International Journal of Applied Evolutionary Computation
Volume 13 • Issue 2

Real-Time Human Action Recognition Using Deep Learning

Houssem Eddine Azzag, École supérieure en informatique de Sidi Bel Abbès, Algeria*
Imed Eddine Zeroual, Université Mohamed-Chérif Messaadia, Souk Ahras, Algeria
Ammar Ladjailia, Université Mohamed-Chérif Messaadia, Souk Ahras, Algeria
https://orcid.org/0000-0001-8364-1791

ABSTRACT

The future of computer vision lies in deep learning to develop machines to solve our human problems. One of the most important areas of research is smart video surveillance. This feature is related to the study and recognition of movements, and it's used in many fields, like security, sports, medicine, and a whole lot of new applications. The study and analysis of human activity is very important to improve because it is a very sensitive field, like in security, the human needs a machine's help a lot; and in recent years, developers have adopted many advanced algorithms to discover the type of movements humans preform, and the results differ from one to another. The most important part of human activity recognition is real time, so one can detect any issue, like a medical problem, in time. In this regard, the authors will use methods of deep learning to reach a good result of recognition of the nature of human action in real time clips.

KEYWORDS

AI, Artificial Intelligence, Computer vision, Deep learning, Human action recognition

1. INTRODUCTION

We, humans, love the way we think. So, we worked on making computers think and learn like us. Artificial intelligence is the field that enables computers to solve human problems using the smartest human methods. Previous methods used to solve problems with computers were by programming each case by writing its detailed program and setting the conditions in tedious detail to solve that case. This is an impossible task with cases that have a large number and variety of possibilities, as well as problems that contain hidden possibilities. So guiding computers has gone from giving them step-by-step instructions to teaching them with examples. Then, the computer's job will be to find the causes of each scenario based on statistics and probabilities, provide us with its conclusions and take the appropriate action for that specific scenario. Human life is almost impossible without machine technology. The developers have recently been interested in studying human activity through video,

DOI: 10.4018/IJAEC.315633 *Corresponding Author

which relies on studying the type of movement, which is a very useful feature in many areas-- with its wide scope, international security services have become fully dependent on these studies, in addition to sports, medicine, etc. The nature of the human being is known to be very mobile, but many movements are difficult to recognize through machine technology. Researchers have devoted great efforts to tackling this particular field (Panousis 2020). The indicated project has been made with HMMs model of recognizing, using MRSA dataset and another and an INDIVIDUAL dataset with accuracy avg 89.45, (Li 2020). However, here they use the model of video recognizing with RGBD, which means video detection with color. (Parisi 2020) In this project, they transform every frame into a structure and study the movements of the structure. Several variants of activity recognition have been made by removing the background method and putting all frames in one picture, and then doing a simple classification from them (Goffredo 2009). But the problem here is that all the progress is used to detect only one activity in a video (Ladjailia 2020); this means if we have a clip of a person at first, he starts walking, and after a few moments, he starts running, we can't detect those two types of movement because they are in the same clip. So we try to solve this problem based on the other result, where we employ a variant with KNN Neural Network to capture different human activity in the same clip.

2. RELATED WORK

Researchers have devoted a lot of efforts in the field of artificial intelligence that studies the recognition of human activity through videos, and in recent decades they have found many different ways (Lei 2019), (Yadav 2019) (Figure 1) to improve prediction results in terms of motion recognition accuracy and recognition speed in live video clips. Some clip-to- skeleton techniques were used (Parisi 2020) to discover differences by creating a skeletal model that simulatesthe movements in the original video. The differences are studied for the structural model and the prediction results are given for the original video. While others used special techniques represented in identifying each section of the body gradually, and this helps many people to get good accuracy, many researchers use a method based on these RGBD colors to color some parts of the body with different colors (Luvizon 2019), (Liu 2020) and use them to detect activity by color (Khowaja 2020). But the best way is to use the person as a shadow without any background or effects (Gnouma 2019); some other solutions are based on placing the skeleton in Space-Time and doing a 3D tracking (Yadav 2019). So, we can say this filed is missing two points: the first one, most of the old research is not in real-time, and the second one is the accuracy of the work according to all these results. It's time to give our way to improve those problems, to get a real-time recognition with high accuracy.

