Human Action Recognition from Body-Part Directional Velocity using Hidden Markov Models

Sid Ahmed Walid Talha, Anthony Fleury, Sebastien Ambellouis

Human Action Recognition from Body-Part Directional Velocity using Hidden Markov Models

Sid Ahmed Walid Talha, Anthony Fleury
Computer Sciences and Automatic Control Dpt.
IMT Lille Douai, Univ. Lille
F-59000 Lille, France
sidahmed.talha@imt-lille-douai.fr, anthony.fleury@imt-lille-douai.fr

Sébastien Ambellouis
Cosys Dept. - Leost
IFSTTAR
F-59666 Villeneuve d'Ascq, France
sebastien.ambellouis@ifsttar.fr

***Abstract*—This paper introduces a novel approach for early recognition of human actions using 3D skeleton joints extracted from 3D depth data. We propose a novel, frame-by-frame and real-time descriptor called Body-part Directional Velocity (BDV) calculated by considering the algebraic velocity produced by different body-parts. A real-time Hidden Markov Models algorithm with Gaussian Mixture Models state-output distributions is used to carry out the classification. We show that our method outperforms various state-of-the-art skeleton-based human action recognition approaches on MSRAction3D and Florence3D datasets. We also proved the suitability of our approach for early human action recognition by deducing the decision from a partial analysis of the sequence.**

***Index Terms*—Human Action Recognition; RGB-D Sensor; Hidden Markov models (HMMs); Gaussian mixture models (GMMs).**

I. Introduction

Human action recognition has been widely studied in the field of computer vision and machine learning. It can be applied in many domains such as video surveillance, video games, ambient assisted living. Several studies exploit image sequences provided by standard cameras to recognize actions [1]–[4]. These approaches deal with 2D information and present some limitations such as color sensitivity, complex background and illumination changes. The arrival of low-cost 3D capturing systems, has motivated researchers to investigate in action recognition using RGB-D sensors. In addition to RGB data, they provide depth maps to address the issues revealed by 2D cameras. Most of these sensors also include embedded skeleton extraction algorithms to provide data.

In real-world applications, early action recognition is very important. For elderly people, a fall may cause injury and induce a long hospitalization. Early recognition of fall before the person hits the ground is necessary to prevent injuries, by triggering the inflation of a wearable airbag for example [5]. In video games applications, latency is important [6] and the prior knowledge of the action is necessary to improve the response time. In the field of video surveillance and security problems [7], a suspicious activity has to be almost instantly detected. This challenge requires a system able to analyze the ongoing human action in real-time and recognize it before it completely appears.

Our propositions are: first, we propose a new feature related to moving direction of body-parts to describe human actions. Then, a classifier using continuous HMMs is learned allowing the calculation of the likelihoods associated to each class for frame received. We evaluate the effectiveness of our method on two reference datasets. We demonstrate that our system is able to recognize the ongoing actions before the end of their execution to provide an early recognition.

The remainder of this paper is organized as follows. Section II reviews existing methods for human action recognition. Section III details the proposed descriptor named *Body-part Directional Velocity* (BDV). Section IV describes the continuous HMMs and the real-time recognition step. Section V discusses the experimental results and finally Section VI concludes the paper and sum up our results.

II. Related work

Approaches using RGB-D cameras for human action recognition can be divided into two major classes by using depth frames of human body movements and using a skeleton sequence.

Li et al. [8] are among the first to work on depth images for action recognition. They proposed an action graph to model the dynamics of the actions and a bag of 3D points to describe salient postures. Gaussian Mixture Models (GMMs) are employed to capture the statistical distribution of the features. In [9], a 4D histogram over depth, time, and space is used to capture the changes of the surface normal orientation (HON4D). The final descriptor is the concatenation of the HON4D computed for each cell. Chen et al. [10] extracted the features using the depth motion maps (DMMs). Each depth frame in a depth video sequence is projected onto three orthogonal Cartesian planes and the absolute difference between two consecutive projected images was accumulated through an entire depth video.

