Features and Classification Schemes for View-Invariant and Real-Time Human Action Recognition

Article *in* IEEE Transactions on Cognitive and Developmental Systems · June 2018

DOI: 10.1109/TCDS.2018.2844279

5 authors, including:

Sid Ahmed Walid Talha
IMT Lille Douai
3 PUBLICATIONS **1** CITATION
SEE PROFILE

Hammouche Mounir
University of Franche-Comté
8 PUBLICATIONS **9** CITATIONS
SEE PROFILE

Enjie Ghorbel
University of Luxembourg
14 PUBLICATIONS **23** CITATIONS
SEE PROFILE

Anthony Fleury
IMT Lille Douai
89 PUBLICATIONS **1,549** CITATIONS
SEE PROFILE

Features and Classification Schemes for View-Invariant and Real-Time Human Action Recognition

Sid Ahmed Walid Talha[1], Mounir Hammouche[1,2], Enjie Ghorbel[1,3], Anthony Fleury*[1], *Member, IEEE,*, Sébastien Ambellouis[1,4]

***Abstract*—Human Action recognition (HAR) is largely used in the field of Ambient Assisted Living (AAL) to create an interaction between humans and computers. In these applications, it cannot be asked to people to act non-naturally. The algorithm has to adapt and the interaction has to be as quick as possible to make this interaction fluent. To improve the existing algorithms with regards to that points, we propose a novel method based on skeleton information provided by RGB-D cameras. This approach is able to carry out early action recognition and is more robust to viewpoint variability. To reach this goal, a new descriptor called Body Directional Velocity is proposed and a real-time classification is performed. Experimental results on four benchmarks show that our method competes with various skeleton-based HAR algorithms. We also show the suitability of our method for early recognition of human actions.**

***Index Terms*—Human action recognition, Human Robot Interaction, skeleton analysis, Body-part Directional Velocity, Hidden Markov Model, Gaussian Mixture Model.**

I. Introduction

According to the recent study proposed by the United Nations, the number of elderly persons is expected to more than triple in 2050 [1]. Therefore, it is more important than ever to design intelligent systems in the field of Ambient Assisted Living (AAL) that aim at offering better quality of life and autonomy to elderly people. Such systems often require a stage of Human Action Recognition (HAR).

Generally, action recognition is considered as the association of two main parts: action description and action classification. While action description aims at extracting the discriminative information of motion from data captured by a sensor, action recognition deals with machine learning approaches in order to estimate models able to attribute correct labels to actions. As it can be noted, the sensor has a major influence on the performances of the system.

For example, Zhu et al. [2] have introduced a human-robot interaction system for elderly and disabled persons to perform the recognition of five different hand gestures and four daily activities. This method is based on wearable sensors attached to one foot and to the waist of the participants. Neural networks have been used for gesture segmentation and Hierarchical Hidden Markov Models (HHMMs) have been trained for the classification task.

In [3], the same kind of sensors worn on the wrist have captured the temperature, the altitude and the acceleration of body-parts. Based on previous studies, features from both time and frequency domains have been calculated. Furthermore, a sliding window has been used for segmentation. Neural networks combined with Support Vector Machines (SVM) have carried out the classification. Wearable sensors provide motion information as long as the user is wearing them. Nevertheless, in real-world scenarios, it is hard to ensure that people wear them continuously. In this context, using a camera is a relevant solution to this uncertainty.

Recently, the emergence of RGB-D sensors such as Microsoft Kinect has renewed the interest of researchers. Further to RGB images, these low-cost sensors provide additional information called depth maps. Moreover, the algorithm introduced by Shotton et al. [4] allows a real-time skeleton extraction from these depth maps that can be integrated to the camera. In this way, RGB-D based action recognition methods can be divided into three groups according to the used information: depth-based, skeleton-based and hybrid methods. Various methods belonging to these groups have been proposed in the state-of-the-art [5]–[12], showing their efficiency in terms of accuracy of recognition.

In the context of AAL, the accuracy is not the only constraint to deal with. Other challenges exist including 1) the robustness to the camera viewpoint variations – the action recognition system should not be affected by the human body orientation changes, 2) the ability to recognize an action as soon as possible i.e. before its end, 3) the computation speed.

This paper describes a novel method for fast, viewpoint invariant and early action recognition. To do that, we have proposed a novel descriptor based on the hierarchical information of the algebraic velocity of skeleton joints, called Body-part Directional Velocity (BDV). Then, an HMM with GMM state-output distributions is used for the classification task. The efficiency of this method is proven by the experiments realized on four well-known and publicly available benchmarks (MSRAction3D [13], UTKinect [14], Florence 3D [15] and Multiview3D [16]). The results show that our method

* Correspondong author, e-mail: anthony.fleury@imt-lille-douai.fr, Tel: +33 3 27 71 23 81.

[1] IMT Lille Douai, Univ. Lille, Unité de Recherche Informatique Automatique (Computer Sciences and Automatic Control Dpt. – URIA), F-59000 Lille, France

[2] Institut FEMTO-ST, Université de Franche-Comté

[3] Normandie Univ, UNIROUEN, ESIGELEC, IRSEEM, 76000 Rouen, France

[4] IFSTTAR, Cosys-Leost, Villeneuve d'Ascq, France.

Published on Jun. 5, 2018. Author preprint. Final version is available at https://doi.org/10.1109/TCDS.2018.2844279

competes with skeleton-based state-of-the-art methods in terms of accuracy for the four previously mentioned datasets. Furthermore, the proposed approach is able to provide its final decision using less than 50% of the total number of frames of each action sequence.