3. PROPOSED MODEL

To be sure the accuracy of our work is high, so we guarantee very little rate of error and transparency at work, our model is based on two important things.

1) ***Static Camera***: No movement of a camera. This means all scenes will be in the same place with the same position and same angle. This thing will also include section analysis by removing the background.
2) ***Classification frame by frame***: Processing or classification frame by frame gives you the result of a type of movement. When we ensure that we have a static camera and we work with frame by frame classification, we guarantee real-time activity detection and good recognizing and accuracy (Table 1).

Figure 1. A different method of Human Action Recognition

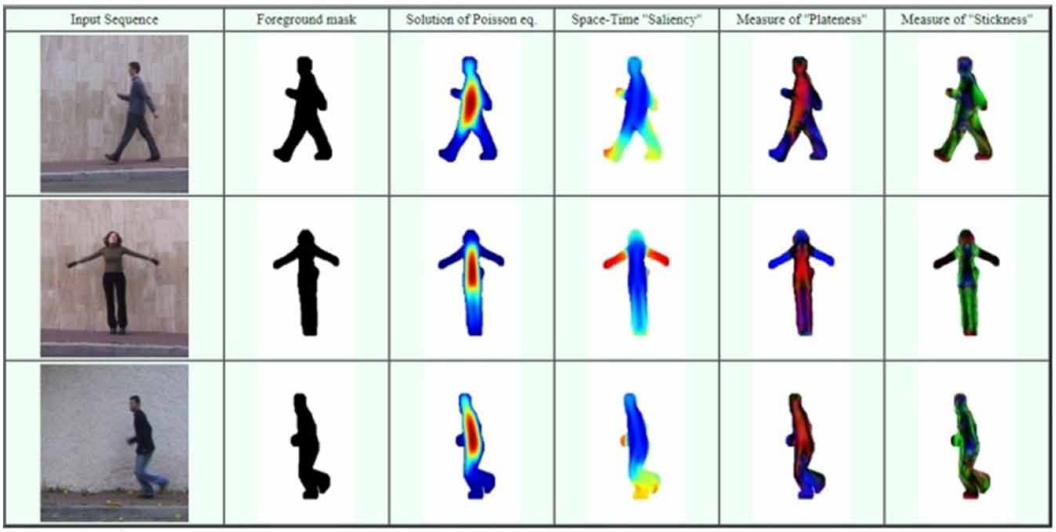


Table 1. Comparative results with recent approaches

Method	Accuracy
Our Method	84.44%
Decision Tree (Kin 2009)	79.56%
LSTM (Rivera 2017)	80%
SVM+CNN (Ferrari 2020)	71.89%

3.1 Model Definition

Let F = {frame(t)[y:y+h, x:x+w] }

t time frames, [x,y,w,h] sizes of structures. Following the definition every frame in any clip we convert them into picture without background (Figure 2). We proposed a conventional neural network (CNN) technique with the model architecture proved in Figure 3, starting with an input layer of shape (40x80x2) as a 2D image that came from model training part. They pass for 5 hidden layers until the final layer is a dense layer with 10 (the number of classes varies according to the number of movements).

When we find our outputs (Pictures from every frame without background) the next step is to select only the person from every frame "Figure 4" after all we go to the step of classification

3.2 Weizmann Dataset

There is a built-in dataset content of 10 human activities (Figure 5), and every activity content 9 different videos. With 180x144 resolution, for this work we modify some things. In this dataset we create some video content with two different activities in the same clip, to test the accuracy of live recognizing.