To deal with the huge amount of data of depth-based approaches and the expensive computation of training, validation and testing steps, the skeleton approaches use higher-level information extracted from each depth frames for a more compact representation. For the Kinect, the extraction of the skeleton is done in real-time using Shotton et al.'s method [11] embedded in the sensor. An human action is thus constituted

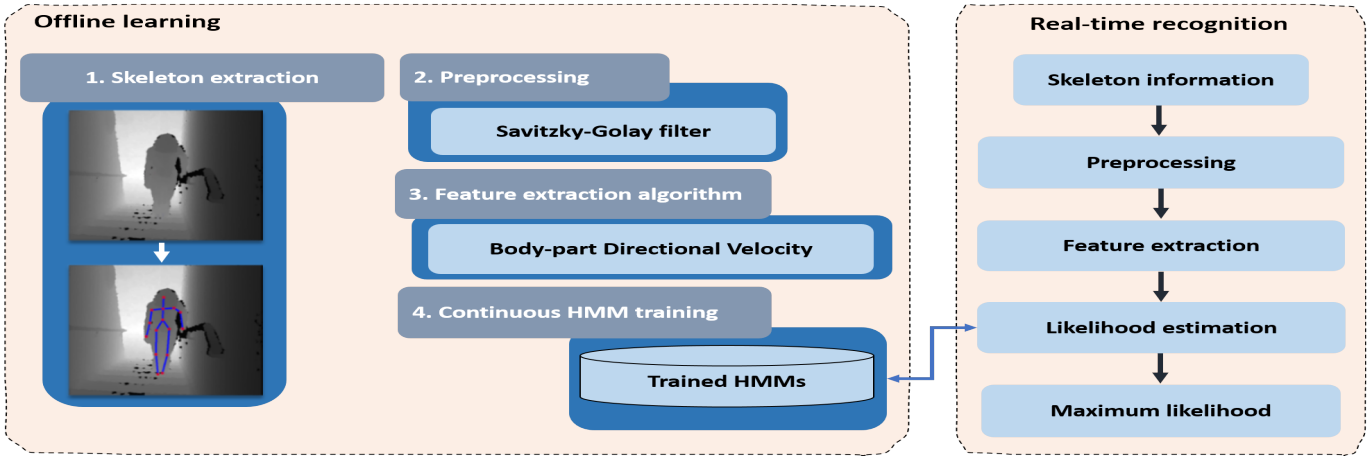

Fig. 1: Overview of the proposed system

of a time series of poses. Each pose then gives the relative position of a set of rigid segments connected by joints. Many recent skeleton-based approaches have shown their ability to recognize actions. Xia et al. [12] uses 3D joints located by spherical coordinates. It computes a posture vocabulary by clustering the spatial histograms of joint location which space is reduced by a linear discriminant analysis. The temporal evolution of the posture sequences has been modeled using discrete hidden Markov models (HMMs).

Some authors have used deep learning techniques. For instance, Huang et al [13] incorporate the Lie group structure into a deep network architecture to learn more suitable features. Du et al. [14] propose a Hierarchical Bidirectional Recurrent Neural Network (HBRNN) to classify human action. They divide the skeleton into five groups of joints representing two arms, two legs and the trunk. Each group is fed into five BRNNs. The generated hidden states are combined and fed into another set of BRNNS as inputs. A softmax classifier layer is used to recognize actions. In Cippitelli et al. [15], the feature extraction step involves a normalization of the skeleton using the Euclidean distance between the torso and the neck joint. Groups of similar postures are defined thanks to clustering and a multiclass Support Vector Machine (SVM) is applied to identify human actions. In [6], Devanne et al. compute the similarity between shapes of skeleton joints trajectories in a Riemannian manifold. The classification is performed using a k-NN-based classifier. Miranda et al. [16] introduce a real-time method based on a spherical coordinates representation of skeleton joints. A multiclass SVM classifier with a tailored pose kernel is performed to identify key poses while a random forest based decision process allows the recognition of gestures from the key pose sequences. In [17] the authors propose a k-NN based approach based a moving pose descriptor containing 3D joint positions, velocities and accelerations.

