Toward a Real Time View-invariant 3D Action Recognition

Mounir Hammouche, Enjie Ghorbel, Anthony Fleury, Sébastien Ambellouis

Toward a real time view-invariant 3D action recognition

Mounir Hammouche[1], Enjie Ghorbel[1], Anthony Fleury[1] and Sébastien Ambellouis[1,2]
[1]*Mines Douai, URIA, F-59508 Douai Cedex, France*
[2]*Ifsttar, Cosys - Leost, F-59666 Villeneuve d'Ascq Cedex, France*
mounirhammouche@gmail.com, {enjie.ghorbel,anthony.fleury}@mines-douai.fr,sebastien.ambellouis@ifsttar.fr

Keywords: Action Recognition, Human Interaction

Abstract: In this paper we propose a novel human action recognition method, robust to viewpoint variation, which combines skeleton- and depth-based action recognition approaches. For this matter, we first build several base classifiers, to independently predict the action performed by a subject. Then, two efficient combination strategies, that take into account skeleton accuracy and human body orientation, are proposed. The first is based on fuzzy switcher where the second uses a combination between fuzzy switcher and aggregation. Moreover, we introduce a new algorithm for the estimation of human body orientation. To perform the test we have created a new Multiview 3D Action public dataset with three viewpoint angles (30°,0°,-30°). The experimental results show that an efficient combination strategy of base classifiers improves the accuracy and the computational efficiency for human action recognition.

1 INTRODUCTION

Action recognition has been an active field of research in the last decades (Laptev and Lindeberg, 2003; Rahmani et al., 2014). It has a strong connection to many fields such as smart surveillance, intelligent human-robot interaction, and sociology (Peng et al., 2014). The activities that have been studied include people interactions, single or group of person(s) activities. Moreover, human action recognition methods can be divided into three categories that are RGB-, skeleton- and depth-based methods (Wang et al., 2012b).

In RGB-based methods, research has focused on learning directly from the colors of the image sequences. However, there are inner limitations of this kind of information, e.g. it is sensitive to illumination changes and background clutters. The accurate RGB-based actions recognition still remains a challenging task (Vemulapalli et al., 2014).

With the recent advances of depth cameras these last years, such as Kinect, depth-based approaches have received a greater attention. Depth-based methods have several advantages compared with RGB-based ones. First, the depth cameras offers 3D map information of the scene, which provides more discerning information to recognize actions. Secondly, depth cameras can work in dark conditions (Xia and Aggarwal, 2013). The advantages of depth cameras also lead to a renewal interest in skeleton-based action recognition, notably after the integration of the robust skeleton estimation algorithm of Shotton et al.(Shotton et al., 2011) in several software. This algorithm makes the estimation of joints from depth video sequence relatively easy, fast and accurate.

Depth-based action recognition approaches provide better results than skeleton-based approaches in the presence of self-occlusion. However, these methods suffer from the lack of precision and are sensitive to viewpoint variation. There are some view-invariant methods based on the extraction of features from pointclouds instead of depth images (Rahmani et al., 2014), but these methods are time consuming and not adapted for real-time applications.

The skeleton estimated from depth images is widely accurate in laboratory settings. However in real conditions, the situations are more complex: Self-occlusion between body segments usually appears and people are not often in front of the camera (Xia and Aggarwal, 2013). These conditions cause difficulties for recognition task (Oreifej and Liu, 2013).

The aim of this paper is to develop a novel human action recognition technique, which combines skeleton- and depth-based action recognition approaches. Foremost, to deal with view-invariant issue of depth-based action recognition, we propose a novel method, robust to viewpoint variation and applicable in real-time. These methods are based on hu-

man body orientation and aggregation of several base classifiers. Furthermore, to combine the skeleton- and depth-based classifiers, we study two strategies, the first being based on fuzzy switcher, and the second on combination between fuzzy switcher and aggregation.

This paper is organized as follows: Section 2, briefly reviews the most relevant skeleton and depth based human action recognition approaches; section 3 introduces a novel Multiview public dataset and provides duration and accuracy evaluations for various state-of-the-art algorithms on our dataset; section 4 gives a description of our proposed method; Section 5 discusses the computational efficiency and the robustness of combination methods to viewpoint variations. Finally, conclusion is given in Section 6.

2 RELATED WORKS

The recent advances of RGB-D cameras have provided a new opportunity for skeleton estimation. However, the extraction of human skeleton has become relatively easier and more accurate (Shotton et al., 2011), leading to a great attention for skeleton-based action recognition approaches. These last years, many researchers focused on developing novel skeleton feature space to characterize human gestures. In (Yang and Tian, 2012), Yang et al. used the relative-joint positions representation as features to characterize the different actions, classification is done using the Naive-Bayes nearest neighbor. Wang et al. (Wang et al., 2012b) used the same relative-joint position features but they focused on the temporal modeling of different actions using Fourier Temporal Pyramid. Furthermore, to make the skeletal data invariant to the location and orientation of the subject in the scene, person-centric coordinate and normalization of the skeleton were adopted by many researches (Vemulapalli et al., 2014; Xia et al., 2012). These processes improve the robustness of skeleton-based action recognition to make them view invariant. Xia et al. (Xia et al., 2012), introduced a view-independent representation of the skeleton obtained by the orientations quantization of the body joints into histograms with respect to the hip center reference. The temporal evolution of this method is modeled using Hidden Markov Model (HMM). Vemulapalli et al. (Vemulapalli et al., 2014), present a new skeletal representation that lies in the Lie group of Special Euclidean group SE(3), describing the rotations and translations between various body segments using 3D euclidean geometric transformations. In order to simplify the approach of classification the authors project the extracted features using lie algebra. Then, they perform classification using a combination of dynamic time warping, Fourier temporal pyramid representation and linear Support Vector Machine (SVM).