Figure 2. Convert clip into pictures

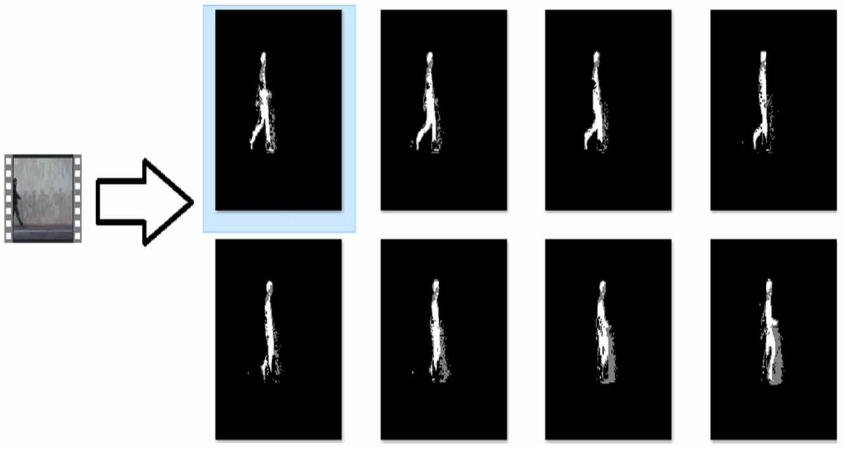


Figure 3. CNN architecture

Layer	Input	Output
InputLayer	40x80x2	40x80x2
Conv2D	40x80x2	40x80x32
MaxPool2D	40x80x32	20x40x32
BatchNorm	20x40x32	20x40x32
Conv2D	20x40x32	20x40x64
MaxPool2D	20x40x64	10x20x64
BatchNorm	10x20x64	10x20x64
MaxPool2D	10x20x128	5x10x128
BatchNorm	5x10x128	5x10x128
Conv24	5x10x128	5x10x128
MaxPool	5x10x128	2x5x128
Flatten	2x5x128	1280
Dense	1280	128
Dense	128	10

Figure 4. Select the place of the person in the picture

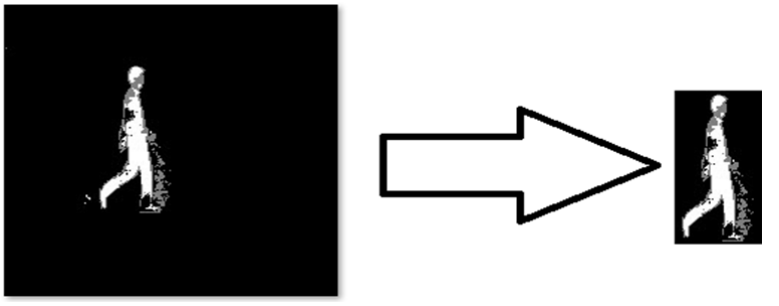
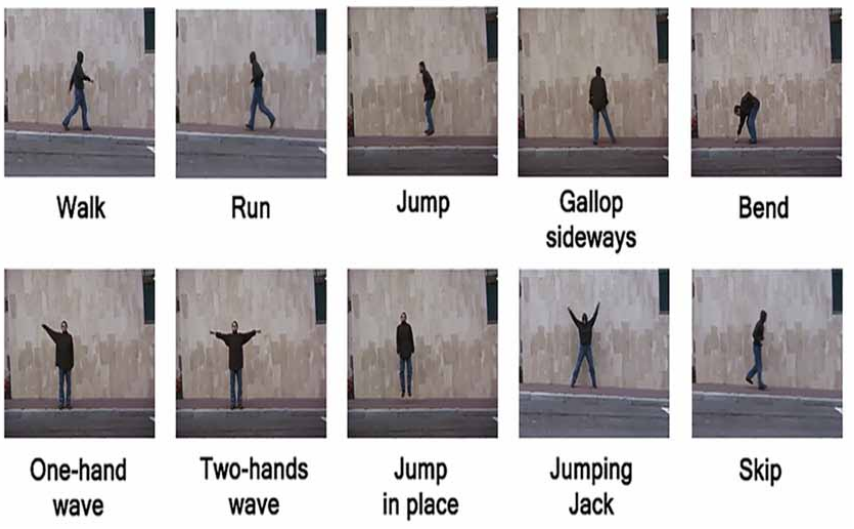