This paper is organized as follows: Section 2 presents a state-of-the-art of action recognition methods developed in the context of Human Computer Interaction. Section 3 introduces the approach designed for viewpoint invariant and early action recognition. Then, Section 4 presents the experiments realized on the four benchmarks. Finally, Section 5 concludes this work and presents future works.

II. Related work

Human action recognition is a very popular topic in the researcher community. As it has been written previously, the proposed approaches can use depth, skeleton information or can benefit from both jointly. The literature states clearly that depth-based methods offer accurate performances on reference datasets. But if skeleton-based methods improve the robustness to viewpoint variability, they fail when actions have little motion differences that the extracted skeleton cannot catch. Thus, to gather both skeleton and depth information seems to be a good solution if an optimal combination of features is defined.

Li et al. [13] have presented one of the first action recognition method using depth images. They proposed an action graph to model the dynamics of actions and a bag of 3D points to describe salient postures that correspond to the nodes in the action graph. Gaussian Mixture Models (GMMs) have been employed to capture the statistical distribution of features. In [6], 4D histograms over depth, time and space have been used to capture the changes of human body normal orientation (HON4D). Chen et al.[17] extracted the features using the Depth Motion Maps (DMMs). Each depth frame has been projected onto three orthogonal Cartesian planes. For each projection, the absolute difference between two consecutive projected images has been accumulated through an entire depth video. In [5], Ohn-Bar et al. built a depth-based descriptor by calculating two histograms according to space and time (HOG2). The classification is performed using linear SVM. Approaches based on depth have to deal with an important amount of data involving an expensive computation in both training and validation stages.

Using Shotton et al.'s algorithm [4], skeleton information can be extracted in real-time from depth maps. Based on this low-dimensional data, many recent skeleton-based approaches have shown their ability to recognize actions. In [14], a spatial histogram of joint locations has been computed and a linear discriminant analysis has been performed to extract the dominant features. A posture vocabulary has been constructed by clustering the histograms with the use of the k-means algorithm. The temporal evolution of postures has been modeled using Discrete Hidden Markov Models (DHMMs). Du et al. [18] proposed a Hierarchical Bidirectional Recurrent Neural Network (HBRNN) to classify human actions. They divided the skeleton into five groups of joints representing two arms, two legs and one trunk. After that, each group has been used to feed five BRNNs. The generated hidden states have been combined and introduced into another set of BRNNS. The results represents the input of the next layer. In [19], the feature extraction has involved normalization of the skeleton using the Euclidean distance between the torso and the neck joints. Then, the k-means algorithm has been applied to group similar postures into clusters, and a vector containing the centroids of each cluster has been created. Finally, a multiclass Support Vector Machine (SVM) has been applied to identify human actions. In [8], Devanne et al. have computed the similarity between shapes of skeleton joint trajectories in a Riemannian framework. Classification has been performed using a k-NN classifier. Miranda et al. [20] introduced a method for real-time gesture recognition. This method proposed to represent each pose by the spherical coordinates of skeleton joints. A multiclass SVM classifier with a tailored pose kernel has been used to identify key poses while a random forest is used to recognize gestures. In [7], Relative Joint Positions (RJP) have been used to describe the skeleton motion. The temporal evolution of these joints has been compared using dynamic time warping (DTW). Then, a Fourier Temporal Pyramid (FTP) has been applied to obtain the final descriptor. Action classification has been performed using a linear SVM classifier. Finally, in [21], the authors define a new local skeleton descriptor that encodes the relative position of joint quadruples and yield a high performing classifier.

In [22], the authors use a set of random forests to fuse the spatio-temporal depth and joints features. Wang et al. [23] compute the histogram of occupancy patterns of a fixed region around each joint in each frame of an action video and use low temporal frequency Fourier components as features to classify the actions. Recently, Shahroudy et al. [24] proposed hierarchical mixed norms to fuse different features and to select the most informative body joints. More recently, the authors of [25] have proposed a more robust method to significant viewpoints changes. The method is modeling the temporal relations between human body-parts and the objects of the environment.

In a previous work [26], we have proposed to gather depth and skeleton streams through a depth estimation of the body orientation and a fuzzy reasoning to decide the better skeleton-based SVM for action recognition (depending on the orientation). The method has been evaluated on Multiview3D dataset [16] specifically designed to evaluate the performances of the algorithm in such conditions. Even if the results of the method were promising, it appears that the skeleton-based part of the algorithm can be improved by proposing a more representative feature vector without taking into account depth information. Several feature vector improvements have been proposed in previous works and we propose a new efficient one in this article. This new feature is called Body-part Directional Velocity descriptor (BDV). Our proposed method exploits this feature combined to a GMM-based HMM classifier. It yields two improvements. Firstly, our method outperforms the well-known skeleton-based techniques and secondly it carries out early action recognition by attributing a label to an instance for each frame before the end of the action.

III. Overview of the Classification Approach

In this section, the proposed method is described in details. As the majority of action recognition methods, the proposed method is composed of two major steps: Action description and Action recognition. Figure 1 presents an overview of this method. To describe actions, a novel frame-by-frame human action descriptor is first introduced, called Body-part Directional Velocity (BDV) which is calculated based on the algebraic value of different body-part velocity. For the classification, a GMM-based HMM classifier is used considering BDV as input, allowing us to obtain an early action recognition system. In the following, these two steps are respectively described in Section III-A and Section III-B.

A. Body-part Directional Velocity (BDV)

This frame-by-frame descriptor has been designed for its suitability for early action recognition. Indeed, to build BDV, a prior knowledge of the whole sequence is not needed. In this section, we detail the calculation of BDV.

