction since 2013. He is the Editor in chief of the international Journal "HMI","MLHC" and a dozen of special issues.