Figure 5. Weizmann Dataset activity



3.3 Model Training

Our training model is based on finding the coordinates of the place of the person in every frame. So, every video will be converted into pictures (Figure 5). In those pictures we have only the structure of the person in the scene without anything. Those pictures are with different resolution, so we decided to create a version of them with fixed resolution-- in our case 40x80 (Figure 6), so then our training data will be those (40x80) pictures.

3.4. Classification Process

The final step in any action recognition is to feed the final data into a classifier, which adopted one of the classification algorithms. Here, we followed a conventional neural network (CNN) with the model architecture proved in Figure 3. We split our data into two parts: test 70%, and 30% training. Also the training part we split it to two parts (test/validation) with the same split ratio as last time 70% and 30%. We start our training, and after about 40 epochs we get very good results (Table 2). Using the Weizman dataset, and also the confusion matrix.

Figure 6. Training images 40x80

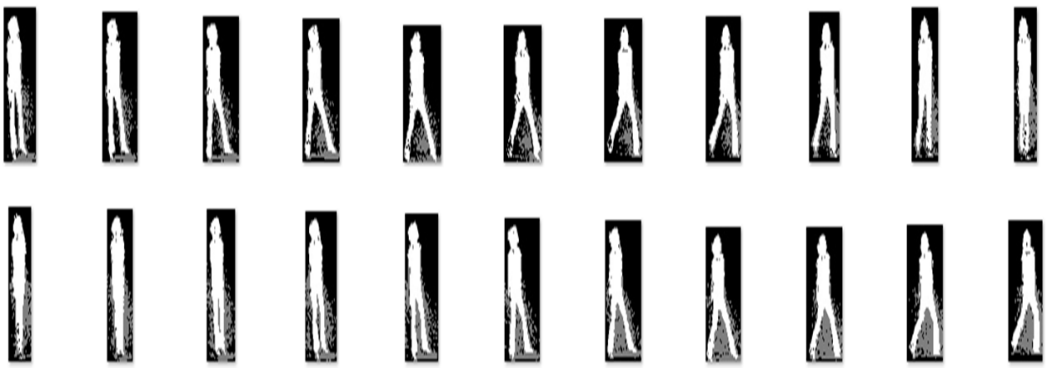


Table 3, after finishing the training part, we turn into the detection part. It's the step that is more important for the user, so after getting our model from training we use it to predict every frame from live stream video. The model gives us results in every frame; for testing our model, we are using a small laptop with 8GB RAM, i5 5th Gen CPU, NVIDIA Gefore 920M, and we get a frame rate of around 40 frames per second. So, we can say that our method can work with good accuracy, and an acceptable FPS to work within any device so our model will be very usable.

4. ACTION RECOGNITION

Movement recognition is the process of watching a person make a movement, and through certain characteristics of this movement, identifying it and differentiating it from other similar movements. Behavior identification is currently one of the largest areas of research, working on the principle of capturing a clip of a specific movement on the ground, whatever its source. Then, modern algorithms that are passed on this section through which a system can tell you what movement was made (Luvizon 2018).

In the classification process part, we do other tests with different clips' content in the same scene. This is for testing real-time tracking, and you can see results in Table 2, where the average of accuracy with our method is 84.44. The highest value is 98.12 in Wave action, and the lowest value is 63.44 (in the skipping, action because it's too similar to the running action), but mostly it is very useful for real-time detecting. For reasons of trying our method in real-time detection, we give some clips with different activities. The good thing is the result is the same as Figure 7, so we give result in every frame (figure 6). It is useful in security detection and so many other fields with a lesser chance of being wrong. We give you a confusion matrix in Table 3 to add more visuals of our model results in different dataset classes.

