*I.J. Intelligent Systems and Applications,* **2019, 3, 42-52**
Published Online March 2019 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijisa.2019.03.05

# 3D Skeleton Action Recognition for Security Improvement

**Adlen Kerboua**
Department of computer science, University of Constantine 2 - A. Mehri, Constantine, Algeria
E-mail: adlen.kerboua@univ-constantine2.dz

**Mohamed Batouche**
Department of computer science, University of Constantine 2 - A. Mehri, Constantine, Algeria
E-mail: mohamed.batouche@univ-constantine2.dz

Received: 02 December 2017; Revised: 11 February 2018; Accepted: 25 April 2018; Published: 08 March 2019

***Abstract*****—**Most of action recognition methods allow achieving high action recognition accuracy, but only after processing the entire video sequence, however, for security issues, it is primordial to detect dangerous behavior occurrence as soon as possible allowing early warnings. In this paper, we present a human activity recognition method using 3D skeleton information, recovered by an RGB-D sensor by proposing a new descriptor modeling the dynamic relation between 3D locations of skeleton joints expressed in Euclidean distance and spherical coordinates between the normalized joints, A PCA dimension reduction is used to remove noisy information and enhance recognition accuracy while improving calculation and decision time. We also study the accuracy of the proposed descriptor calculated on limited few first frames and using limited skeleton joint number, to perform early action detection by exploring several classifiers. We test this approach on two datasets, MSR Daily Activity 3D and our own dataset called INDACT. Experimental evaluation shows that the proposed approach can robustly classify actions outperforming state-of-the-art methods and maintain good accuracy score even using limited frame number and reduced skeleton joints.

***Index Terms*****—**Action recognition, RGB-D sensor, skeleton joint, classification.

## I. INTRODUCTION

Inspired from the efficient action recognition method proposed by X. Yang et al. [1] we propose a new action recognition descriptor, using both Euclidean distance and spherical coordinates of 3D skeleton joints for action recognition. We develop a descriptor containing discriminating features for human activity classification, by adopting the differences of 3D normalized skeleton joint coordinate in both temporal (across frames) and spatial (in the same frame) domains to model the dynamics of every joint and the configuration of all available joints by modeling distances (Euclidean plan) and angles (spherical coordinates) between them. Skeleton joints are compact information comparing to any other like depth or RGB but can be affected in video surveillance applications by high camera mounting or human subject occlusion, to overcome these cons, we propose to recover missing skeleton joints in preprocessing step by interpolating those of neighbors' frames. Unlike the traditional trajectory-based methods, our descriptor is capable of modeling actions through more informative and more compact body joints without background noise; it is discriminating and simpler to compute. This choice helps us to reduce the length of the descriptor for more calculation efficiency and enables making fast action recognition using four classifiers (decision tree, ensemble subspace discriminant, bagged tree and SVM).

The proposed method can deduce relevant information for monitoring security in the industrial field. Indeed, as we can get to know the danger level of performing actions at an early stage to avoid the dangerous consequences, this technique can be associated with a fire and a spark detection method like the method proposed by [2,3] to increase robustness and give a proper alert for a real situation.

To classify actions rapidly, we compute light descriptor that encodes information from the first few frames and limited skeleton joint number. Hence, we studied the impact of this limitation on the action recognition accuracy, reducing both the time taken by system to observe distinctive frames for making correct classification and reduce the time of calculation due to reduced size of the descriptor.

This paper has two contributions:

- First, we propose a new descriptor containing the combination of two light information representing distance and spherical information encoding 3D skeleton joint locations captured by an RGB-D sensor to robustly classify actions in near real time.
- Second, we study action recognition anticipation for security purposes by reducing descriptor complexity to contain only the most representative skeleton joints in just first few frames.

This improvement decreases significantly the amount of information needed to build distinctive descriptor for processing time efficiency while keeping a decent accuracy. This method can also handle activity with low amplitude of movement because this method is not based on movement detection algorithm like optical flow or Gaussian mixture model.

The remainder of this paper is organized as follows. Section 2 reviews some related works. Section 3 gives a detailed presentation of our dataset INDACT specifically for the industrial field. Section 4 is dedicated to the description of our activity recognition approach which includes a preprocessing phase (joints normalization), and the construction of a descriptor based on merging spherical and distance information. Some experimental results, using challenging MSR Daily Activity 3D dataset, and INDACT dataset, are presented in section 5. Finally, conclusions and future work are draw.

## II. Related Work

The rich information provided by the new generation of RGB-D low cost sensor like the Microsoft Kinect can be used for human activity recognition applications. They capture the visible spectrum (RGB) as a conventional camera, simultaneously as depth information (D) with high frequency which can be easily exploited to build list of features for human activity labeling application. This type of device finds large fields of applications, from the e-care at home for older people monitoring, video surveillance indexing, to working behavior analyzing for productivity and security improvement in the industrial domain. For the last kind of application, it will be very interesting to be able to detect dangerous action occurrences as soon as possible to avoid accidents. State of art in action recognition domain offers a large range of detecting methods, giving more and better accuracy scores, especially those combining both RGB and Depth information.

Many approaches have been proposed for human activity recognition. These techniques have been surveyed in [4-6]. The event of depth sensor permits adding 3D information in real time to action recognition possibility. Compared to a conventional RGB camera, the depth camera has several advantages, Depth images are insensitive to changes in lighting conditions or in point of view. Moreover, depth information eases human skeleton detection as proposed by [7-9]. This technological development allows exploring research in action recognition based on RGB-D information field. The main idea is to extend classical methods proposed for RGB action recognition, by adding 3D information. In [10] authors propose a Bag of 3D Points model by sampling a set of 3D points from a body surface to describe the posture being performed on each frame and projected by mapping the depth map onto three orthogonal Cartesian planes. In reference [11], the authors used a Histogram of 3D Joint Locations (HOJ3D) to represent posture and to compute spherical coordinates from skeleton joints to get a view-invariant descriptor. The temporal information is then coded by Discrete Hidden Markov Models (HMM). Sung et al. [12] used both RGB frame and depth map to recognize human daily activities, skeleton joints were used to model body pose, hand position, and motion information. They extracted Histogram of Oriented Gradients (HOG) features from the region of interest in gray level images and depth maps to characterize the appearance information.

Depth Motion Maps DMM where obtained [13] by concatenating the projected 3D depth maps onto three planes. HOG was then computed from DMM as a global representation of human action. In the same way, Actionlet mining algorithm was proposed by [14] to perform a selection of skeleton joints. In addition to joint-based feature, they also made use of depth maps to characterize object shape and appearance.