Recently, depth-based action recognition has made great progress. Wang et al. (Wang et al., 2012a), utilize a sparse coding approach to encode the Random Occupancy Pattern (ROP) features to recognize the different actions. The ROP features is shown to be robust to occlusion. The approach proposed in (Yang et al., 2012) projects depth maps of action instances into three orthogonal planes. To represent the action, the authors use the Histogram of Oriented Gradients (HOG) descriptor to characterize the Depth Motion Maps (DMM), which is the accumulation of motion energy through the entire sequences of each projection. Due to its computational simplicity, the same approach in (Yang et al., 2012) is adopted in the work of (Chen et al., 2013), while the latter modifies the procedure to obtain DMMs without including HOG descriptor. As a result, the computational complexity of the feature extraction process is greatly reduced. Similarly to (Yang et al., 2012), Oreifej et al. (Oreifej and Liu, 2013) describe the depth sequence using a histogram that captures the distribution of the normal to the surface orientation in the 4D space (HON4D), composed of time, depth, and spatial coordinates. Ohn-Bar et al. (Ohn-Bar and Trivedi, 2013), propose a new descriptor for spatio-temporal feature extraction from depth images called HOG^2 in which they evaluated the extracted features in a bag-of-words scheme using linear SVM.

Rather than the recent skeleton-based approaches which are robust to viewpoint variation, the view-invariance is still a major challenge for depth-based action recognition. In fact, there are recently some researches that make the depth-based more robust to the view angle variation, such as the work of Hossein et al. (Rahmani et al., 2014),they processed directly the pointclouds of depth sequence using Histogram of Oriented Principal Components (HOPC) descriptor, which is robust to noise, viewpoint, scale and action speed variations. However, the process of pointclouds are very time consuming which makes this method so far to be applicable in real-time.

A hybrid solution combining depth and skeleton information is used in many research works. Most of them use various depth descriptors around the joints as features such as: (Wang et al., 2012b; Yang et al., 2012), and (Ohn-Bar and Trivedi, 2013). Wang et al., in (Wang et al., 2012b), use 3D joints position and local occupancy patterns as features. Ohn-Bar et al.(Ohn-Bar and Trivedi, 2013) propose to use the HOG^2 descriptor around each joints instead of whole depth pixels. Indeed, these methods are strongly de-

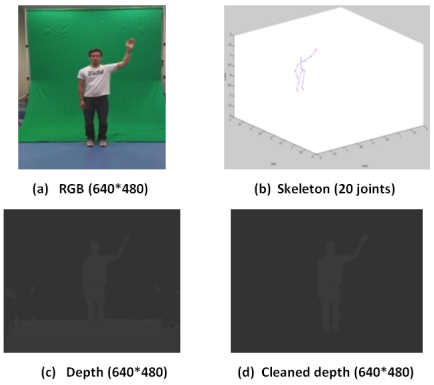

Figure 1: Snapshots of four information available in MinesDouai_3D dataset.

pendent on the accuracy of skeleton estimation, thus they do not make a substantial contribution in the case of self-occlusions or when the skeleton estimation failed.

3 BASE CLASSIFIER FOR HUMAN ACTION RECOGNITION

The goal of this section is to evaluate and analyze the most relevant depth and skeleton base classifiers of action recognition with respect to viewpoint variation, to further show the contribution of combination methods to enhance the recognition accuracy and timing performance.

3.1 Mines Douai Multiview 3D Dataset

Several 3D dataset have been created in the last years for action recognition. However, most of them are acquired from only one viewpoint angle which prevent us from measuring the effect of person's orientation. Recently, Rahmani et al. (Rahmani et al., 2014), created a multiview dataset but this dataset is not well structured. Thus, the creation of new publicly available multiview dataset to enhance the study of view-invariant action recognition became a necessary task. MinesDouai_Multiview_3D dataset contains four types of information (Figure 1): RGB, depth, cleaned depth and human skeleton data, captured with only one kinect camera. For each acquisition, the subject performs the same action twice in three orientations (30°,0°,-30°) in order to produce the dissimilarity between the three viewpoint angles even for the same subject.