Table 2. Accuracy Result of some Weizmann Activity

Activity	Acc 1%	Acc 2%	Acc 3%	Avg
Walk	83.56%	88.88%	87.33%	86.59%
Rum	87.80%	71.42%	82.36%	80.52%
Jump	86.84%	89.18%	92.13%	89.38%
Skip	63.44%	71.14%	76.52%	70.36%
Wave	98.12%	96.29%	96.33%	96.91%
Other	82.23%	80.14%	86.27%	82.88%
Results	83.66%	82.84%	86.82%	84.44%

Table 3. Confusion matrix for human action recognition: results for cross-matching of different classes

	1	2	3	4	5
1	**49**	5	2	6	3
2	7	**22**	6	11	14
3	7	8	**11**	8	12
4	12	9	6	**14**	18
5	8	12	6	7	**37**

5. HUMAN ACTION RECOGNITION IN VIDEOS

The main goal of human action recognition in videos is to identify the unknown actions happening. This goal is achieved by analyzing the frames of these videos to form and build a set of data that can be classified efficiently in terms of accuracy, speed, and simplicity. The main structure of human action recognition consists of two main stages: human object tracking, and action classification. The first stage has to answer the question of how to detect or segment and track the human object in each frame of the video sequences. The second stage has to answer the question of how to classify the data from the first stage by applying an effective classification algorithm (Hussein 2019).

6. EXPERIMENTAL RESULTS

To test the effectiveness of the proposed method, we conducted experiments on the aforementioned Weizmann dataset. As we mentioned earlier, the dataset contains 10 human activities, each containing 9 different sections. In our case, we used 6 videos for training and the remaining three for test and validation.

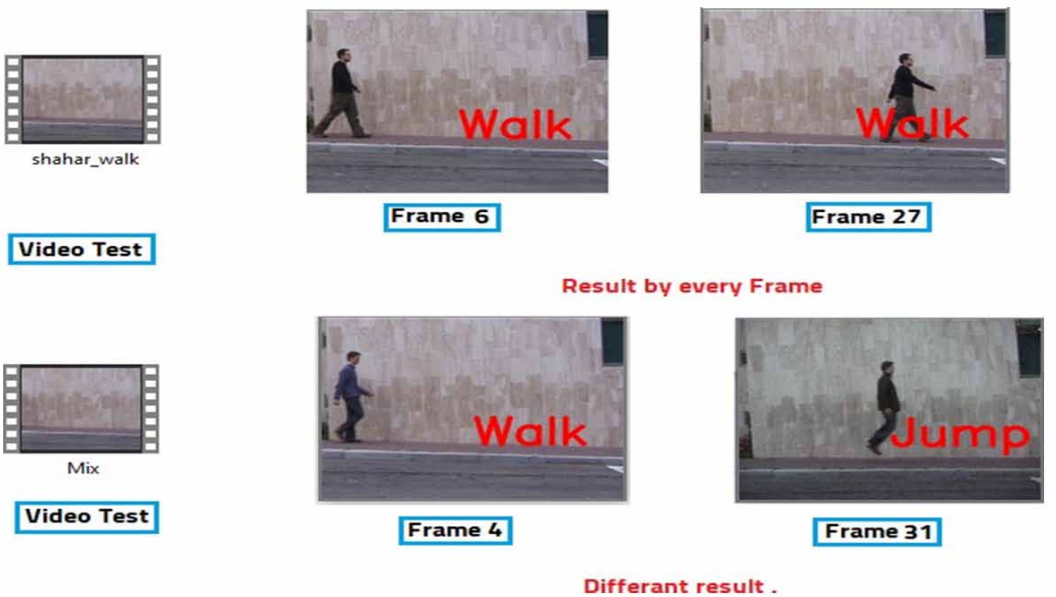
7. FUTURE RESEARCH DIRECTION

Despite the great efforts made in this research, it does not reach perfection, so it is necessary to work carefully in developing this research in the future through the following points:

7.1 Work with Movement Camera

One of the problems that the authors encountered in this research is that the authors work only in static cameras, so background subtraction is possible, but when you work with movement camera background subtraction is very hard-- and in some cases, we can say it's impossible. So, it's a huge challenge to work for human activity recognition with a movement camera (Yang Tian 2016).