The method described in [1] propose also efficient skeleton based action feature descriptor, Eigen Joints, for action recognition. They designed action feature descriptor by adopting the difference (Cartesian distance) between 3D skeleton joints in temporal and spatial domains to explicitly model the dynamics of each individual joint and the configuration of all available joints. Wu et al. [15] proposed to use polar coordinates as the one description of activity features. Eweiwi et al. [16] also suggested a combination of joint spherical coordinates to encode spatial information and joint velocities feature to keep temporal information over a fixed set of K frames as 3d or 2d histograms. However, the limited frame window losses the overall dynamic of action.

Most of the recognition systems proposed in literature need to process entire video sequences to perform action recognition. While in real time video surveillance, the system must require only few observations as possible. Schindler et al. [17] studied how many frames were needed to classify an action with acceptable robustness. They concluded that limited action subsequence with a few first frames can encode enough information as the entire video. In [1], the authors have addressed this aspect. They evaluated the minimum number of frames sufficient to enable accurate action recognition, and they find that the first 30–40% frames allowed achieving comparable recognition accuracies to the ones using entire video sequences.

## III. Details of the Approach

Our approach for action recognition using skeleton information is partially inspired from a study conducted by Yang et al. [1], this contribution has been followed by several other variants using a combination of the spatial and the temporal relation of 3D skeleton joints [9,10].

Because of the ubiquity of noise in depth map given by Microsoft Kinect used by skeleton joints estimation algorithm developed in [7], we got some frames with missing skeleton information. To solve this problem, we must perform a preprocessing step to recover missing skeleton information by interpolating smoothed joints belonging to neighbor frames. Besides, a normalization process is done to make the method cross-subject by making human subject limbs dimensions' equals.

Before the detailed presentation of the process, we denote the overall structure of algorithm. It can be divided into two parts: the first one deals with the preprocessing, while the second part concerns essentially the construction of feature descriptor used to train the classifiers. An overview of our algorithm is illustrated in Fig.1. ach step will be detailed in the remainder of this paper.

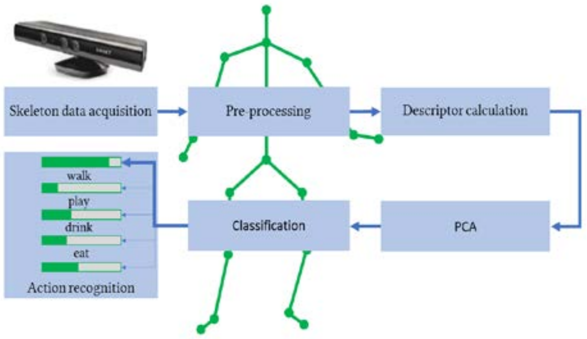


Fig.1. Overview of the approach.

## A. Data preprocessing

Before starting the constructing of the descriptor, we must perform two preprocessing tasks:

- First, to reduce noise in joint estimation, we smooth their positions over frames by applying local regression using weighted linear least squares and a 2nd degree polynomial model. In the same time, according to individual speed, similar gestures can be performed in different time duration by each subject, resulting different number of frames. To uniform the skeleton information over frames, we use cubic interpolation of the values of neighboring grid points in each respective dimension. This method allows removing wrongly estimated joints (Fig.2).

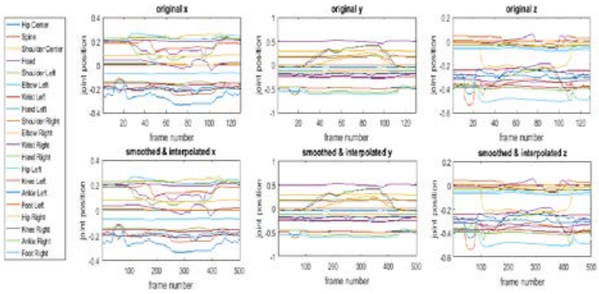


Fig.2. Skeleton joint smoothing and frames number standardization by interpolation.

- Second, given variations in body sizes which can cause intra-classes variations and confuse the classifier, and to compensate those variations, we follow the method presented in [18] by imposing the same limbs lengths for skeletons of all individuals in the dataset (Fig.3). We use fixed distances to link joints in the skeleton according to the body symmetry D={d1, d2, d3, d4, d5, d6, d7, d8, d9, d10, d11}, starting from the root node (hip center joint), moving forward to the branches, this process keeps the joint angles unchanged, while the same limbs will have the same length across subjects and frames.

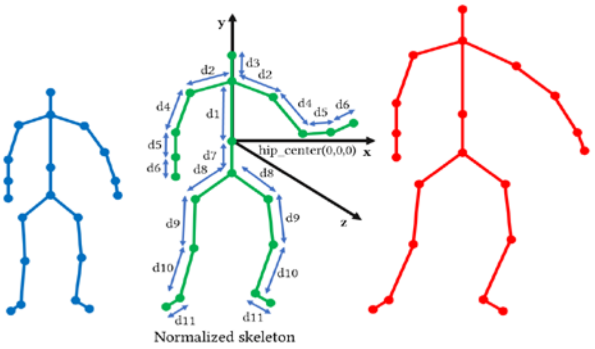


Fig.3. Body sizes standardization and skeleton translation.

To overcome the camera position variation, we apply a translation to make the hip center as the origin of skeleton joints.

## B. Construction of distance and spherical descriptors

The accuracy of activity detection method depends widely on the robust construction of distinctive feature descriptor. It must contain enough distinguished information to correctly classify actions. All must be light for calculation efficiency. For this aim we choose to mix two types of features from Euclidean plan and spherical coordinates. We also study the effect of reducing the number of frames/joints considered for the calculation of the descriptor. To reach good performance, we use more action information characteristic of each action including both distance and spherical coordinates of each joint. with this improvement, we obtain more action characteristics. In practice, in distance domain we compute Euclidean distances and spherical coordinates (angles) including:

- The overall action dynamics representing distances between all joints in frame *C* and those belonging to the first frame.
- The descriptor of motion property containing joints differences between the current frame C and its preceding frame *C-1*.
- The static distance posture of the body part by calculating pair-wise joints differences in current frame *C*,

Let suppose that $jo_i^c$ is the 3D coordinate of joint *i* representing 3D coordinates of body parts centroids in frame *C* (RGB-D sensor gives 20 joints by frame), we have:

$$jo_i^c = \left[x_i^c, y_i^c, z_i^c\right]; c \in \{1, \ldots, n\}; i = 1, \ldots, 20 \qquad (1)$$

We calculate Distance Descriptor vector *DD^c^* for frame *C* as in:

$$DD^c = \sqrt{\left(\left(x_i^c - x_j^d\right)^2 + \left(y_i^c - y_j^d\right)^2 + \left(z_i^c - z_j^d\right)^2\right)}; \quad (2)$$
$$d \in \{1, c-1, c\}; i, j = 1, 2, ..., 20$$

Analogically, in the spherical domain, we compute the Spherical Descriptor vector $SD^c$ for frame *C*:

$$SD^c = \left[az^c, el^c\right]; \quad (3)$$

Containing both azimuth $az^c$ and elevation $el^c$ as below:

$$az^c = \arctan\left(y_i^c - y_j^d, x_i^c - x_j^d\right); \quad (4)$$
$$d \in \{1, c-1, c\}; i, j = 1, 2, ..., 20$$

$$el^c = \arctan\left(z_i^c - z_j^d, \sqrt{\left(\left(x_i^c - x_j^d\right)^2 + \left(y_i^c - y_j^d\right)^2\right)}\right);$$
$$d \in \{1, c-1, c\}; i, j = 1, 2, ..., 20 \quad (5)$$

The final $DSD^c$ Distance/Spherical Descriptor:

$$DSD^c = \left[DD^c, SD^c\right]; c \in \{1, ..., n\} \quad (6)$$

The final descriptor contains the concatenation of both distance and spherical information obtained over frames that includes sufficient amount of information to classify actions into their respective classes using four classifiers.