A skeleton sequence $\mathbf{p}$ represents a series of N temporal ordered poses as described by Equation (1) (such that p_t refers to the skeleton pose at a given time t).

$$\mathbf{p} = [\mathbf{p}_1, \mathbf{p}_2, ..., \mathbf{p}_t, ..., \mathbf{p}_N] \tag{1}$$

At each time t, $\mathbf{p}_t$ is composed of a set of n joint positions, as described by Equation (2) (such that p_t^i is the ith joint positions at a time t).

$$\mathbf{p}_t = [\mathbf{p}_t^1, \mathbf{p}_t^2, ..., \mathbf{p}_t^i, ..., \mathbf{p}_t^n] \tag{2}$$

First, since the 3D skeleton data are not always accurate due to the noise and the occlusions, a pre-processing of smoothing is carried out. Therefore, a Savitzky-Golay filter is applied to all joint positions as follows: $\forall (i,t) \in [\![1,n]\!] \times [\![1,N]\!]$,

$$\mathbf{P}_t^i = \frac{1}{35}(-3\mathbf{p}_{t-2}^i + 12\mathbf{p}_{t-1}^i + 17\mathbf{p}_t^i + 12\mathbf{p}_{t+1}^i - 3\mathbf{p}_{t+2}^i) \tag{3}$$

where $\mathbf{P}_t^i$ refers to the position of the joint i at a time t after the filtering process.

Then, the velocity at a time t of each joint i, considered as a very discriminative feature, is computed as in [12] using Equation (4).

$$\mathbf{V}_t^i = \mathbf{P}_{t+1}^i - \mathbf{P}_{t-1}^i \tag{4}$$

Since different motions imply the movement of different joints, we propose to divide the human body into five body-parts, namely, left arm (B_1), right arm (B_2), left leg (B_3), right leg (B_4) and spine (B_5), as illustrated in Figure 2. The set of body-parts is therefore denoted by $B = \{B_1, B_2, B_3, B_4, B_5\}$.

Then, for every body-part, the negative and positive velocities of associated joints are summed and respectively denoted by $D_{B_l}^+$ and $D_{B_l}^-$ in Equation (5) and Equation (6). The separation of negative and positive values is very informative since it is related to the direction of the motion.

$$\mathbf{D}_{B_l}^+(t) = \sum_{i \in B_l} (\mathbf{V}_t^i \geq 0) \tag{5}$$

$$\mathbf{D}_{B_l}^-(t) = \sum_{i \in B_l} (\mathbf{V}_t^i < 0) \tag{6}$$

The final descriptor obtained at a time t, denoted by $\mathbf{D}(t)$ is obtained following Equation (7).

$$\mathbf{D}(t) = \bigcup_{l=1}^{5} [\mathbf{D}_{B_l}^+(t), \mathbf{D}_{B_l}^-(t)] \tag{7}$$

Therefore, the size of the proposed BDV descriptor is equal to $d_D = 30$

B. Classification using HMM based on GMM

As previously explained, we compute Body-part Directional Velocity features which are adapted to early recognition of human action. In this work, HMMs with GMMs state-output distributions (illustrated in Figure 3) are used to achieve our goal. A Hidden Markov Model (HMM) [27] is a statistical model used to describe the evolution of observable events. It is especially used to model time sequential data for speech, gesture and activity recognition. HMM is based on two stochastic

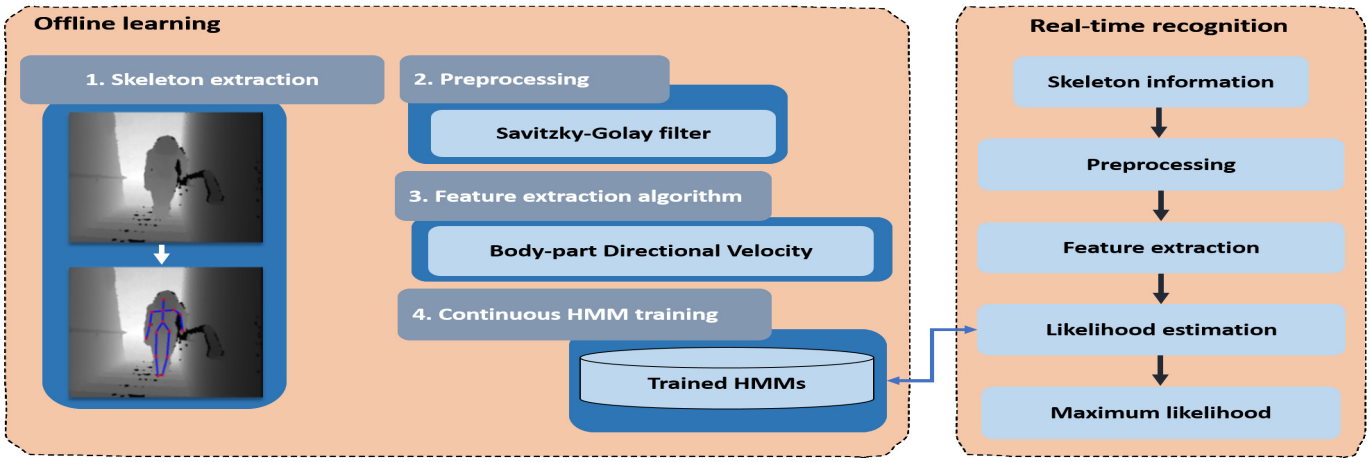


Fig. 1: Overview of the proposed system

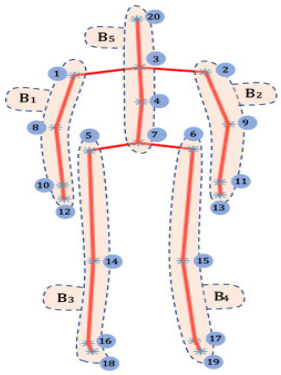


Fig. 2: Human body divided into five different body-parts (B_1, B_2, B_3, B_4 and B_5)