Figure 7. Prediction phase



7.2 Next Action Prediction

Recognizing the next movement of a person includes identifying the current movement of that person, so when you can detect the movement that the person doing now you can predict the next movement. It's very helpful in detecting human intent (Ryoo 2011), it can be used, for example, to detect driver behavior to avoid accidents (Fox 1997).

7.3 Detect Object and His Movement

In this method, you can't detect the placement of the person, only his action. In the future, the authors can think of building a system that can know the location and movement of a person at the same time (Liem Gavrila 2014), (Kellokumpu 2010).

7.4 Detect New Action

Recognizing a movement that the model has never seen before is one of the biggest challenges in the field of deep learning, and this is also a challenge for the authors in their upcoming research to create a system capable of knowing new movements that it has not gone through before and save them and use them for the coming times (Davis 1997).

8. CONCLUSION

The security field is one of the most important in the world because of the importance of what people have as data, money, and everything; also, the government needs to avoid any problems. We need to improve this important field, so our method is one of many methods that is used to add computer vision to the security filed to help humans.. Despite the different theories and methods of studying motion, it is still impossible to reach a hundred percent of the results in determining what a movement is, so we can say a computer alone can't compensation a human in a security work, but they can help him a lot.

Also, movement detection is not only used in the security field, it can be used in many other fields, like sports and medicine. But as we say about security, it can't replace the whole human and work alone. But we conclude from this study that we can obtain a practical and effective methods in several areas of real-time human action recognition, although some basic elements are required to improve the accuracy and progress of work effectively and permanently. Our experiments provided strong empirical evidence for the effectiveness of our approach. Note that the studied model, (1) gives considerable accuracy of recognition, and (2) and emphasizes the importance real-time tracking. In our future actions, we will be determined to be able to achieve a higher level of compatibility.

REFERENCES

Al-Ali, S., Milanova, M., Al-Rizzo, H., Fox, V. L. (2015). Human action recognition: contour-based and silhouette-based approaches. In *Computer Vision in Control Systems-2* (pp. 11-47). Springer, Cham.

Davis, J. W., & Bobick, A. F. (1997, June). The representation and recognition of human movement using temporal templates. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (pp. 928-934). IEEE.

Ferrari, A., Micucci, D., Mobilio, M., & Napoletano, P. (2020). On the personalization of classification models for human activity recognition. *IEEE Access: Practical Innovations, Open Solutions*, *8*, 32066–32079. doi:10.1109/ACCESS.2020.2973425

Fox, G. K., Bowden, S. C., & Bashford, G. M., & SmithD. S. (1997). Alzheimer's disease and driving: Prediction and assessment of driving performance. *Journal of the American Geriatrics Society*, *45*(8), 949–953.

Gnouma, M., Ladjailia, A., Ejbali, R., & Zaied, M. (2019). Stacked sparse autoencoder and history of binary motion image for human activity recognition. *Multimedia Tools and Applications*, *78*(2), 2157–2179. doi:10.1007/s11042-018-6273-1

Goffredo, M., Bouchrika, I., Carter, J. N., & Nixon, M.S. (2009). Self-calibrating view-invariant gait biometrics. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 40*(4), 997-1008.

Hussein, N., Gavves, E., & Smeulders, A. W. (2019). Reception for complex action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 254-263). IEEE.