As we use 20 joints in each frame, it might result a features dimension of $DD^c$ vectors containing 3*20² = 1200 pairwise comparisons. As spherical descriptor $SD^c$ contains two comparisons (azimuth and elevation), $SD^c$ vector contains 2*3*20² = 2400 features. In total, we have about 3600 elements per frame in the $DSD^c$ vector. As skeleton joints are high-level information recovered from noisy depth maps, this large dimension might be redundant and noisy. Hence, we choose to apply PCA dimension reduction algorithm to remove redundancy and disturbances in the final descriptor. We keep only 512 strongest features that contain over 99% of the significant information.

### C. Used classifiers

We have explored four types of classifiers, including decision tree, ensemble subspace discriminant, bagged trees and SVM, all available in Matlab's Statistics and Machine Learning Toolbox™

All parameters used are depicted in Table 1. Our application needs to make a trade-off between the speed of training, memory usage and accuracy, so we test those classifiers in both classification accuracy, training and testing times.

Table 1. Parameters used for the four classifiers.

| Classifier | Parameters |
|---|---|
| Simple decision tree | Maximum number of splits = 22 & split criterion ginis diversity index. |
| Ensemble subspace discriminant | Learner type nearest neighbors with 200 learners and 200 subspace dimensions. |
| bagged trees | Ensemble method bag & learner type decision tree & number of learners 380. |
| SVM | SVM classifier kernel function cubic multiclass method one-vs-all & no data standardization. |

To make a fair comparison with the state-of-the-art results, we adopt the same experimental protocol as used in [1,10,11,14].

We use Cross-Subject validation technique to evaluate our classification performance in making predictions on new data not included in the training process. We apply Cross Subjects Leave-One-Out Cross validation CS-LOOCV to partition data into *k* subsets (or folds) each one containing actions performed by a single subject, where k is equal to the number of subjects in the data. One subset is used to validate the model trained using the remaining subsets. This process is repeated *k* times such that each subset is used exactly once for validation. The average cross-validation error is used as a performance indicator to avoid the overfitting problem.

## IV. Experiments

The algorithm is mainly written in Matlab. We use Image Acquisition Toolbox™, Image Processing Toolbox™ and Statistics and Machine Learning Toolbox™ and includes several C++ functions integrated as Mex files. The application runs on a Core i3 processor of 2.4 GHz and 6 GB of Ram, the average sampling rate of skeleton information is 10Hertz.

The experiment consists of several scenarios. It starts with an evaluation of discriminating strength of our descriptors using four classifiers. Then we study how many frames are necessary to obtain a good action recognition score. Finally, we investigate the importance of every skeleton joint in the action recognition accuracy.

We conduct our experimentation on two datasets, MSR Daily Activity 3D dataset and our own dataset specialized in working behaviors.

### A. Classification accuracy on MSR Daily Activity 3D dataset

First, we evaluate the performance of our enhanced descriptor for action recognition task on MSR Daily Activity 3D dataset [14] The dataset is a challenging benchmark for RGB-D action recognition providing 16 actions types chosen to cover human daily activities in the living room including: drink, eat, read book, call cellphone, write on a paper, use laptop, use vacuum, cheer up, sit still, toss paper, play game, lay down on sofa, walk, play guitar, stand up, sit down.

Ten subjects performed each action. Every subject performed an activity in two different poses: sitting and standing. The total number of the activity samples is 320. Some RGB and 3D skeleton information examples of activities are shown in Fig.4 and Fig.5 respectively. The dataset contains color RGB and Depth frames and 20 3D skeleton joints information extracted by the skeleton tracker from noisy depth and gave some missing skeleton data. We also note that the actions in this dataset are more complex and in general, they require interactions with objects. Thus, this dataset is more challenging.

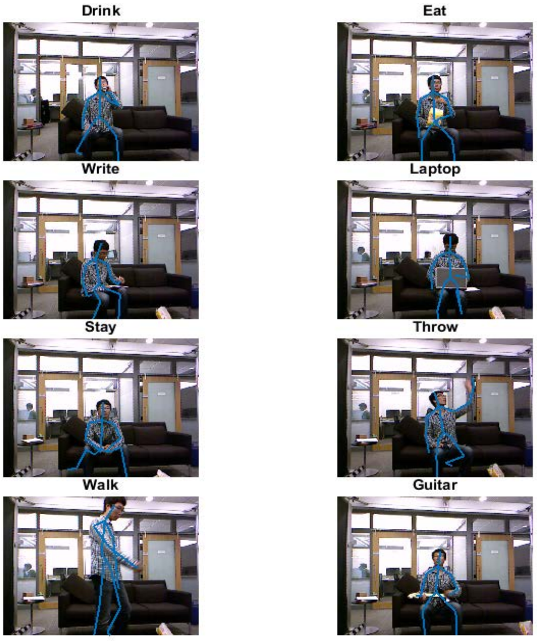


Fig.4. Color frame and skeleton examples in screen coordinates from MSR Daily Activity 3D dataset.

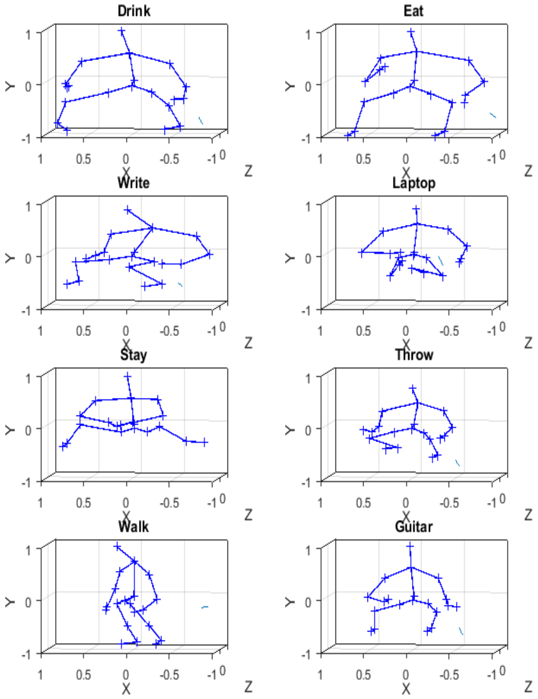


Fig.5. Skeleton examples in 3D coordinates from MSR Daily Activity 3D dataset.