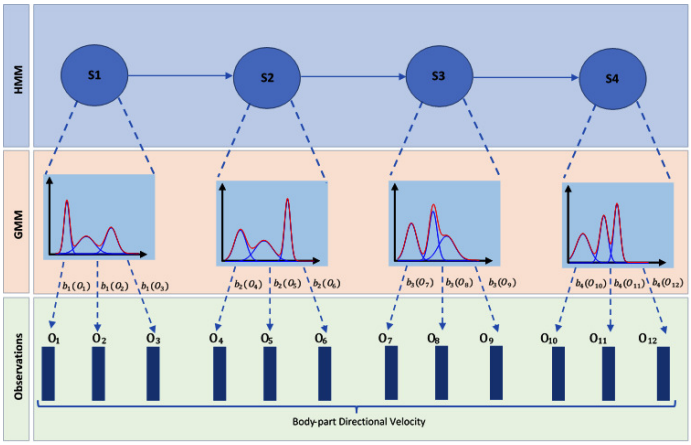


Fig. 3: Architecture of the used GMM-based HMM system

processes. The first process is an observable process which represents the sequence of observed symbols. The second one is unobservable (hidden) and can be indirectly inferred by analyzing the sequence of observed symbols. In this work, an HMM is learned for each action a.

For each HMM^a learned for the action a, let us denote by:

- N^a the number of states in the model,
- M the number of observation symbols,
- $S^a = \{s_1^a, s_2^a, ..., s_N^a\}$ the set of distinct states,
- $V = \{v_1, v_2, ..., v_M\}$ the observation alphabet,
- $Q^a = \{q_1^a, q_2^a, ..., q_T^a\}$ the T states from S^a,
- $O = \{o_1, o_2, ..., o_T\}$ the T observations from the alphabet V corresponding to Q^a states.

Each HMM^a can be written in a compact form and denoted by $\lambda^a = (\pi^a, \mathbf{A}^a, \mathbf{B}^a)$.

π^a is the vector of initial state distribution:

$$\pi^a = \{\pi_i\}, \pi_i = P(q_1 = s_i)_{1 \leq i \leq N^a} \quad (8)$$

$\mathbf{A}^a$ is the matrix of state transition probability distribution:

$$\mathbf{A}^a = \{a_{ij}\}, a_{ij} = P(q_{t+1} = s_j | q_t = s_i)_{1 \leq i,j \leq N^a} \quad (9)$$

such that a_{ij} represents the transition probability from state i to state j.

$\mathbf{B}^a$ is the matrix of observation symbol probability distribution:

$$\mathbf{B}^a = \{b_{ik}\}, b_{ik} = P(o_t = v_k | q_t = s_i)_{1 \leq i \leq N^a, 1 \leq k \leq M} \quad (10)$$

such that b_{ik} represents the probability of the k^{th} observation realization from the state i.

The Discrete HMM (DHMM) considers that the observations are discrete symbols from a finite alphabet. Therefore the extracted features are quantized by using unsupervised classification algorithm. In [14], the vector quantization is performed by clustering the features into k clusters using k-means algorithm. The symbol number and the centroid of each cluster form a codebook. The vector quantization involves the degradation of the model, leading to poor accuracy. To overcome this problem, continuous probability distribution functions are used to model Body-part Directional Velocity descriptors as depicted by Equation (11),

$$b_{ik}(o_t) = \sum_{r=1}^{N_g} w_{ir} g_r(o_t, \mu_{ir}, C_{ir})_{1 \leq i \leq N^a, 1 \leq k \leq M} \quad (11)$$

such that w_{ir}, μ_{ir} and C_{ir} respectively represent the weight, the mean vector and the covariance matrix of the r^{th} Gaussian model in the state i.

N_g is the number of mixture densities. In our experiments, we fix it empirically to $N_g = 3$. We recall that d_D is the dimension of the descriptor BDV.

The probability density function employed is a mixture of multivariate Gaussian (GMMs), where each one is defined as follows:

$$g_r(o_t, \mu_{ir}, C_{ir}) = \frac{1}{(2\pi)^{d_D/2} |C_{ir}|^{1/2}} e^{\frac{-1}{2}(o_t - \mu_{ir})^T C_{ir}^{-1} (o_t - \mu_{ir})} \quad (12)$$

As specified before, HMM^a is separately trained for each action a.

The likelihood estimation of the feature vector sequence is calculated for each HMM^a, at each time t using the forward algorithm. Then the HMM presenting the highest probability is selected to get the correct label a^*, as described by Equation (13).

$$a^*(t) = \arg\max_{a \in A} (P(o_t | \lambda_i)) \quad (13)$$

IV. Experiments

In this part, we present the experiments performed on two well-known benchmarks, namely, MSRAction3D and Florence 3D. We show the effectiveness of our approach, not only in terms of rapidity of calculation and accuracy but also in terms of observational latency (the necessary time of observation required DHMM to recognize each action). Experiments are also conducted on datasets containing different human body orientations, namely, UTKinect-Action and Multiview3D. The

obtained results demonstrate the robustness of our method to viewpoint variation.

This section is divided into two parts. In the first part, our method is compared to state-of-the-art approaches. Descriptors computed on complete sequences are used for testing. The second part concerns the early recognition of human action. For this purpose, descriptors are computed on incomplete sequences.