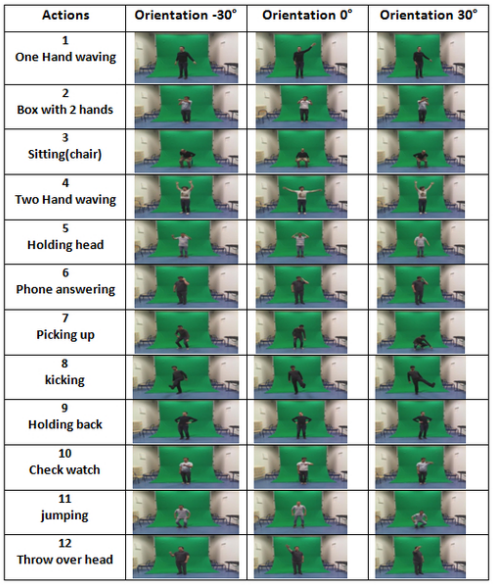

Figure 2: Samples frames of Multiview MinesDouai_3D dataset.

This dataset includes 12 actions: one-hand waving, boxing, setting, two-hand waving, holding head, phone answering, picking up, kicking, holding back, check watch, jumping, and throw over head. Each action is performed by 8 actors. Figure 2 shows some sample frames of our dataset captured from different angles.

3.2 Skeleton-Based Action Recognition

The extraction of a skeleton from depth information was a big challenge in the last decades. Recently a robust skeleton estimation algorithm proposed by Shotton et al. (Shotton et al., 2011) regenerates the interest on skeleton-based action recognition. To study the impact of viewpoint variation on the accuracy of action recognition, we conducted our experimentation with the most relevant state-of-the-art skeleton-based approaches. Foremost, we started by Actionlet algorithm (Wang et al., 2012b) and then we tested a four alternative skeletal representations on our dataset using a combination of dynamic time warping, Fourier Temporal Pyramid representation and linear SVM. This algorithm pipeline is proposed by Vemulapalli et al. (Vemulapalli et al., 2014). The four alternative skeletal representations tested on our dataset are: Joint positions (JP), Pairwise relative positions of the joints (RJP), Joint angles (JA), and Individual body part locations (BPL).

To make the skeleton invariant to the orientation and the location in the scene, Vemulapalli et al. (Vemulapalli et al., 2014) pre-process this information with

the following operations:

- For scale-invariant: the authors consider person-centric coordinate as reference and they apply the normalization by taking one of the skeletons as reference;
- For view-invariance: a rotation is applied to the skeleton such that the vector from right hip to left hip is parallel to the global x-axis.

For fair comparison between different skeleton representations, we have used the same experimental setting with all the representations. The obtained results are summarized in Table 1.

3.3 Depth-Based Action Recognition

The skeleton estimation from RGB-D image is accurate under experimental settings, but the estimation of the skeleton remains limited. As previously stated, it has difficulties to correctly work in the presence of self-occlusion between body segments and the estimation completely fails, for instance, in video surveillance when the body is not in front of the camera (Xia and Aggarwal, 2013). Designing an efficient depth representation for action recognition can give better results in these adverse conditions. After, reviewing the most relevant state-of-the-art depth-based action recognition approaches, we choose three algorithms: DMM (Chen et al., 2013), HON4D (Oreifej and Liu, 2013), and HOG^2 (Ohn-Bar and Trivedi, 2013). This selection is based on the recognition accuracy presented in their papers with different benchmarks, and on execution time. We evaluated these three algorithms on our dataset. The experimental results are presented in Table 1.

3.4 Discussion and evaluation of base classifiers

In order to evaluate the various skeleton- and depth-based classifiers, we have divided our dataset into 2 groups. Each group contains 4 subjects with three orientations (2*3 subgroups) as shown in Figure 3. To perform the test, one subgroup is used for training and the other for testing. Table 1 summarizes the results. The "same view test" refers to the average of accuracy rates obtained when the train and test subgroupes are from the same viewpoint angle. The "cross-view test" refers to the case where train and test subgroups are from different viewpoint angles. Table 1 also contains execution duration for each algorithm.

Note that the depth-based algorithms are quietly inaccurate in comparison with skeleton-based ones and especially for cross-view tests. Moreover,

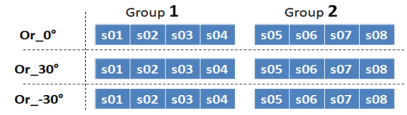

Figure 3: Six subgroups used for classification.

Table 1: Comparison between the state-of-art results with MinesDouai_Multiview_3D Dataset (bold indicates highest rate).

		Cross Validation Test (4*4)		Duration (s)
		Same View	Cross-View	
Depth	DMM	0.782	0.603	1.413
	HON4D	0.893	0.766	19.21
	HOG^2	0.878	0.742	3.083
Skel.	Actionlet	0.871	0.697	0.139
	Joint Position (JP)	0.960	0.881	0.595
	Relative Joints (RJP)	**0.977**	**0.927**	1.357
	Joint Angles (JA)	0.913	0.721	2.016
	SE3_Lie_Algebra	0.967	0.883	1.212

HON4D gives the best results in both same-view and cross-view tests among depth-based approaches. HOG^2 gives a very acceptable result compared with DMM and with HON4D (there are only a difference of 2%). To put things into perspective with this little difference between HOG^2 and HON4D, it has to be noted that the first widely outperforms the second in terms of computation time. These performances are recorded using a PC with 2.53GHz Intel Xeon CPU with 24 GB RAM (only 4GB was used).