III. Body-part Directional Velocity (BDV)

Our new approach is illustrated in Figure 1. It fulfills an early human action recognition by providing a frame-by-frame decision. As emphasized by authors in [18], preprocessing step is very important to reach high performance. A new descriptor has been designed for its suitability for early action recognition application. Indeed, the whole skeleton sequence is not needed in order to build BDV. In this section, we detail the calculation of BDV descriptor.

A skeleton sequence $\mathbf{p}$ represents a series of N temporal ordered poses as described by Equation 1. At an instant t, the skeleton pose is referred to $\mathbf{p}_t$.

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

At each instant t, $\mathbf{p}_t$ is composed of a set of n joint position, as described by Equation (2) (where p_t^i represents the position of ith joint position at an instant 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 preprocessing of smoothing is carried out. Therefore, a Savitzky-Golay filter is applied to all joint positions as described below. $\forall (i,t) \in [\![1,n]\!] \times [\![1,N]\!]$,

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

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

Then, the velocity at an instant t of each joint i, considered as a very discriminative feature, is computed as in [17] 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, right arm (B_1), left arm (B_2), right leg (B_3), left leg (B_4) and spine (B_5), as illustrated in Figure 2. The set of body-parts is therefore denoted by $B = \{B_1, B_2, B_3, B_4, B_5\}$.

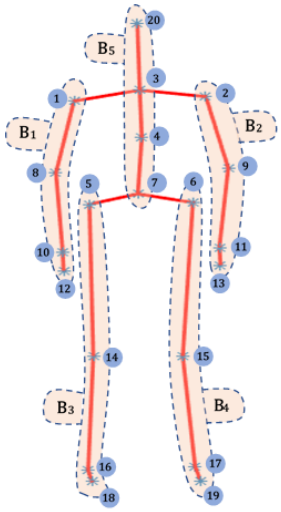

Fig. 2: Human body divided into five different body-parts (B_1, B_2, B_3, B_4 and B_5)

Then, for every body-part, the sum of the negative and positive velocity of associated joints are respectively computed and are respectively denoted by $D^+_{B_l}$ and $D^-_{B_l}$ as depicted by Equation (5) and (6). The separation of negative and positive values is very informative since it indicates the direction of the motion.

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

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

The final descriptor obtained at an instant t, denoted by $\mathbf{D}(t)$ is computed as described by Equation (7).

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

Therefore, the dimension of the proposed descriptor BDV descriptor denoted by D(t) at an instant t is equal to $d_D = 30$.

IV. Classification using continuous HMMs

In this work, HMMs with GMMs state-output distributions (illustrated in Figure 3) are used to model the BDV distribution of each human action. A Hidden Markov Model (HMM) [19] 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 processes, one is an observable process which represents the sequence of observed symbols. The second process is unobservable (hidden) and can be indirectly inferred by analyzing the sequence of observed symbols. In this work, an HMM is learned for every action a.

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

- N^a: Number of states in the model.
- M: Number of observation symbols.
- $S^a = \{s^a_1, s^a_2, ..., s^a_N\}$: Set of distinct states.
- $V = \{v_1, v_2, ..., v_M\}$: Observation alphabet.
- $Q^a = \{q^a_1, q^a_2, ..., q^a_T\}$: T states from S^a.
- $O = \{o_1, o_2, ..., o_T\}$: T observations from alphabet V corresponding to Q^a states.

Each HMM^a can be written in a compact form as

$$\lambda^a = (\pi^a, A^a, B^a) \tag{8}$$

where π^a is the vector of initial state distribution:

$$\pi^a = \{\pi_i\}, \pi_i = P(q_1 = s_i)_{1 \leq i \leq N^a} \tag{9}$$

and A^a is the matrix of state transition probability distribution, represent transition from state i to state j:

$$A^a = \{a_{ij}\}, a_{i,j} = P(q_{t+1} = s_j | q_t = s_i)_{1 \leq i,j \leq N^a} \tag{10}$$

B^a is the matrix of observation symbol probability distribution, represent the probability of observation k being generated from the state i:

$$B^a = \{b_{ik}\}, b_{ik} = P(o_t = v_k | q_t = s_i)_{1 \leq i \leq N^a, 1 \leq k \leq M} \tag{11}$$