MSRAction3D dataset represents one of the most used benchmark for RGB-D based human action recognition. It is composed of depth maps and skeleton sequences. This dataset has been collected by Microsoft Research and includes 20 actions: *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up and throw*. Each action is performed by 10 subjects 2 or 3 times for a total of 567 sequences. To fairly compare our method to state-of-the-art approaches, the dataset has been divided into three subsets: AS1, AS2, AS3. The training and testing steps are done in each subset separately, and the average recognition obtained is reported. Also, the cross-splitting of [13] is followed, for which the data realized by half of the subjects have been used for training while the rest of the data has been kept for the testing step.

Florence 3D dataset has been collected at the university of Florence. It contains depth maps and skeleton sequences. It includes 9 different actions: *arm wave, drink from a bottle, answer phone, clap, tight lace, sit down, stand up, read watch, bow*. Each action is performed by 10 subjects, 2 or 3 times, for a total of 215 sequences. The main challenge of this dataset is its high intra-class variation: the same action is performed using left or right hand. We followed the experimental protocol of [15], where a leave-one-subject-out cross-validation is used.

UTKinect-Action dataset contains 10 subjects performing 10 different actions. Each subject performed every action two times. The 10 actions include: *walk, sit down, stand up, pick up, carry, throw, push, pull, wave and clap hands*. The dataset contains an additional challenge compared to MSRAction3D and Florence 3D: it is collected using different human body orientations with respect to the camera (right view, frontal view and back view). In order to compare our method to the state-of-the-art approaches, we again followed a leave-one-subject-out cross-validation protocol.

Multiview3D includes 8 subjects performing 12 actions: *one-hand wave, boxing, sitting, two-hand wave, holding head, phone answering, picking up, kicking, holding back, check watch, jumping, and throw over head*. Each subject performs the same action twice for three orientations (30°,0°,-30°).

A. Human action recognition

Table I reports the recognition rates compared with state-of-the-art methods on MSRAction3D dataset. The results presented below show that the introduced methodology achieves an average score of 92.9% of accuracy outscoring most of the previous methods.

Furthermore, Table II compares our method to state-of-the-art approaches on Florence3D dataset. Our approach presents one of the best accuracy score compared with literature methods. It registers 90.32% of accuracy (slightly less than the approach of Vamulapalli et al.[7] which presents a score of 90.88% of accuracy).

Algorithm	AS1 (%)	AS2 (%)	AS3 (%)	Overall (%)
Li et al. [13]	72.90	71.90	79.20	74.70
Venkataraman et al. [28]	77.50	63.10	87.00	75.90
Chen et al. [17]	96.20	83.20	92.00	90.50
Miranda et al. [20]	96.00	57.10	97.30	83.50
Chaaraouia et al. [29]	91.59	90.83	97.28	93.23
Vemulapalli et al. [7]	95.29	83.87	98.22	92.46
Du et al. [18]	93.33	94.64	95.50	94.49
Cippitelli et al. [19]	79.50	71.90	92.30	81.50
Liu et al. [30]	86.79	76.11	89.29	84.07
Ours	91.40	91.07	96.23	**92.90**

TABLE I: Accuracy of different methods on MSRAction3D dataset

In this way, the results obtained on two benchmarks prove that our method competes with recent skeleton based state-of-the-art approaches.

B. Viewpoint invariance

The viewpoint variation is related to the change of human body orientation of the person that performs an action with respect to the camera. In fact, the subject may be in front of the camera which is the ideal condition to make human action recognition. However, in a real-world context, the subject is not exactly in front of the camera, making the recognition task more complex. Thus, to test the robustness of our method to viewpoint variation, experiments are conducted on two datasets collected using different viewpoints: UTKinect and Multiview3D datasets.

We have to define how we consider the viewpoint. Both datasets are acquired with different viewpoints but the difference is in the granularity of these viewpoints. As explained before, the "ideal" configuration would be when the person faces the camera. It is defined by the fact that the two points of the hips of the person are parallel with the line defined by the camera. As it is a depth camera with two acquisition devices, this line is defined by the segment connecting the two acquisition cone origins. In Multiview3D dataset, the hips line is parallel or has a 30 or $-30°$angle with the camera line around the vertical axis. In UTKinect, the same line of the hips is considered but only the qualitative orientation (aligned, right view, left view or back view) is given.

Table III reports recognition rates on UTKinect dataset compared with state-of-the-art methods. It shows that our method outperforms various state-of-the-art approaches and achieves a score of 91.1% of accuracy. Even if Vemulapalli

Algorithm	Accuracy (%)
Seidenari et al. [15]	82.00
Anirudh et al. [31]	89.67
Devanne et al. [8]	87.04
Cippitelli et al. [19]	76.10
Vemulapalli et al. [7]	90.88
Ours	**90.32**

TABLE II: Accuracy of different methods on Florence3D dataset

Algorithm	Accuracy (%)
Zhu et al. [32]	87.90
Slama et al. [10]	88.50
Xia et al. [14]	90.92
Vemulapalli et al. [7]	97.08
Ours	**91.10**

TABLE III: Accuracy of different methods on UTKinect dataset

Angle data train	Angle data test		
	0°	30°	-30°
0°	92.71	90.63	90.10
30°	90.63	90.70	87.10
-30°	92.19	87.00	92.23
(0° 30° -30°)	92.71	91.15	91.67