The skeleton-based algorithms show high accuracy of recognition and robustness to viewpoint variations compared with depth-based approaches. The Relative Joint Position (RJP) provides the best improvement on our dataset, achieving a maximum of 97.7% accuracy. The Actionlet algorithm (Wang et al., 2012b) shows lower results in cross-view test.

The results of depth-based approaches in cross-view tests show the limitations of these algorithms to handle viewpoint variation. These results can be explained by the fact that the appearance of each action varies widely from one angle to the others. It is then preferable to make various base classifiers trained separately with different angles. The idea is to predict the action of an unknown subject with the appropriate base classifier which corresponds to his/her orientation.

Among several difficulties of skeleton-based approaches we can mention the lack of precision and self-occlusions caused by body parts. Note that the results in Table 1 are achieved using skeleton information acquired under experimental settings. To study the effects of instability of skeleton estimation on the performance of action recognition, we simulated self-occlusion and instability of skeleton by adding noise to some joints randomly (like in the realistic sce-

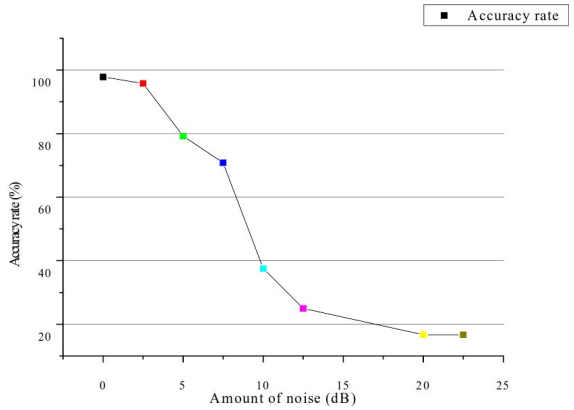

Figure 4: Recognition rate accuracy computed with respect to the amount of noise variation.

nario). Then we conducted the test using the Relative Joint Position representation (RJP), which gave the best results with our dataset. We conducted the experimentation by increasing iteratively the amplitude of the noise up to 25 dB. The obtained results are depicted in Figure 4. We can see clearly that the accuracy of action recognition using skeleton-based approach decreased rapidly when the amount of noise is greater than 3dB.

Therefore, to make the action recognition more robust, the skeleton instability should be considered by switching to another source of information such as depth which are more robust to self-occlusion. One of the simplest and efficient methods that can be used to estimate the accuracy of the skeleton data is the *JointTrackingState* information offered by Matlab™(Matworks, MA, USA) Image Acquisition Toolbox, which provides the extent of joint estimation accuracy of each frame (confidence index).

By computing the mean of confidences indexes of skeleton joints (*Conf*) for a video sequence using equation 1, we can deduce if the action recognition using skeleton information is affected by the noise or not.

$$\text{SkeletonsJointsAccuracy} = \sum_{i=1}^{p}\sum_{j=1}^{n}\frac{conf(i,j)}{n \cdot p \cdot max(conf)} \quad (1)$$

s.t. n is the number of skeleton joints, p is the number of frames in video sequence. max(conf) is the maximum reachable value for the confidence index.

4 PROPOSED ALGORITHM

In this section, we describe our method for human action recognition. In the first step, four base classifiers are used, one for skeleton-based approach and three for depth-based approach, each one corresponding to a different angle (i.e. 0°, 30°, -30°). These base classifiers work in parallel to estimate the actions performed by a subject using two different sources of information (depth and skeleton). We could extend our algorithm to cover more than three orientations, by acquiring more samples. For depth-based classifiers, we choose the human motion descriptor HOG^2, which is the fastest and that gave an acceptable accuracy in same-view test, as seen previously. For the Skeleton-based classifier, we choose the Relative Joint Position representation which gave a high accuracy. The second step of our proposed method is the combination of these base classifiers outputs to achieve more accurate action recognition. The details of the introduced methods are highlighted in the following subsections.

4.1 Depth-skeleton Fuzzy switcher Algorithm (DSFSA)

To design the overall action recognition framework, we must consider the skeleton instability and the human body orientation. A fusion algorithm that takes into account the limitations of each base classifier is therefore necessary. We found that Fuzzy Switcher is well suited for this task, as depicted in Figure 5. The proposed DSFSA Algorithm is an expert rule-based method for choosing the best base classifier. Our algorithm uses two inputs and one output, as shown in Figure 6.

Here, the term "fuzzy switcher" refers to the process of combining two sets of information to produce a better output (Singhala et al., 2014). That is, we need to use the skeleton information when all joints (or at least most of them) are accurate. In the presence of self-occlusion, which may lead to degradation of skeleton-based classifiers, we have to eliminate its use temporally by switching to depth-based one. Furthermore, as explained in the previous section, it is more accurate to predict the actions of a subject with the depth-based classifier that corresponds to the subject orientation with respect to the camera. Thus, we also switch between different depth-based classifiers to produce better action recognition.