The Discrete HMM (DHMM) considers that the observations are discrete symbols from a finite alphabet. The symbols are obtained by applying unsupervised classification algorithm to extracted features. In [12], it is performed by clustering method as k-means. 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 features as depicted by Equation (12).

$$b_i(o_t) = \sum_{r=1}^{N_g} w_{ir} g(o_t, \mu_{ir}, C_{ir})_{1 \leq i \leq N^a, 1 \leq k \leq M} \tag{12}$$

Where w_{ir}, μ_{ir} and C_{ir} represent respectively the weight, the mean vector and the covariance matrix of the Gaussian model r 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 represent 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:

$$g(o_t, \mu_{ir}, C_{ir}) = \frac{1}{(2\pi)^{d_D/2} |C_{ir}|^{1/2}} e^{\frac{-1}{2}(o_t - \mu_{ir})^T C^{-1}_{ir} (o_t - \mu_{ir})} \tag{13}$$

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

The likelihood estimation of the feature vector sequence is calculated for each HHM^a, at each instant t using the forward

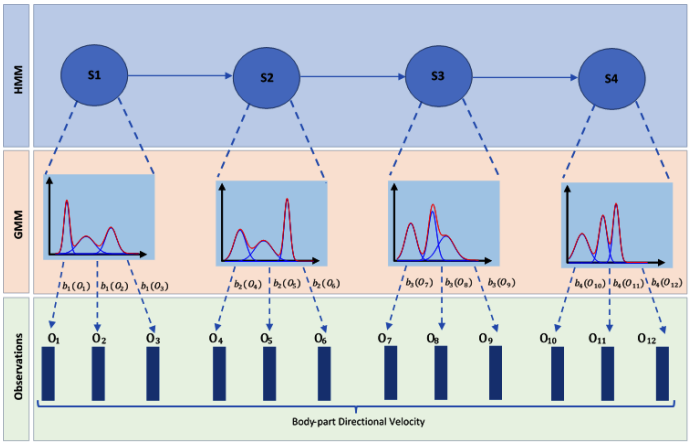

Fig. 3: Architecture of the used HMM-GMM system

algorithm. then the most likely HMM is selected as the correct label a^*, as described by Equation (14).

$$a^*(t) = \arg\max_{a \in A}(P(O|\lambda_i)) \tag{14}$$

V. Experimental results

This section is divided into two parts. In the first part, our method is compared to state-of-the-art approaches using MSRAction3D [8] and Florence3D [20] datasets. Descriptors computed on the entire sequences are used for testing. The second part concerns the early recognition of human action. For this purpose, descriptors are computed in real-time on incomplete sequences. Our tests are performed on two popular datasets, namely MSRAction3D [8] and Florence3D datasets [20].

MSRAction3D dataset represents one of the most used benchmark for RGB-D based human action recognition. It is formed by depth maps and skeleton sequences. This dataset has been collected by Microsoft Research and includes 20 actions. Each action is performed by 10 subjects 2 or 3 times for a total of 567 sequences. As discussed in [8] and to realize a fair comparison with state-of-the-art methods, the dataset has been divided into three subsets: AS1, AS2, AS3 as shown in Table I. The training and testing is therefore done in each subset separately, and the average recognition obtained is reported. Also, the cross-splitting of [8] is followed: the data realized by half of the subjects have been used for training, while the rest of the data has been used for testing.

Florence 3D dataset has been collected at the university of Florence. It contains depth maps and skeleton sequences. It includes 9 different actions. Each action is performed by 10 subjects, 2 or 3 times, for a total of 215 sequences. We followed the experimental protocol of [20], where a leave-one-out subject validation is performed.