Table 2. and Fig.6 (in graphic form) illustrate the accuracies, training and testing times of different methods. Both distance and spherical descriptors gives good scores separately with all the classifiers expect using decision tree. The best one is about 94.06% of good recognition rate using SVM, which is a good result considering the difficulties in this dataset. If we train the classifier on the distance/spherical descriptor, the best accuracy increases to 95.31% using the same classifier. We note also a good time parameter of all descriptors, which is less than few seconds knowing that dataset contains several thousands of frames to process. The minimum training time is given by SVM with less than 0.5 seconds. When we consider the overall performance, SVM is the best classifier.

Table 2. Accuracy and processing time evaluation of distance and spherical descriptors individually and together using four classifiers on MSR Daily Activity 3D dataset.

| Classifier | Parameter | Descriptor | | |
|---|---|---|---|---|
| | | Distance | Spherical | Both |
| Simple decision tree | Accuracy | 70.62 % | 75.00 % | 65.93 % |
| | Training time | 0.62 s | **0.36 s** | 0.53 s |
| | Testing time | **1.67 s** | **1.60 s** | **2.33 s** |
| Ensemble subspace discriminant | Accuracy | 89.37 % | 91.87 % | 87.81 % |
| | Training time | 2.31 s | 2.16 s | 2.33 s |
| | Testing time | 24.03 s | 22.62 s | 24.29 s |
| bagged trees | Accuracy | 91.56 % | 93.75 % | 92.18 % |
| | Training time | 9.87 s | 9.59 s | 10.54 s |
| | Testing time | 52.57 s | 48.63 s | 59.21 s |
| SVM | Accuracy | **94.06 %** | **94.06 %** | **95.31 %** |
| | Training time | **0.53 s** | 0.40 s | **0.49 s** |
| | Testing time | 2.14 s | 1.91 s | 2.51 s |

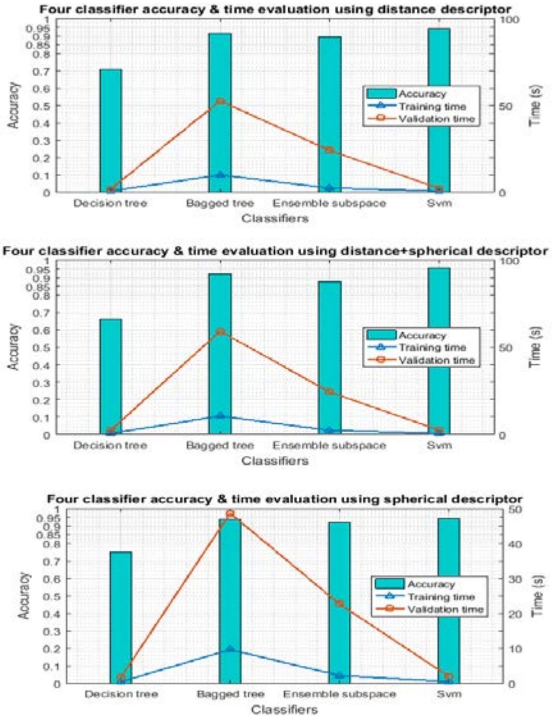


Fig.6. Accuracy and processing time evaluation of distance and spherical descriptors individually and together using four classifiers on MSR Daily Activity 3D dataset.

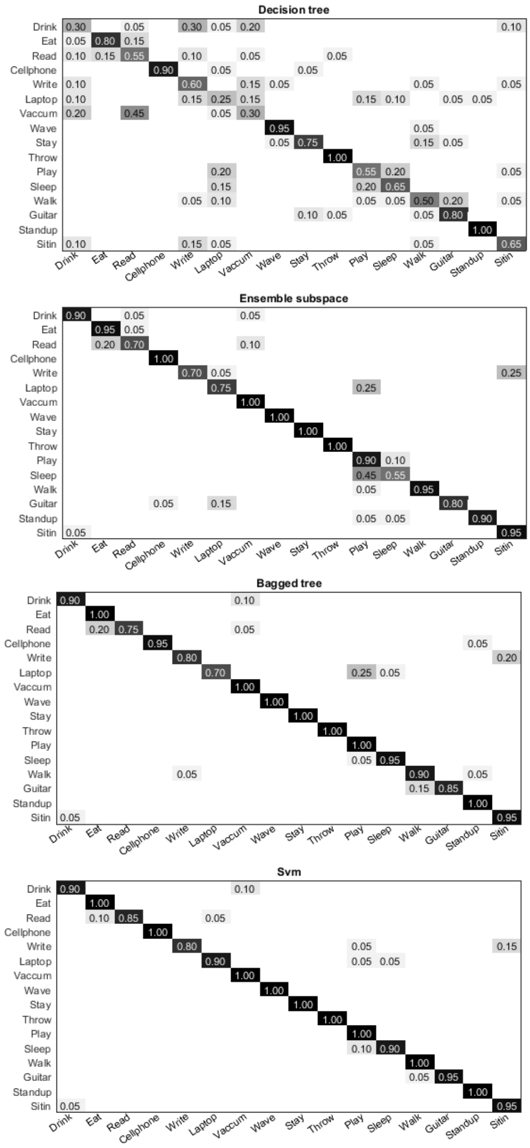


Fig.7. Confusion matrices of four classifiers using distance/spherical descriptor on MSR Daily Activity 3D dataset.

In Table 2. we compare the accuracy of the distance, spherical and distance/spherical descriptor. It is shown that the fusion of both descriptors consistently outperforms each of them, they contain complementary information and their fusion improves the action recognition performance.

Fig.7 shows the confusion matrices for the recognition method with a fusion of distance/spherical descriptor using four classifiers. In all the cases, good accuracy is achieved for all activities except using decision tree due to its simple decision mechanism.

Also, we compare our approach with the state-of-the-art methods using skeleton information including Eigen Joints [1], Bag-of-3D-points [10], HOJ3D [11] and Actionlet Ensemble [14] using the CS LOOCV protocol on the same dataset. Table 3. shows the overall accuracies. It reveals that our method significantly outperforms state-of-art.