TABLE IV: Accuracy (%) on Multiview3D dataset

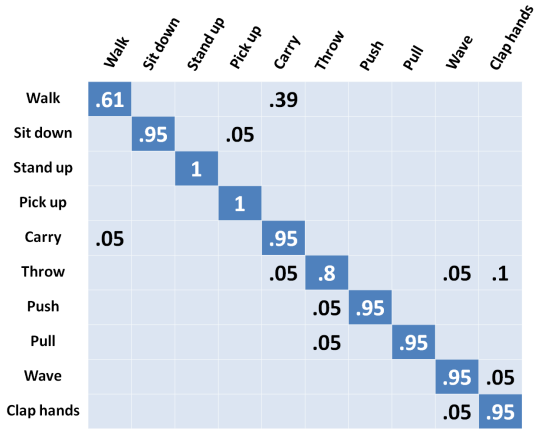


Fig. 4: Confusion matrix obtained on UTKinect dataset

et al.'s approach [7] presents a better accuracy, this method remains very time consuming to compute its decision. In [33], the computational time of [7] has been evaluated on MSRAction3D dataset. They reported that the approach takes an average of 17.61s to compute a descriptor (mean time on a sequence), knowing that MSRAction3D contains segmented videos composed of 12 to 54 frames. The algorithm was tested on a laptop with a CPU Intel Core i7 and with a RAM of 4 GB (similar to ours). Therefore, one can say that our method presents more interesting performances in terms of time processing. We will discuss this point in details at the end of this section.

The confusion matrix is presented in Fig. 4. We can notice that almost all actions are well recognized. However, we observe some confusion between the actions *walk* and *carry*. The main reason is that in some cases, subjects walk while performing the action "carrying". In our approach, we only exploit the skeleton data and we do not include information related to object interactions.

In Multiview3D dataset, the actions are observed from 3 views (30°, 0°, -30°). The tests are carried out under two settings: same view testing (training data and testing data have the same view) and multi-view testing (training data and testing data have different views). Table IV reports the obtained accuracy for both experiments. It shows that when the same view is used for training and testing we reach an average score of 91.88% accuracy. When different views are used for training and testing we achieve an average score of 89.61%. The obtained results for the tests are very close (difference of 2.27%). These results demonstrate the robustness of our method to viewpoint variation.

C. Early recognition of human actions

To perform early recognition of human action, the skeleton data are generated frame-by-frame. At each new data received, our system is able to give a decision.

For each action, we evaluate the percentage of necessary frames to recognize the ongoing action. For instance, Figure 5a illustrates the likelihoods obtained as outputs for each HMM during the execution of the action "Side kick" that belongs to dataset MSRAction3D. In that case we notice that based only on 20% of the global sequence, the likelihood of the HMM corresponding to the action "Side kick" exceeds likelihoods of other classes until the end of the action. In this example, the action is recognized very early. We also present another example in Figure 5b showing the recognition of the action "Walk" that belongs to the dataset UKkinect over time. After observing 10% of the global sequence, the likelihoods corresponding to the actions "walk" and "carry" clearly exceeds the likelihoods of other classes. The reason is that these actions are the only ones in UTKinect that involve principally the movement of legs (the subjects walk when they perform the action "carry"). After 10% of observed frames, we notice a small difference between these actions. In the confusion matrix presented in Fig. 4, it can be seen that the action "walk" is sometimes confused with the action "carry". Indeed, it is very difficult to discriminate them, even after observing the whole sequence.

To visualize the global results, we use boxplots depicting graphically data distributions through the smallest observation, lower quartile, median, upper quartile and the largest observation. The Interquartile range (IQR) represents the difference between the third and the first quartiles, illustrated by the length of the box. Comparing to the mean and the standard deviation, the median and the IQR are robust to outliers and non-normal data.

To evaluate our method in terms of early action recognition, we propose the following graph presented in Figure 6. This latter shows the boxplot distributions of the observed frames needed to recognize actions. We performed our tests on three subsets of MSRAction3D. Globally, we obtained different distributions. The median value separating the highest half of distribution from the lowest one is represented by a segment inside the rectangle. For the three subsets, Med varies in $[4\%, 52\%]$. The maximum value corresponds to the class "Hammer" equal to 52% which is very promising. That means that our system recognizes the half of the actions of each class at almost the middle of the sequence. The minimum values of median are obtained for both classes "forward punch" and "side kick". These actions are recognized very quickly. It should be due to the properties of Body-part Directional Velocity descriptor which takes into account the direction of

body-part motion. The action "Forward punch" is the only class of the subset AS1, where only one arm is moving in a forward direction. For the class "Side kick", only one leg is moving in the side direction. In the boxplots, the IQR indicates the percentage spread of necessary frames for recognizing an action with $IQR \in [3\%, 45\%]$. These results might be due to the intraclass correlation, which reflects the execution variability present in each class. The maximum value observed belongs to the class "High arm wave" with 91% of needed frames. Outliers are also shown by the boxplots. They are plotted individually using the "+" symbol. They represent the observation values that are distant from other observations. Let us denote by q_1 and q_3 the first and the third quartiles. Outliers are greater than $q_3 + 1.5 \times (q_3 - q_1)$ or less than $q_1 - 1.5 \times (q_3 - q_1)$. The highest value is observed in the class "Side kick" with 98%. We note that outliers may be misleading.

The three subsets share the following actions, "High throw", "Pick up and throw", "Forward kick" and "Tennis serve". The boxplots of the first two classes show that depending on which subset these actions belong (AS1 or AS3), the distribution of the required percentage of frames is different. On the other hand, the classes "Forward kick" and "Tennis serve" present the same distribution in the boxplots. It should be related to the interclass variability in each subset. An action in subset with similar actions is more challenging and need more percentage of frames, while, an action in a subset with distinct actions is recognized quickly and needs less frames.