1) Human action recognition: Table II reports the recognition accuracy compared with the state-of-the-art methods on MSRAction3D dataset. The results presented show that our method achieves a score of 92.9% of accuracy outscoring most

AS1	AS2	AS3
[a02] Horizontal arm wave	[a01] High arm wave	[a06] High throw
[a03] Hammer	[a04] Hand Catch	[a14] Forward kick
[a05] Forward punch	[a07] Draw X	[a15] Side kick
[a06] High throw	[a08] Draw tick	[a16] Jogging
[a10] Hand clap	[a09] Draw circle	[a17] Tennis swing
[a13] Bend	[a11] Two-hand wave	[a18] Tennis serve
[a18] Tennis serve	[a12] Side boxing	[a19] Golf swing
[a20] Pick up and throw	[a14] Forward kick	[a20] Pick up and throw

TABLE I: Three subsets of actions from MSR Action3D dataset: AS1, AS2, AS3

Algorithm	AS1 (%)	AS2 (%)	AS3 (%)	Overall (%)
Li et al. [8]	72.90	71.90	79.20	74.70
Venkataraman et al. [21]	77.50	63.10	87.00	75.90
Chen et al. [10]	96.20	83.20	92.00	90.50
Miranda et al. [16]	96.00	57.10	97.30	83.50
Chaaraouia et al. [22]	91.59	90.83	97.28	93.23
Vemulapalli et al. [23]	95.29	83.87	98.22	92.46
Du et al. [14]	93.33	94.64	95.50	94.49
Cippitelli et al. [15]	79.50	71.90	92.30	81.50
Liu et al. [24]	86.79	76.11	89.29	84.07
Ours	91.40	91.07	96.23	**92.90**

TABLE II: Accuracy of different methods on MSRAction3D dataset

of the previous methods. As presented in the confusion matrix (Figure 5), most of the actions are well recognized. Confusion occurs only among very similar actions. For example, for "hand catch" action (a04), the good classification rate is 0.5. One explanation is that the "hand catch" action is mostly characterized by the catching part of the sequence and this moment is not captured by the skeleton that has just one joint per hand.

A comparison with the state-of-the-art methods is presented in Table III for Florence 3D dataset.

As presented in this table, our approach performs well in terms of accuracy in comparison with the literature methods. As shown in the confusion matrix presented in Figure 4, it

Algorithm	Accuracy (%)
Seidenari et al. [20]	82.00
Anirudh et al. [25]	89.67
Devanne et al. [6]	87.04
Cippitelli et al. [15]	76.10
Vemulapalli et al. [23]	90.88
Ours	**90.32**

TABLE III: Accuracy of different methods on Florence3D dataset

	Arm wave	Drink	Phone call	Clap	Tight lace	Sit down	Stand up	Read watch	Bow
Arm wave	.92	.04	0	0	0	0	0	.04	0
Drink	.04	.77	.15	0	0	0	0	.04	0
Phone call	0	.18	.73	0	0	0	0	.09	0
Clap	0	0	0	.94	0	0	0	.06	0
Tight lace	0	0	0	0	1	0	0	0	0
Sit down	0	0	0	0	0	1	0	0	0
Stand up	0	0	0	0	0	0	1	0	0
Read watch	.09	.04	.04	.13	0	0	0	.70	0
Bow	0	0	0	0	0	0	0	0	1

Fig. 4: Confusion matrix obtained on Florence3D dataset

	a02	a03	a05	a06	a10	a13	a18	a20
a02	.92	0	0	0	0	0	0	.08
a03	0	1	0	0	0	0	0	0
a05	0	.18	.73	0	0	0	.09	0
a06	0	.18	0	.82	0	0	0	0
a10	0	0	0	0	1	0	0	0
a13	0	0	0	0	0	1	0	0
a18	0	0	0	0	0	0	1	0
a20	0	0	0	0	0	.22	0	.78

	a01	a04	a07	a08	a09	a11	a12	a14
a01	.76	0	0	0	.08	.08	.08	0
a04	.25	.5	.25	0	0	0	0	0
a07	0	0	1	0	0	0	0	0
a08	0	0	0	.93	0	0	.07	0
a09	0	0	0	0	1	0	0	0
a11	0	0	0	0	0	1	0	0
a12	0	0	0	0	0	0	1	0
a14	0	0	0	0	0	0	0	1

	a06	a14	a15	a16	a17	a18	a19	a20
a06	.73	0	0	0	0	.27	0	0
a14	0	1	0	0	0	0	0	0
a15	0	.09	.91	0	0	0	0	0
a16	0	0	0	1	0	0	0	0
a17	0	0	0	0	1	0	0	0
a18	0	0	0	0	0	1	0	0
a19	0	0	0	0	0	0	1	0
a20	0	0	0	0	0	0	0	1