Table 3. CS LOOCV recognition accuracies of our method and the state-of-the-art on MSR Daily Activity 3D.

| Methods | Accuracy score |
|---|---|
| Eigen Joints [1] | 83.3% |
| Bag-of-3D-points [8] | 74.7% |
| HOJ3D [9] | 79.0% |
| Actionlet Ensemble [12] | 85.7% |
| Our proposed approach | **95.3%** |

### B. *Classification accuracy on INDACT dataset*

This RGB-D database for human action recognition in the industrial field is proposed in [19] and called INDACT for INDustrial human ACTivity, including RGB, Depth and skeleton information for human action analysis.

INDACT mainly focuses on daily activities accomplished by human worker in industrial context. It contains 15 actions classified into three subsets following the danger level of each one.

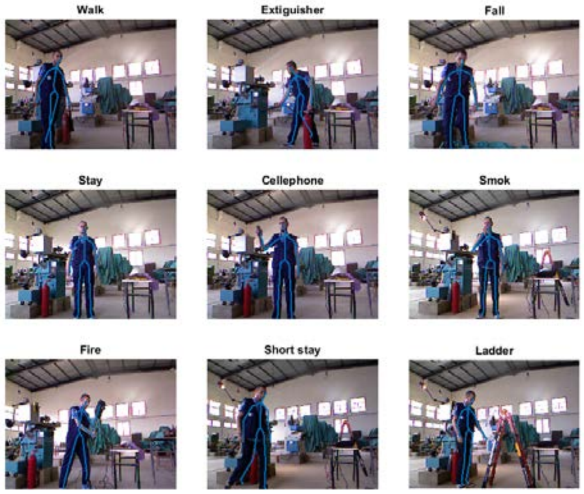


Fig.8. Color frame and skeleton examples in screen coordinates from INDACT dataset.

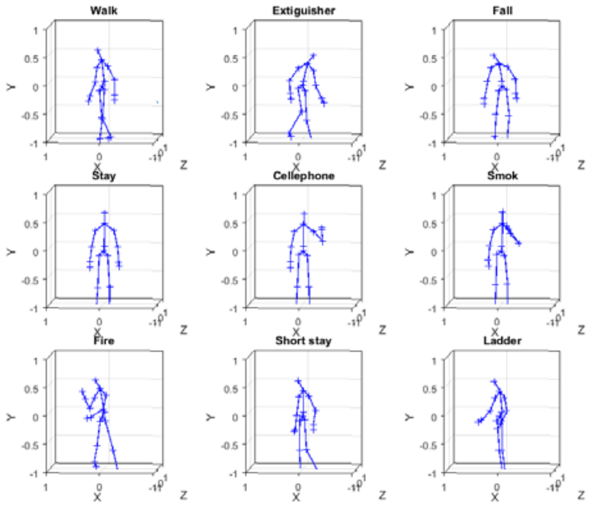


Fig.9. Skeleton examples in 3D coordinates from INDACT dataset.

The dataset includes 360-labeled video with both color, depth and 20 skeleton joint locations in screen and world coordinates with average duration of ten seconds for each action.

Some example frames from this dataset are illustrated in terms of both color and skeleton in screen coordinates in Fig.8 and in 3D coordinates in Fig.9.

Both Table 4. and Fig.10, present the performances of distance descriptor on INDACT dataset giving recognition accuracy of 97.50%. When we use spherical descriptor, we get more than 95.55% and we obtain 96.38% of good recognition score by mixing the two descriptors.

Table 4. Accuracy and processing time evaluation of distance and spherical descriptors individually and together using four classifiers on INDACT dataset.

| Classifier | Parameter | Descriptor | | |
|---|---|---|---|---|
| | | Distance | Spherical | Both |
| Simple decision tree | Accuracy | 84.16 % | 86.94 % | **90.00 %** |
| | Training time | 0.83 s | **0.79 s** | 2.04 s |
| | Testing time | **3.33 s** | 3.50 s | 8.54 s |
| Ensemble subspace discriminant | Accuracy | **97.50 %** | 93.05 % | 94.16 % |
| | Training time | 2.32 s | **2.14 s** | 2.19 s |
| | Testing time | 22.55 s | **21.43 s** | 22.11 s |
| bagged trees | Accuracy | 93.88 % | **95.00 %** | 94.72 % |
| | Training time | 11.59 s | **11.32 s** | 14.27 s |
| | Testing time | 61.71 s | **59.86 s** | 71.57 s |
| SVM | Accuracy | **97.50 %** | 95.55 % | 96.38 % |
| | Training time | **0.41 s** | 0.42 s | 0.69 s |
| | Testing time | **2.14 s** | 2.20 s | 3.29 s |

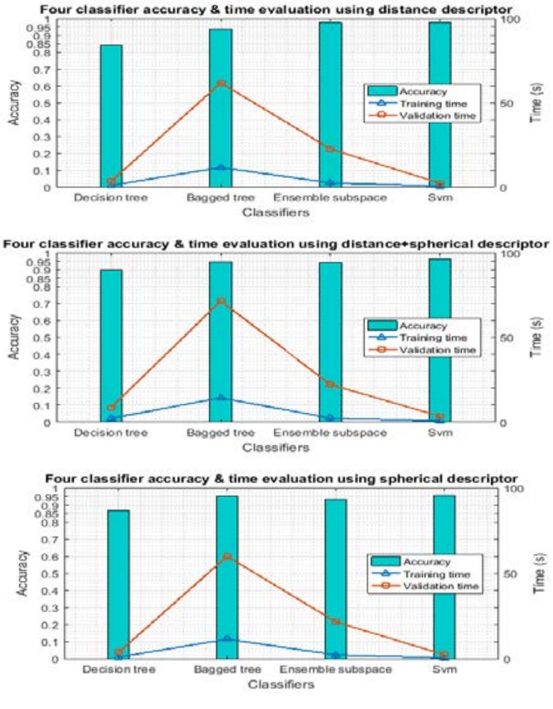


Fig.10. Accuracy and processing time evaluation of distance and spherical descriptors individually and together using four classifiers on INDACT dataset.

The confusion matrix of distance/spherical descriptor in Fig.11 proves the good accuracy score using four classifiers. We confirm the good performance of all descriptors as well as good training and testing time.