To describe the relationship between the input and the output, the following set of rules is applied :

$$\begin{cases} \text{Skel.\_Acc.} = \text{Good} \implies \text{SW} = \text{SVM}_1 \\ \text{Skel.\_Acc.} = \text{Bad} \wedge \text{Orient.} = \text{Front} \implies \text{SW} = \text{SVM}_2 \\ \text{Skel.\_Acc.} = \text{Bad} \wedge \text{Orient.} = \text{Right} \implies \text{SW} = \text{SVM}_3 \\ \text{Skel.\_Acc.} = \text{Bad} \wedge \text{Orient.} = \text{Left} \implies \text{SW} = \text{SVM}_4 \end{cases}$$

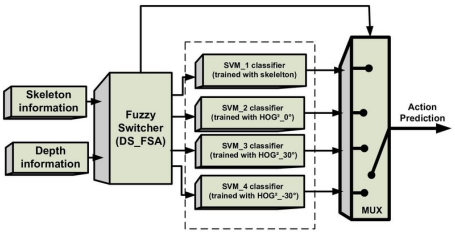

Figure 5: Overall architecture of the Depth-skeleton Fuzzy Switcher Algorithm.

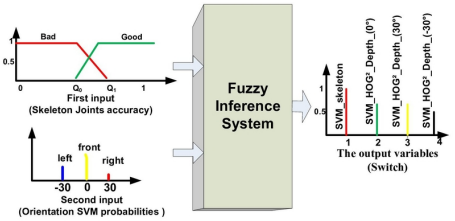

Figure 6: Architecture of the fuzzy logic system.

The fuzzy rules are directly derived from the two basic rules defined at the beginning of this subsection (related to skeleton accuracy and human body orientation). In our case, the output of the fuzzy inference system, SW (switch), is a dimensionless weighting factor that emphasizes one of the four classifier as the best for this recognition.

As shown in the Figure 6, the skeleton joint accuracy input is represented by two membership functions. The parameters (Q_0, Q_1) are estimated using the Fuzzy C_mean (FCM) algorithm. FCM is a method of clustering which provides the cluster centers and the degree of affiliation for each data point (Dwi Ade Riandayani, 2014). This information can be used to construct fuzzy inference system. Fuzzy partitioning is accomplished by an iterative optimization of the following objective function (Eq. 2),

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^m ||x_i - c_j||^2, \quad 1 \leq m < \infty \qquad (2)$$

s.t. m is a real number greater than one, x_i is the i^{th} measured data, c_j is the center of the cluster, and u_{ij} is the membership's degree of x_i in cluster j. The update of cluster centers c_j and the membership u_{ij} are determined by equation 3.

$$u_{ij}^m = \frac{1}{\sum_{k=1}^{C} (\frac{||x_i - c_j||}{||x_i - c_k||})^{\frac{2}{m-1}}}, \qquad c_j = \frac{\sum_{i=1}^{N} u_{ij}^m \cdot x_i}{\sum_{i=1}^{N} u_{ij}^m} \qquad (3)$$

When the condition $max_{ij}|u_{ij}^{k+1} - u_{ij}^k| < \Im$ is satisfied the iteration will stop, where $\Im$ is the ending criterion between [0,1], and k are the iteration steps. This process converges to a local minimum of J_m. In our experimentation, we have set the ending criterion $\Im = 1e-5$ and the number of iteration k=100. We obtained the couple (Q_0, Q_1)=(0.356, 0.716).

Several popular methods for defuzzification exist in the literature such as max-membership principle, centroid method, weighted average method, center of sums, etc. (Singhala et al., 2014). In our algorithm, the output of the fuzzy inference system is a dimensionless weighting factor. Therefore, the weighted average defuzzification technique is well suited, and is given by the following algebraic expression (eq. 4):

$$Defuz = \frac{\sum_{i=1}^{4} P[i] \cdot W[i]}{\sum_{i=1}^{4} W[i]} \qquad (4)$$

where *Defuz* is the output of fuzzy inference system, *P[i]* is the extremum value of i^th output membership function and W is the weight of the i^th rule.

4.2 Estimation of human body orientation

To recognize the action of subjects with the appropriate base classifier that corresponds to his/her orientation, an accurate and robust orientation estimator is required. In fact, it is a challenging task for many researches, due to the wide variety of poses, actions and body size which degrade the accuracy of estimation. However, the existing methods in the literature are always based on gait cue considering that the body orientation is nearly parallel to the moving trend (Shinmura et al., 2015; Liu et al., 2013). Indeed, for human-robot interaction usually the subjects are in standing position and their actions or gestures are limited to the movement of some body's segments, such as *check a watch*. These kinds of actions make the previous methods unable to handle the different body orientations. To address these challenges, we propose a new method based on a RGB-D sensor.

The proposed method is inspired from the work of (Ozturk et al., 2009) in which they used RGB stream of a camera mounted on the top of a roof. Then, the combination of Shape Context and Scale-Invariant Feature Transform (SIFT) features are used to estimate the body orientation by matching the upper region of the body with predefined shape templates.

In our method we used the depth data to capture the body orientation by projecting the depth of each frame onto an orthogonal Cartesian planes (Yang et al., 2012; Chen et al., 2013). Then, we calculated Depth Map (DM) of top-view projection to characterize the orientation as shown in Figure 7. The DM is obtained by accumulating the absolute difference between two consecutive depth images across n frames

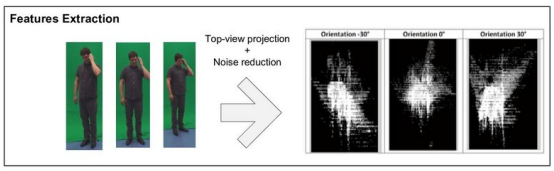

Figure 7: Three examples of DMs generated from different view-point samples of Mines_douai_3D dataset.