Fig. 5: Confusion matrices obtained on MSRAction3D for (from left to right) AS1, AS2 and AS3

can be observed that the classes (5, 6, 7, 9) are perfectly classified, while the classes 2 and 3 (drink from a bottle versus answer phone), present a high confusion rate. We can explain it because both classes involve human-object interaction which is not captured by the skeleton data.

2) Early recognition of human actions: To perform early recognition of human action, we propose to compute a classification likelihood at each frame. Our system extracts the corresponding BDV descriptors for each frame and applies a forward algorithm to estimate the likelihood for each HMM. The class that gives the maximum likelihood corresponds to the inferred ongoing action.

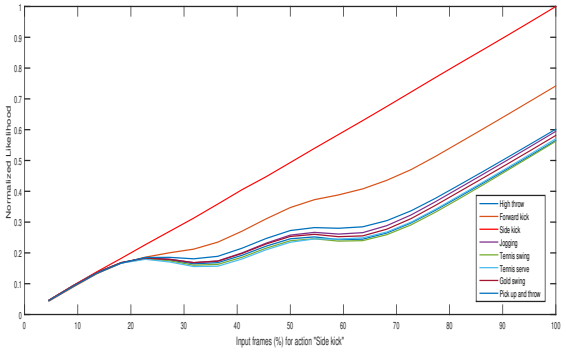

Fig. 6: Likelihood during performing the action "Side kick"

To evaluation the early recognition performance for our approach, we select the percentage of necessary frames to recognize the ongoing action. For instance, Figure 6 illustrates the output likelihoods obtained for each HMM in real-time during the execution of the action "Side kick". In this case "Side kick" action is recognized early: from 20% of the global sequence, the likelihood of the HMM corresponding to the action "Side kick" exceed other HMMs until the end of the action.

As presented on Figure 7, we visualize the early recognition property of the method for MSRAction3D dataset, thanks to a boxplots representation of the smallest number of frames for recognition, the lower quartile, the median, the upper quartile and the largest number of frames. The Interquartile range (IQR) represents the difference between the third and the first quartile, i.e. 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.

Globally, we observe that the time property of the recognition depends on the subset. The median value separating the higher half of distribution from the lower half is represented by a segment inside the rectangle. For the subset AS1, $Med \in [4\%, 52\%]$, for the subset AS2 $Med \in [15\%, 42\%]$, and for the subset AS3 $Med \in [4\%, 41\%]$. The maximal values of the medians for each subset indicate that the half of each class can be recognized with 52% of seen frames for AS1, 42% for AS2 ,and 41% of AS3. 52% means that our system recognizes the half actions of each class at almost the middle of the ongoing action. The fastest recognized classes for the three subsets are the actions "Forward punch", "Two-hand wave" and "Side kick" with respectively 4%, 15% and 4% as a median value. These promising results are related to the quality of our feature extraction algorithm.