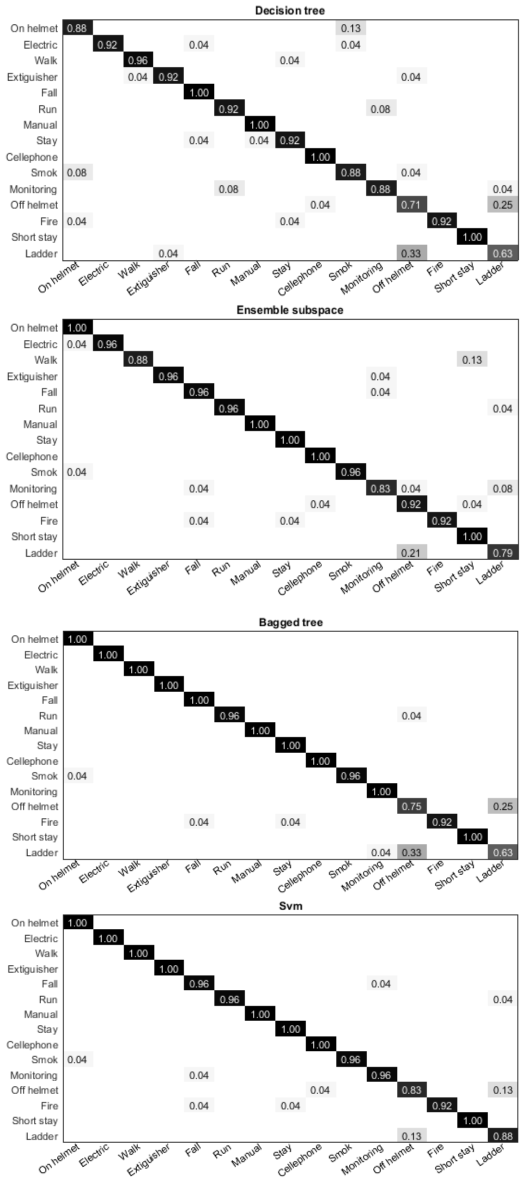


Fig.11. Confusion matrices of four classifiers using distance/spherical descriptor on INDACT dataset.

## C. *How many frames are sufficient to get good detection?*

For monitoring security of human workers in industry, we need to know the danger level of performing actions in real time. This kind of system is affected by two main factors. First, the time taken to observe sufficient frames for making a good prediction. Second the time to make a good decision.

We reduce the number of frames needed to extract enough discriminating descriptor, and then we use a quick classification technique. To evaluate the minimum frame number necessary to conduct a proper recognition we built distance/spherical descriptor for each frame number, and then we evaluate recognition score for every one of them on both MSR Daily Activity 3D and INDACT dataset as illustrated in Fig.12 and Fig.13.

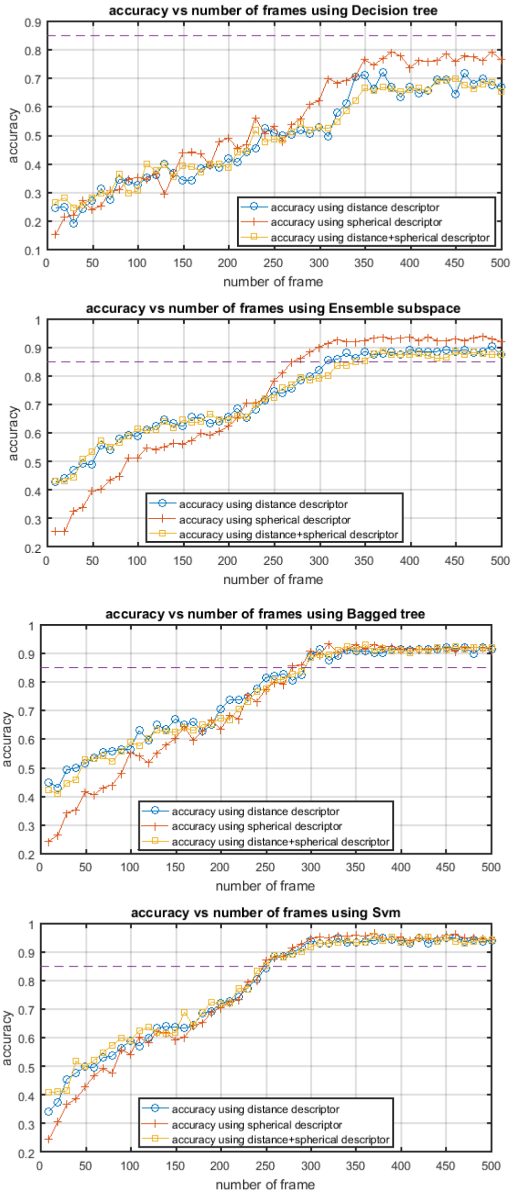


Fig.12. Accuracy vs frame number using distance/spherical descriptor on MSR Daily Activity 3D dataset.

As shown in Fig.13 with INDACT dataset, in most cases just the first 25–40% frames are sufficient to achieve good recognition (>85%), when on MSR Daily Activity 3D dataset, we need to include more than half of frames to get a comparable score due to noisy skeleton information in this dataset (Fig.12).

Passing this number of frames, the gains decrease by adding more frames to descriptor calculation process. These results are highly important for activity recognition systems when decisions have to be made before the happening of the entire action to prevent dangerous situations.

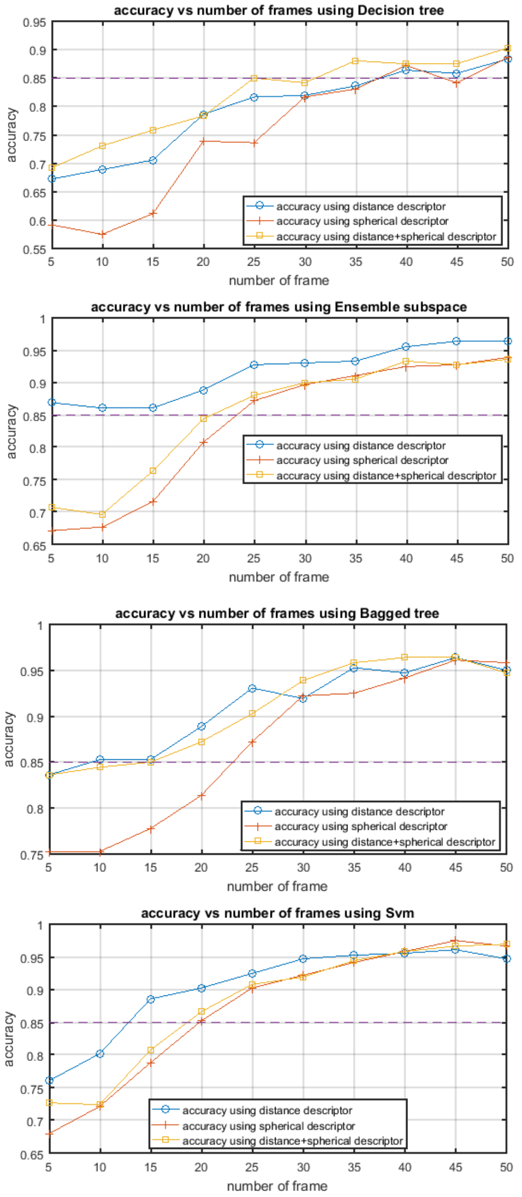


Fig.13. Accuracy vs frame number using distance/spherical descriptor on INDACT dataset.