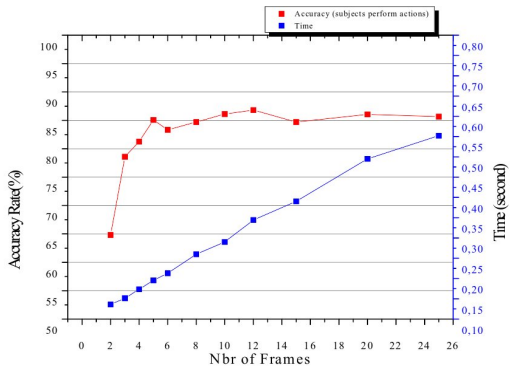

Figure 8: Accuracy and timing performances computed with respect to the number of frame variation.

of a video sequence as following (eq 5):

$$DM_{top} = \sum_{i=a}^{b} |map^{i}_{top} - map^{i-1}_{top}| \qquad (5)$$

where i is the frame index, map^{i}_{top} top-view projection of the i^{th} frame, and $[a,b] \in [2,n]^2$ denote the first and last frame indexes.

We have fixed the same size of the DMs for all samples. For training and testing feature sets, principal component analysis (PCA) was used to reduce the dimensionality in order to optimize computational time.

To test the proposed algorithm, one half of the samples of our dataset are used for training and the other half for testing. An average recognition rate of 86.1% (248/288) was achieved. Figure 8 describes the changes in the accuracy rate of orientation estimation and the computing times, with respect to the number of frames used to compute the DMs. When 2 frames were used to characterize the human body orientation, the estimation was lower. When the number of frames exceed 5, the estimation accuracy remains around 85%.

4.3 Aggregation

Instead of using human body orientation as discriminative criterion to switch between the three depth-based classifiers, we propose another method of combination based on outputs aggregation as shown in Figure 9. Indeed, we only focus on the aggregation of depth-based classifiers without including the skeleton-based classifier. On the one side, the switched method based on the skeleton instability estimation (eq. 1) has shown its robustness. On the other side, it is difficult to know the variation of accuracy of skeleton-based classifier due to its rapid degradation in the case of self-occlusion or when the estimation of the skeleton failed. The aggregation of the three depth-based classifiers is based on the class likelihood provided by each base-classifier (SVMs). The class likelihood information are obtained using *libsvm*.

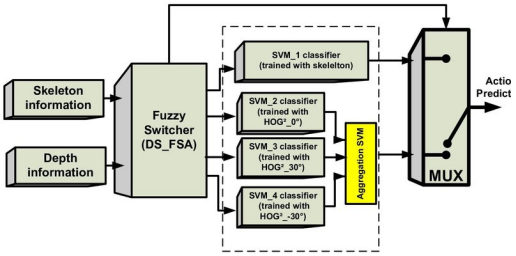

Figure 9: Overall architecture of the DS_FS Algorithm with Aggregation strategy.

After training the three base-classifiers separately, we consider two linear combination techniques: the Majority Voting and the Investment.

4.3.1 Majority Voting (Kim et al., 2002)

It is the simplest method for combining various SVMs. Let f_k be the output of the k^{th} SVM, C_j be the label of the j^{th} class, and N_j is the number of SVMs whose decision is the j^{th} class. The final decision for an input vector x is given by:

$$f_{mv}(x) = \underset{j}{\operatorname{argmax}}\, N_j \qquad (6)$$

4.3.2 Investment

Rather than the voting approach that treats all depth-based classifiers equally, investment technique aims to infer reliability degree for each base-classifier (Li et al., 2015; Pasternack and Roth, 2010). In this approach the classifiers uniformly "invest" their reliability through their claimed values (the class). The confidence in each claim increases proportionally with a non-linear function G(x)= x^g where $g = 1.2$ (eq 7),

$$\mathrm{B}(v) = \left(\sum_{s \in S_v} \frac{w_s}{|V_s|} \right)^{1.2} \qquad (7)$$

where v is the class, S_v is the set of depth-based classifiers that provide this class, and $|V_s|$ is the number

of claims made by the classifier s. The depth-based classifiers trustworthiness is computed by the sum of the confidence in their decisions, weighted by the ratio of trust previously contributed to each (relative to the other base classifiers s'), as described in (eq 8).

$$w_s = \sum_{v \in V_S} \mathrm{B}(v) \cdot \frac{\frac{w_s}{|V_s|}}{\sum_{s' \in S_v} \frac{w_{s'}}{|V_{s'}|}} \quad (8)$$

5 EXPERIMENTAL RESULTS

In order to evaluate the performance of the proposed methods, several tests were performed using fuzzy switcher algorithm and the proposed methods of aggregation (Majority Voting and Investment) on our dataset. Firstly, we highlight the advantages of using various depth-based classifiers to reduce execution time and enhance accuracy of depth-based action recognition. Then, we present the results of combination between depth and skeleton-based classifiers.