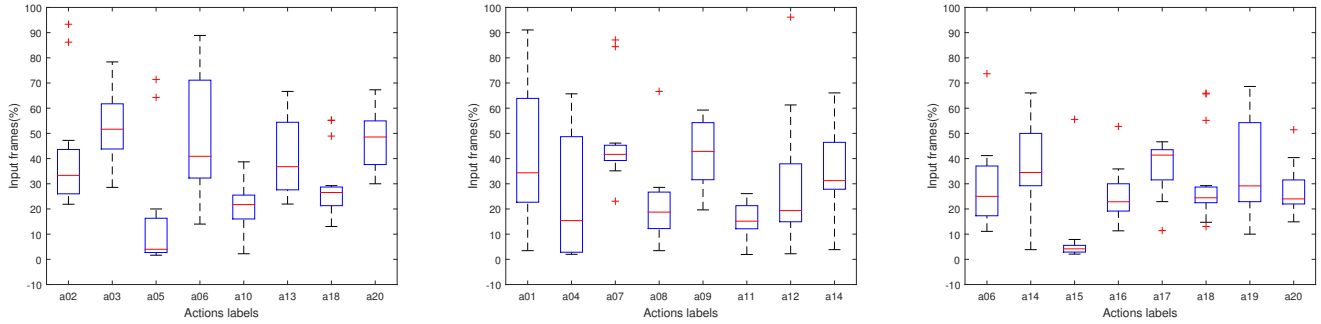

Fig. 7: Boxplot distributions of input necessary frames to recognize actions in the three subsets of MSRAction3D

In the boxplots, the IQR indicates the variability of the number of frames required for action recognition. The classes "High throw", "Hand Catch" and "Golf swing" present the largest IQRs in each subset with respectively 40%, 45% and 32%. This might be due to the complexity and the variability to perform the actions from one person to another. The lowest IQRs in each subset is achieved for the classes "Hand clap", "Draw X " and "Side kick" with respectively 10%, 7% and 3%. These actions are performed with low variability by the subjects consequently the system recognized the classes with almost the same part of percentage frames.

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.

VI. Conclusion

In this paper, we present a new method that performs early recognition of human action based on skeleton joints extracted from 3D depth data. A novel real-time feature extraction algorithm called Body-part Directional Velocity (BDV) is proposed and a Hidden Markov Models classifier with Gaussian Mixture Models state-output distributions is trained to classify human actions. Our system has been designed to provide, at each frame, a recognition likelihood of an ongoing action. To compare our method to the state of the art, we apply it to complete actions i.e. by using all frames of each action sequence. The experimental results obtained on two reference datasets show that our approach is effective and outperforms various well-known skeleton-based human action recognition techniques. The second part of experiment deals with the ability of the method to produce an early recognition. The performance is measure by calculating the number of frames required to recognize each action. The obtained results show promising performance for all actions and even by recognizing them before the end of their execution. Some classes have been recognized with only 4% of the frames and most of the actions do not need more than 50% to be classified.

References

[1] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36.

[2] K. Mikolajczyk and H. Uemura, "Action recognition with motion-appearance vocabulary forest," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[3] J. W. Davis and A. F. Bobick, "The representation and recognition of human movement using temporal templates," in *Computer Vision and Pattern Recognition, 1997. Proceedings., 1997 IEEE Computer Society Conference on*. IEEE, 1997, pp. 928–934.

[4] L. Wang and D. Suter, "Learning and matching of dynamic shape manifolds for human action recognition," *IEEE Transactions on Image Processing*, vol. 16, no. 6, pp. 1646–1661, 2007.

[5] T. R. Bennett, J. Wu, N. Kehtarnavaz, and R. Jafari, "Inertial measurement unit-based wearable computers for assisted living applications: A signal processing perspective," *IEEE Signal Processing Magazine*, vol. 33, no. 2, pp. 28–35, 2016.

[6] 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.

[7] W. Niu, J. Long, D. Han, and Y.-F. Wang, "Human activity detection and recognition for video surveillance," in *Multimedia and Expo, 2004. ICME'04. 2004 IEEE International Conference on*, vol. 1. IEEE, 2004, pp. 719–722.

[8] 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.

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

[10] 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.

[11] 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.

[12] 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.

[13] Z. Huang, C. Wan, T. Probst, and L. Van Gool, "Deep learning on lie groups for skeleton-based action recognition," *arXiv preprint arXiv:1612.05877*, 2016.

[14] 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.

[15] 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.

[16] 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.

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

[18] A. Holzinger, B. Malle, M. Bloice, M. Wiltgen, M. Ferri, I. Stanganelli, and R. Hofmann-Wellenhof, *On the Generation of Point Cloud Data Sets: Step One in the Knowledge Discovery Process*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014, pp. 57–80.

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

[20] 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.

[21] 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.

[22] 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.

[23] R. Vemulapalli, F. Arrate, and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 588–595.

[24] 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.

[25] 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.