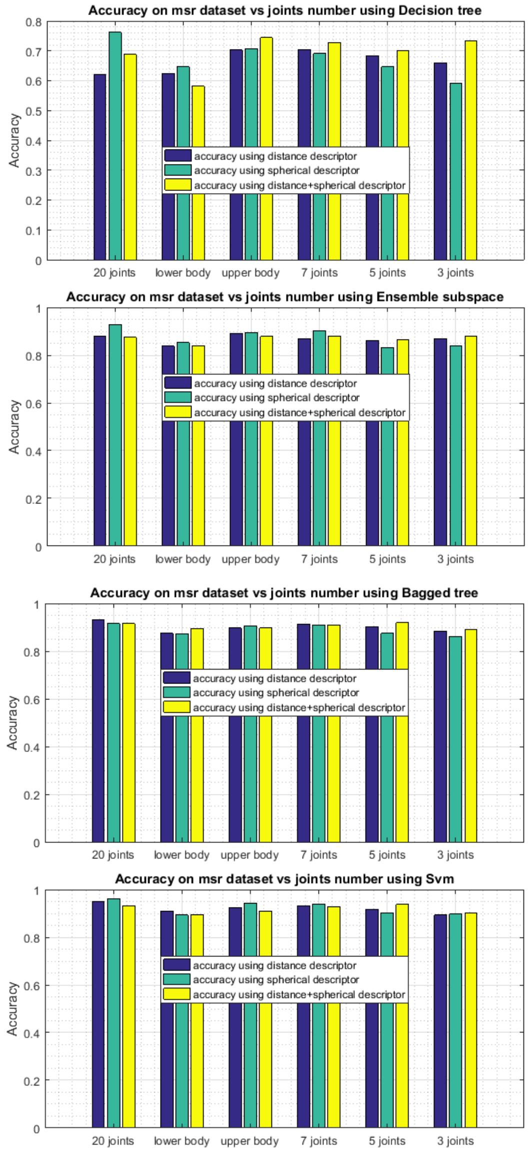
### D. *Are all joints important for good detection?*

Microsoft Kinect in its first version provides 20 skeleton joints in each frame. To speed up the recognition by decreasing the amount of information included in the descriptor, we test the importance of each one of those joints in action recognition process. We compute the distance/spherical descriptor using various numbers of joints including:

- 3 joints = {head, left hand, right hand};
- 5 joints = {head, left hand, right hand, left foot, right foot};
- 7 joints = {head, left shoulder, left hand, right shoulder, right hand, left foot, right foot};

- Upper body joints = {center hip, spine, center shoulder, head, left shoulder, left elbow, left wrist, left hand, right shoulder, right elbow, right wrist, right hand};
- Lower body joints = {center hip, spine, left hip, left knee, left ankle, left foot, right hip, right knee, right ankle, right foot};
- 20 joints = {center hip, spine, center shoulder, head, left shoulder, left elbow, left wrist, left hand, right shoulder, right elbow, right wrist, right hand, left hip, left knee, left ankle, left foot, right hip, right knee, right ankle, right foot};

Fig.14 and Fig.15 show accuracy scores on both datasets. We notice that every joints combination performs well; even the combination which uses only three skeletal joints, including the head and the two hands gives a close score if compared to the best one.



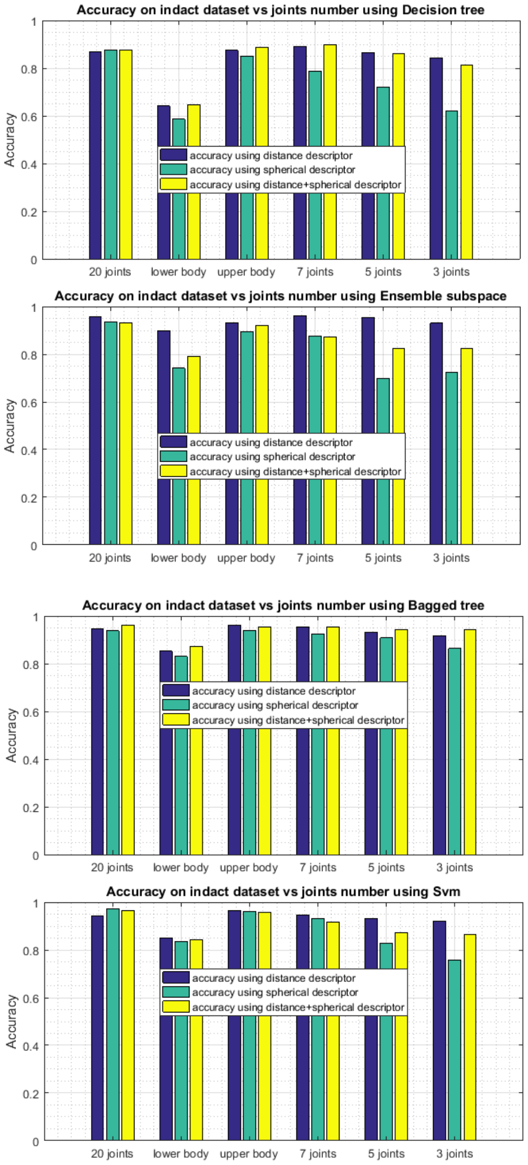
Accuracy on msr dataset vs joints number using Decision tree

Accuracy on msr dataset vs joints number using Ensemble subspace

Accuracy on msr dataset vs joints number using Bagged tree

Accuracy on msr dataset vs joints number using Svm

Fig.14. Accuracy vs joints number on MSR Daily Activity 3D dataset.



Accuracy on indact dataset vs joints number using Decision tree

Accuracy on indact dataset vs joints number using Ensemble subspace

Accuracy on indact dataset vs joints number using Bagged tree

Accuracy on indact dataset vs joints number using Svm

Fig.15. Accuracy vs joints number on INDACT dataset.

Using only three joints to compute the descriptor gives an accuracy of 94.44% using bagged tree on INDACT dataset and about 90.31% using SVM on MSR Daily Activity 3D dataset compared to the best score obtained by including all available joints by 96.38% of good classification using SVM on INDACT dataset and 95.31% using SVM on MSR Daily Activity 3D dataset. In same time we reduce significantly the complexity of the descriptor features vector that generate only $3*3^2+2*3*3^2 = 81$ features compared to $3*20^2+2*3*20^2 = 3600$ using all available joints, which is important for real time applications.

## V. Conclusion and Future Work

We introduced a new action recognition method to warn the occurrence of dangerous situations in an industrial context, we use a compact and light descriptor built from few first frames and few skeleton joints to categorize each action. We observed that only the first few frames and a limited number of joints are sufficient to make good decisions proved on our own specific dataset INDACT as well as on a public dataset MSR Daily Activity 3D.

In future work, we plan to study the interaction between human and tools in industrial environment and try to take benefits from object identification techniques to improve the action recognition process and to decrease the detection time. We are also interested by proposing a body gesture detection system [20] to analyze workers' degree of compliance with regulation and the impact on productivity and health especially during the execution of repetitive tasks.