5.1 Combination of Depth-base-classifiers

To show the advantages of combining various depth-based classifiers in term of computation and accuracy performances, we conducted the test with the proposed fuzzy switcher algorithm using only the human body orientation and the two methods of aggregation without including skeleton-based classifier.

Firstly, we train three SVMs (depth-based), separately, with 4 subjects in each viewpoint orientation. Then, we choose 4 unseen subjects that perform 12 actions with different viewpoint angles and we test the three trained depth-based classifiers independently. The results are presented in Table 2.

Secondly, we conduct the tests with fuzzy switcher algorithm using subjects orientation angles. When we use the real subjects orientation labels (given by the ground truth), the average accuracy of the recognition is 95.14%. This result decreases by 2% (to 93.74 %), when we use the estimated subjects orientations provided by the proposed human body orientation estimator. The estimator achieves a 88.19% accuracy rate in this test. The recognition rates are also calculated using both aggregation methods. The majority voting aggregation gives the worst accuracy of 90.97%, among the other combination techniques, whereas the Investment aggregation gives the best results with rate of 95.14%.

Finally, we test another approach based on a single SVM trained with the three orientations. The features used for training and testing are the concatenation of HOG^2 and DMs features. This concatenation gives a good rate of accuracy as shown in Table 2. However, this methods takes a long time for training and for testing compared with the previous approaches, due to the large size of the features vectors and of the training set.

Table 2 shows clearly that the combination of different base-classifiers enhances the accuracy. As far as execution time is concerned (Table 3), the extraction of features from depth sequence using HOG^2 descriptor takes only 0.1s/frame. For the whole recognition pipeline, including extraction of features and classification, it reaches a maximum of 3s/sequence for average of 35 frames. These results make the combination approaches applicable in real-time, rather than the previous view-invariant depth-based algorithms that are more time consuming. Furthermore, the fuzzy system and skeleton instability estimation take less than 20 ms and the aggregation block takes only 0.13s. These results confirm the appropriateness of our algorithm for real time applications.

5.2 Combination between depth and skeleton-base-classifiers

In the second stage of our experimentation, we included the skeleton information. We conducted our test using Depth_skeleton Fuzzy switcher Algorithm (DS_FSA). In the following, we present only the case when the skeleton information are not accurate by adding a Gaussian White noise of 5dB to half of the testing samples alternatively (cf. Figure 10). Furthermore, the viewpoint angles of testing samples are taken randomly as shown on Figure 11.

Figure 12 shows the output of Fuzzy Inference System. When the skeleton estimation is good the DS_FSA switch to the skeleton-based classifier. In the case of skeleton instability (skeletons are noisy), the DS_FSA detects this instability easily, and does not use the skeleton information by switching to one of the three depth-based classifiers. Similarly, for the other methods of aggregation, the DS_FSA switches to the output of aggregation block, as seen on Figure 9. The results are detailed in Table 4.

We tested two variants of combination methods for DS_FSA. On one hand, we included the subject orientation angles provided by our proposed human body orientation estimator. On the other hand, we applied the investment aggregation method. The results show that the combination of depth and skeleton information notably enhances the accuracy. Moreover, the use of investment aggregation provides the best average accuracy of 97.91%. As result, the combina-

Table 2: Recognition rates for various test of depth-base-classifiers.

	HOG^2_SVM_2 (0°)	HOG^2_SVM_3 (30°)	HOG^2_SVM_4 (-30°)	Estimation of Orientation	Fuzzy Switcher (only depth)	Aggregation		One SVM
						Voting	Invest	
Test1	91.67	88.88	88.88	86.11%	97.22	97.22	97.22	97.22
				Truth 100%	97.22			
Test2	88.88	83.33	75	83.33%	88.88	86.11	91.67	91.67
				Truth 100%	91.67			
Test3	80.55	77.77	77.77	83.33%	91.67	86.11	91.67	91.67
				Truth 100%	94.44			
Test4	97.22	80.55	77.77	100%	97.22	94.44	100	97.22
				Truth 100%	97.22			
Mean	89.58	83.63	79.66	88.19	93.74	90.97	95.14	94.44
				Truth 100%	95.14			

Table 3: Time Performances.

Algorithm	Time consumption (sec.)
Skeleton-based classifier (RJP)	1.357 (per video sequence)
Depth-based classifier (HOG^2)	3.083 (per video sequence)
Skeleton accuracy	0.006
Body-orientation Estimator (10 frames)	0.290
Fuzzy system	0.0181
Aggregation (investisment)	0.1312
Aggregation + vote	0.1067

Table 4: Comparison with the different combinations.

Algorithm	Accuracy
SVM_1_Skelton_RJT (1/2 skeletons noisy 5dB)	**79.17**
SVM_2_HOG^2_0°	**89.58**
SVM_3_HOG^2_30°	**83.63**
SVM_4_HOG^2_-30°	**79.66**
Fuzzy + **viewpoint estimation** (Depth+Skelton)	**95.83**
Fuzzy + **Aggregation (Investment)**	**97.91**

tion between the DS_FSA and investment aggregation is the most appropriate framework that provides the best accuracy and timing performance.