Stage	Processing time (ms)
Pre-processing	0.03 ± 0.006/frame
BDV feature extraction	0.2 ± 0.05/frame
Likelihood per HMM	1.5 ± 0.2/frame

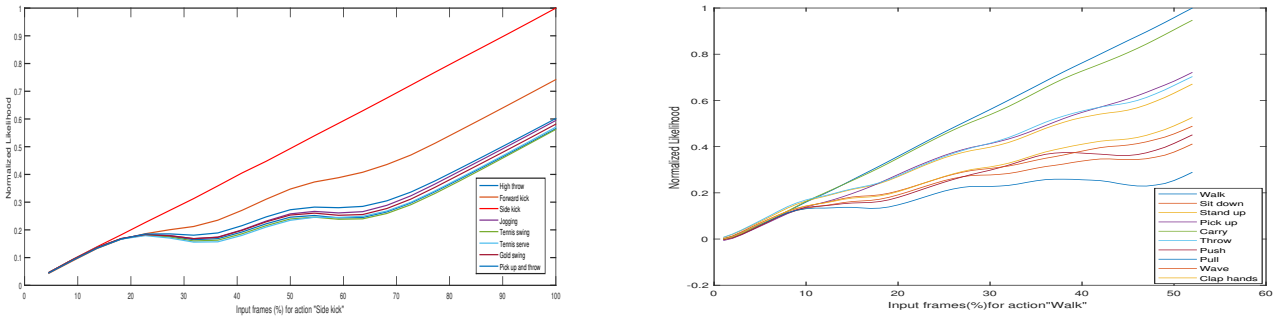
TABLE V: The computational time for each stage

Finally to prove that our approach is suitable for early recognition, we report the processing time needed for each frame. Before, we should underline that the sampling frequency of a Kinect sensor is $30Hz$. This means that it captures a depth map each $33.33ms$. In [4], Shotton et al. indicate that optimized implementation of skeleton extraction algorithm operates in less than $5ms$. Therefore, to carry out a real-time early recognition, the computational time for each frame must be less than $33.3ms$ for a depth-based approach and than $28.33ms$ for skeleton-based approach. In [17], the depth-based proposed approach reached $9.6ms$ on average. In [8], $50ms$ processing time is needed for the whole action recognition process. Chaaraoui et al [29] method showed a low computational time with $1.85ms$. The computational time of the proposed approach is evaluated on MSRAction3D dataset. All experiments are implemented in real-time using Matlab on a CPU Intel Core i5 2.60 GHz and 4 GB of RAM. The computational time is evaluated by the built-in MATLAB function tic-toc which provides $1\mu s$ resolution [34]. For each stage of our method, we have calculated the average and the standard deviation of computational time, as reported in the table V. We show that our approach achieve a low computational time with an average of $1.73ms$ for all stage. It is lower than $28.33ms$, the limit to perform real-time recognition. Therefore, our approach realizes a real-time early recognition.

V. Conclusion

In this paper, we have proposed a novel approach to perform human action recognition using RGB-D sensors. We have focused our work on two major challenges for robotic applications, the robustness to viewpoint changes and the early recognition property.

A novel real-time feature extraction algorithm called Body-part Directional Velocity (BDV) has been proposed, and a Hidden Markov Models (HMMs) classifier with Gaussian Mixture Models (GMMs) state-output distributions has been trained to classify human actions. To show the robustness of our approach, we have first tested it using all frames to perform the action recognition task. The experimental results on two public datasets have demonstrated that our approach is effective and that it outperforms various state-of-the-art skeleton-based human action recognition approaches by reaching an average accuracy of 92.9% and 90.32% on both benchmarks. The second part of experiments has involved early recognition of human actions. In our study, we focused on the percentage of necessary frames to recognize each action. We analyze the distribution of the percentage of needed frames for each class to perform the classification and we show promising



(a) Side kick

(b) Walk

Fig. 5: Evolution of the likelihood during performing actions side kick and walk.

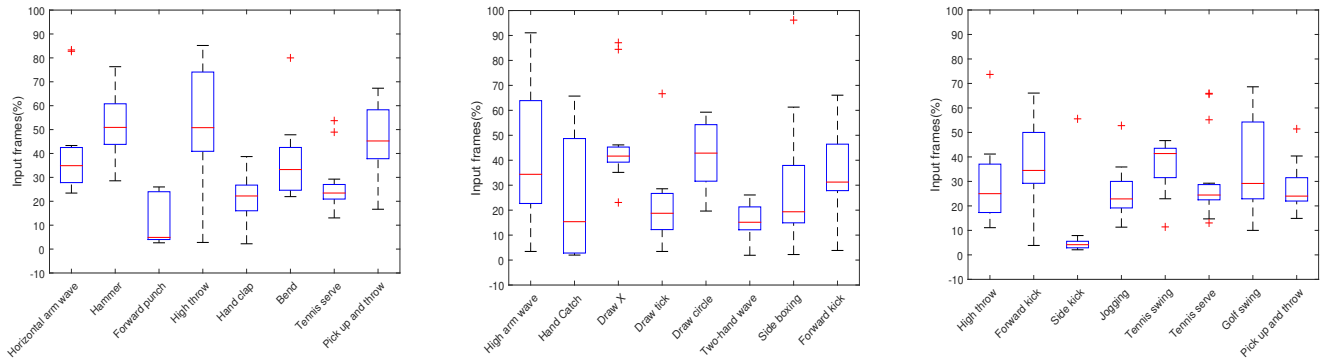


Fig. 6: Boxplot distributions of input necessary frames to recognize the different actions in the three subsets of MSRAction3D (from left to right AS1, AS2 and AS3). The first axis represent the different actions of the datasets and the second axis the percentage of frame to obtain the decision.

performance: some classes have been recognized with only 4% of the whole sequence. Others need more frames but most of them do not need more than 50% of the frames of an action sequence.

Robustness to view point variability remains a very important problem. In the future, we will study if convolutional network approaches applied in the BDV feature space can bring some improvements to this problem. Moreover, we will fuse our new skeleton method with a depth-based reasoning as we began in a previous work [26] to finally define action recognition in an unified framework.