Another way to improve this work is to consider exploring new machine learning approach like deep learning techniques specially adapted to process datasets that contains a large amount of data like in NTU RGB+D [21] that contains over than 56 000 actions. The important volume of data needs to take advantage of the computing power offered by GPUs and HPC. This orientation towards graphical units is particularly adapted while using datasets captured by Kinect sensor in its second version able to provide RGB-D images in high definition and more than 25 skeleton joints as well as infrared images that can be used for fire or heat detection in challenging vision conditions (darkness, smoke).

## References

[1] X. Yang and Y. Tian, “Effective 3D Action Recognition Using Eigenjoints,” Journal of Visual Communication and Image Representation, Vol. 25, No. 1, pp. 2-11, 2014, doi: 10.1016/j.jvcir.2013.03.001.

[2] O. Maksymiv, T. Rak and D. Peleshko, “Video-based Flame Detection using LBP-based Descriptor: Influences of Classifiers Variety on Detection Efficiency,” International Journal of Intelligent Systems and Applications, Vol.9, No.2, pp. 42-48, 2017.

[3] K.C. Manjunatha, H.S. Mohana and P.A Vijaya, “Implementation of Computer Vision Based Industrial Fire Safety Automation by Using Neuro-Fuzzy Algorithms,” International Journal of Information Technology and Computer Science, vol.7, no.4, pp.14-27, 2015.

[4] G. Cheng, Y. Wan, A. Saudagar, K. Namuduri and B. Buckles, “Advances in human action recognition: A survey,” In preprint arXiv: 1501.05964, 2015. https://arxiv.org/abs/1501.05964.

[5] T. Hassner, “A Critical Review of Action Recognition Benchmarks,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPR), pp. 245-250, 2013, doi: 10.1109/CVPRW.2013.43.

[6] L. Presti and M. La Cascia, “3D Skeleton-based Human Action Classification: Survey,” Pattern Recognition, Vol. 53, pp. 130-147, 2016, doi: 10.1016/j.patcog.2015.11.019.

[7] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman and A. Blake, “Real-time human pose recognition in parts from single depth images,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1297-1304, 2011, doi: 10.1145/2398356.2398381.

[8] R. Gishick, J. Shotton, P. Kohli, A. Criminisi and A. Fitzgibbon, “Efficient regression of general activity human poses from depth images,” Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 415-422, 2011, doi: 10.1109/ICCV.2011.6126270.

[9] M. Sun, P. Kohli and J. Shotton, “Conditional regression forests for human pose estimation,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3394-3401, 2012, doi: 10.1109/CVPR.2012.6248079.

[10] W. Li, Z. Zhang and Z. Liu, “Action recognition based on a bag of 3D points,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 9-14, 2010, doi: 10.1109/CVPRW.2010.5543273.

[11] L. Xia, C. Chen and J. Aggarwal, “View invariant human action recognition using histograms of 3D joints,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 20-27, 2012, doi: 10.1109/CVPRW.2012.6239233.

[12] J. Sung, C. Ponce, B. Selman and A. Saxena, “Unstructured human activity detection from RGBD images,” Proceedings of IEEE International Conference on Robotics and Automation (ICRA), pp. 842-849, 2012, doi: 10.1109/ICRA.2012.6224591.

[13] X. Yang, C. Zhang and Y. Tian, “Recognizing actions using depth motion maps-based histograms of oriented gradients,” Proceedings of the 20th ACM international conference on Multimedia, pp. 1057-1060, 2012, doi: 10.1145/2393347.2396382.

[14] J. Wang, Z. Liu, Y. Wu and J. Yuan, “Mining Actionlet ensemble for action recognition with depth cameras,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1290-1297, 2012, doi: 10.1109/CVPR.2012.6247813.

[15] H. Wu, W. Pan, X. Xiong and S. Xu, “Human activity recognition based on the combined SVM&HMM,” Proceedings of IEEE International Conference on Information and Automation (ICIA), pp. 219–224, 2014, doi: 10.1109/ICInfA.2014.6932656.

[16] A. Eweiwi, M. Cheema, C. Bauckhage and J. Gall, “Efficient pose-based action recognition,” Asian Conference on Computer Vision (ACCV), pp. 428-443, 2014, doi: 10.1007/978-3-319-16814-2_28.

[17] K. Schindler and L. Gool, “Action snippets: how many frames does human action recognition require? Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1-8, 2008, doi: 10.1109/CVPR.2008.4587730.

[18] M. Zanfir, M. Leordeanu and C. Sminchisescu, “The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection,” Proceedings of IEEE International Conference on Computer Vision (ICCV), pp. 2752-2759, 2013, doi: 10.1109/ICCV.2013.342.

[19] A. Kerboua, M. Batouche and A. Debbeh, “RGB-D & SVM action recognition for security improvement,” Proceedings of the ACM Mediterranean Conference on Pattern Recognition and Artificial Intelligence (MedPRAI), pp. 137-143, 2016, doi: 10.1145/3038884.3038907.

[20] J. Medina-Catzin, A. Gonzalez, C. Brito-Loeza, V. Uc-Cetina, “Body Gestures Recognition System to Control a Service Robot,” International Journal of Information Technology and Computer Science, Vol.9, No.9, pp. 69-76, 2017.

[21] A. Shahroudy, J. Liu, T. Ng and G. Wang, “NTU RGB+D: A Large-Scale Dataset for 3D Human Activity Analysis,” Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1010-1019, 2016, doi: 10.1109/CVPR.2016.115.

## Authors’ Profiles

**Adlen Kerboua** is a Ph.D. candidate in the Department of Computer Science at University of Constantine 2 - A. Mehri, Algeria. In the field of artificial intelligence. He received Engineer degree in computer science from Constantine University (Algeria) in 1999, a Master degree in intelligent systems from Savoy University (France) in 2004 and the degree of magister in artificial intelligence from Tebessa University (Algeria) in 2011. His research interests are Vision, Action Recognition and Artificial Intelligence.

**Mohamed Batouche**, Ph.D. Is full Professor and head of the Department of Computer Science at University of Constantine 2 - A. Mehri, Algeria. He received Engineer degree in Computer Science from Constantine University (Algeria), MSc and PhD degrees in Computer Science from Nancy- INPL University (France). He published over then 250 papers in refereed international journals and conferences. His research interests are in the field of Complex Systems, Nature Inspired Computing, Bioinformatics, Machine Learning, Deep Learning, Cloud Computing, and Big Data.

**How to cite this paper:** Adlen Kerboua, Mohamed Batouche, "3D Skeleton Action Recognition for Security Improvement", International Journal of Intelligent Systems and Applications(IJISA), Vol.11, No.3, pp.42-52, 2019. DOI: 10.5815/ijisa.2019.03.05