6 CONCLUSION

In this paper we introduced a novel approach of combining skeleton and depth information based on the amalgamation of several base classifiers. Furthermore, we proposed a new method for the estimation of human body orientation which is completely independent of human gait direction. In order to perform view dependency tests, we have created a new multiview public dataset[1]. The results showed that our proposed method is robust to viewpoint variation and skeleton accuracy degradation. Furthermore, due to its computation time efficiency, it is suitable for real-time application.

Future work includes the extension of our algorithm to more than three viewpoint orientations. Since it is not a trivial task to acquire sufficient data for each viewpoint angle, an efficient solution can be used to extend any 3D action dataset to different viewpoint using point-clouds transformation.

[1]Information to dowload the dataset can be found at: `http://ia.ur.mines-douai.fr/en/datasets/`

REFERENCES

Chen, C., Liu, K., and Kehtarnavaz, N. (2013). Real-time human action recognition based on depth motion maps. *Journal of Real-Time Image Processing*, pages 1–9.

Dwi Ade Riandayani, Ketut Gede Darma Putra, P. W. B. (2014). Comparing fuzzy logic and fuzzy c-means (fcm) on summarizing indonesian language document. *Journal of Theoretical and Applied Information Technology*, 59(3):718–724.

Kim, H.-C., Pang, S., Je, H.-M., Kim, D., and Bang, S.-Y. (2002). Support vector machine ensemble with bagging. In *Pattern recognition with support vector machines*, pages 397–408.

Laptev, I. and Lindeberg, T. (2003). Space-time interest points. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 432–439.

Li, Y., Gao, J., Meng, C., Li, Q., Su, L., Zhao, B., Fan, W., and Han, J. (2015). A survey on truth discovery. *arXiv preprint arXiv:1505.02463*.

Liu, W., Zhang, Y., Tang, S., Tang, J., Hong, R., and Li, J. (2013). Accurate estimation of human body orientation from rgb-d sensors. *IEEE Transactions on Cybernetics*, 43(5):1442–1452.

Ohn-Bar, E. and Trivedi, M. (2013). Joint angles similarities and hog2 for action recognition. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 465–470.

Oreifej, O. and Liu, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recogni-*

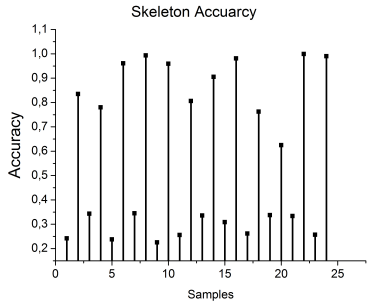

Figure 10: Variation of sketeleton accuracy estimation.

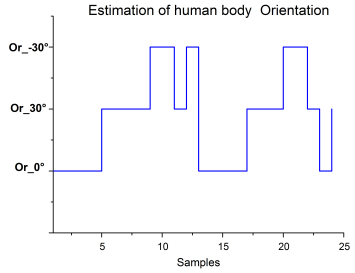

Figure 11: Estimation human body Orientation using DM algorithm.

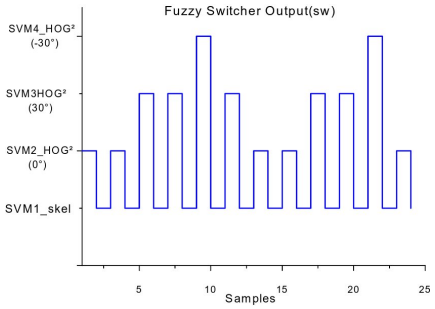

Figure 12: Output of Depth skeleton Fuzzy switcher Algorithm (DS_FSA) during testing

tion (CVPR), 2013 IEEE Conference on, pages 716–723.

Ozturk, O., Yamasaki, T., and Aizawa, K. (2009). Tracking of humans and estimation of body/head orientation from top-view single camera for visual focus of attention analysis. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1020–1027.

Pasternack, J. and Roth, D. (2010). Knowing what to believe (when you already know something). In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 877–885.

Peng, X., Wang, L., Qiao, Y., and Peng, Q. (2014). A joint evaluation of dictionary learning and feature encoding for action recognition. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 2607–2612.

Rahmani, H., Mahmood, A., Q Huynh, D., and Mian, A. (2014). Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision ECCV 2014*, volume 8690 of *Lecture Notes in Computer Science*, pages 742–757.

Shinmura, F., Deguchi, D., Ide, I., Murase, H., and Fujiyoshi, H. (2015). Estimation of human orientation using coaxial rgb-depth images. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 113–120.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 1297–1304.

Singhala, P., Shah, D. N., and Patel, B. (2014). Temperature control using fuzzy logic. *International Journal of Instrumentation and Control Systems*, 4(1):110.

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

Wang, J., Liu, Z., Chorowski, J., Chen, Z., and Wu, Y. (2012a). Robust 3d action recognition with random occupancy patterns. In *Computer vision–ECCV 2012*, pages 872–885.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012b). Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297.

Xia, L. and Aggarwal, J. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2834–2841.

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

Yang, X. and Tian, Y. (2012). Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 14–19.

Yang, X., Zhang, C., and Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proc. of the 20th ACM international conference on Multimedia*, pages 1057–1060.