Kellokumpu, V., Zhao, G., & Pietika¨inen, M. (2010, July). Dynamic textures for human movement recognition. In *Proceedings of the ACM International Conference on Image and Video Retrieval* (pp. 470-476).

Khowaja, S. A., & Lee, S. L. (2020). Semantic image networks for human action recognition. *International Journal of Computer Vision*, *128*(2), 393–419. doi:10.1007/s11263-019-01248-3

Kim, E., Helal, S., & Cook, D. (2009). Human activity recognition and pattern discovery. *IEEE Pervasive Computing*, *9*(1), 48–53. doi:10.1109/MPRV.2010.7 PMID:21258659

Ladjailia, A., Bouchrika, I., Merouani, H. F., Harrati, N., & Mahfouf, Z. (2020). Human activity recognition via optical flow: Decomposing activities into basic actions. *Neural Computing & Applications*, *32*(21), 16387–16400. doi:10.1007/s00521-018-3951-x

Li, M., Chen, S., Chen, X., Zhang, Y., Wang, Y., & Tian, Q. (2021). Symbiotic graph neural networks for 3d skeleton-based human action recognition and motion prediction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Liem, M. C., & Gavrila, D. M. (2014). Joint multi-person detection and tracking from overlapping cameras. *Computer Vision and Image Understanding*, *128*, 36–50.

Liu, T., Dong, X., Wang, Y., Dai, X., You, Q., & Luo, J. (2020). Double-layer conditional random fields model for human action recognition. *Signal Processing Image Communication*, *80*, 115672.

Luvizon, D., Picard, D., & Tabia, H. (2020). Multi-task deep learning for real-time 3D human pose estimation and action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1. doi:10.1109/TPAMI.2020.2976014 PMID:32091993

Luvizon, D. C., Picard, D., & Tabia, H. (2018). 2d/3d pose estimation and action recognition using multitask deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5137-5146). IEEE.

Panousis, K. P., Chatzis, S., & Theodoridis, S. (2020). Variational Conditional-Dependence Hidden Markov Models for Human Action Recognition.

Parisi, G. I. (2020). Human action recognition and assessment via deep neural network self-organization. In *Modelling Human Motion* (pp. 187–211). Springer. doi:10.1007/978-3-030-46732-6_10

Qing, L. (2019). A Survey of Vision-Based Human Action Evaluation Methods. *Sensors,* 4129.

Rivera, P., Valarezo, E., Choi, M. T., & Kim, T. S. (2017). Recognition of human hand activities based on a single wrist imu using recurrent neural networks. *Int. J. Pharma Med. Biol. Sci*, *6*(4), 114–118. doi:10.18178/ijpmbs.6.4.114-118

Ryoo, M. S. (2011, November). Human activity prediction: Early recognition of ongoing activities from streaming videos. In *International Conference on Computer Vision* (pp. 1036-1043). IEEE.

Yadav, S. K., Singh, A., Gupta, A., & Raheja, J. L. (2019). Real-time Yoga recognition using deep learning. *Neural Computing & Applications*, *31*(12), 9349–9361. doi:10.1007/s00521-019-04232-7

Yang, X., & Tian, Y. (2016). Supernormal vector for human activity recognition with depth cameras. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(5), 1028–1039.

Houssem Eddine Azzag is a computer science student in École supérieure en informatique de Sidi Bel Abbès. Houssem is 21 years old, and works in artificial intelligence and computer vision fields.

Imed Eddine Zeroual is a computer science student in University of Souk Ahras. Imed is 24 years old, and works in artificial intelligence and computer vision fields.

Ammar Ladjailia was born in Souk Ahras, Algeria, on 11/29/1977. He graduated in Computer Engineering from the University of Annaba in 2000. He obtained a master's degree in computer science (artificial intelligence) from the University of Annaba in 2003. In 2019, he will obtain a doctorate in computer science from the same university, while working as a lecturer in computer science at Souk Ahras University. His research focuses on the recognition of human activities, automated and intelligent visual surveillance, and image and video processing.