REFERENCES

[1] United Nations, New York, Ny. Dept. of Economic and Social Affairs, *World population ageing, 1950-2050*. United Nations Publications, 2002.

[2] C. Zhu and W. Sheng, "Wearable sensor-based hand gesture and daily activity recognition for robot-assisted living," *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, vol. 41, no. 3, pp. 569–573, 2011.

[3] S. Chernbumroong, S. Cang, A. Atkins, and H. Yu, "Elderly activities recognition and classification for applications in assisted living," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1662–1674, 2013.

[4] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook, and R. Moore, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM*, vol. 56, no. 1, pp. 116–124, 2013.

[5] E. Ohn-Bar and M. M. Trivedi, "Joint angles similarities and HOG2 for action recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*. IEEE, 2013, pp. 465–470.

[6] O. Oreifej and Z. Liu, "Hon4d: Histogram of oriented 4D normals for activity recognition from depth sequences," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 716–723.

[7] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3D skeletons as points in a Lie group," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 588–595.

[8] M. Devanne, H. Wannous, S. Berretti, P. Pala, M. Daoudi, and A. Del Bimbo, "3-d human action recognition by shape analysis of motion trajectories on riemannian manifold," *IEEE transactions on cybernetics*, vol. 45, no. 7, pp. 1340–1352, 2015.

[9] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[10] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, "Accurate 3d action recognition using learning on the grassmann manifold," *Pattern Recognition*, vol. 48, no. 2, pp. 556–567, 2015.

[11] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*. IEEE, 2014, pp. 804–811.

[12] M. Zanfir, M. Leordeanu, and C. Sminchisescu, "The moving pose: An efficient 3D kinematics descriptor for low-latency action recognition and detection," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 2752–2759.

[13] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3D points," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 9–14.

[14] L. Xia, C.-C. Chen, and J. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*. IEEE, 2012, pp. 20–27.

[15] L. Seidenari, V. Varano, S. Berretti, A. Bimbo, and P. Pala, "Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 479–485.

[16] M. Hammouche, S. Ambellouis, and A. Fleury, "Uria kinect dataset," http://ia.ur.mines-douai.fr/en/datasets/, 2015.

[17] C. Chen, K. Liu, and N. Kehtarnavaz, "Real-time human action recognition based on depth motion maps." *J. Real-Time Image Processing*, vol. 12, no. 1, pp. 155–163, 2016.

[18] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1110–1118.

[19] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using skeleton data from rgbd sensors," *Computational intelligence and neuroscience*, vol. 2016, p. 21, 2016.

[20] L. Miranda, T. Vieira, D. Martínez, T. Lewiner, A. W. Vieira, and M. F. Campos, "Online gesture recognition from pose kernel learning and decision forests," *Pattern Recognition Letters*, vol. 39, pp. 65–73, 2014.

[21] G. Evangelidis, G. Singh, and R. Horaud, "Skeletal Quads: Human Action Recognition Using Joint Quadruples," in *International Conference on Pattern Recognition*. Stockholm, Sweden: IEEE, 2014, pp. 4513 – 4518. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00989725

[22] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "Real time action recognition using histograms of depth gradients and random decision forests," in *IEEE Winter Conference on Applications of Computer Vision*, 2014, pp. 626–633.

[23] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Learning actionlet ensemble for 3d human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 5, pp. 914–927, 2014.

[24] A. Shahroudy, T. T. Ng, Q. Yang, and G. Wang, "Multimodal multipart learning for action recognition in depth videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 10, pp. 2123–2129, 2016.

[25] H. Rahmani and M. Bennamoun, "Learning action recognition model from depth and skeleton videos," in *The IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

[26] M. Hammouche, E. Ghorbel, A. Fleury, and S. Ambellouis, "Toward a Real Time View-invariant 3D Action Recognition," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP)*, Roma, Italy, 2016. [Online]. Available: https://hal.archives-ouvertes.fr/hal-01332468

[27] L. Rabiner and B. Juang, "An introduction to hidden markov models," *ieee assp magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[28] V. Venkataraman, P. Turaga, N. Lehrer, M. Baran, T. Rikakis, and S. Wolf, "Attractor-shape for dynamical analysis of human movement: Applications in stroke rehabilitation and action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2013, pp. 514–520.

[29] A. A. Chaaraoui, J. R. Padilla-López, P. Climent-Pérez, and F. Flórez-Revuelta, "Evolutionary joint selection to improve human action recognition with rgb-d devices," *Expert systems with applications*, vol. 41, no. 3, pp. 786–794, 2014.

[30] Z. Liu, C. Zhang, and Y. Tian, "3d-based deep convolutional neural network for action recognition with depth sequences," *Image and Vision Computing*, vol. 55, pp. 93–100, 2016.

[31] R. Anirudh, P. Turaga, J. Su, and A. Srivastava, "Elastic functional coding of human actions: From vector-fields to latent variables," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3147–3155.

[32] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3d action recognition," in *Computer vision and pattern recognition workshops (CVPRW), 2013 IEEE conference on*. IEEE, 2013, pp. 486–491.

[33] E. Ghorbel, R. Boutteau, J. Bonnaert, X. Savatier, and S. Lecoeuche, "A fast and accurate motion descriptor for human action recognition applications," in *Pattern Recognition (ICPR), 2016 23rd International Conference on*. IEEE, 2016, pp. 919–924.

[34] M. Knapp-Cordes and B. McKeeman, "Improvements to tic and toc functions for measuring absolute elapsed time performance in matlab," in *Matlab Technical Articles and Newsletters*. The MathWorks Inc., 2011.