Skeleton point trajectories for human daily activity recognition

A. Chan-Hon-Tong, N. Ballas, C. Achard, B. Delezoide, L. Lucat, P. Sayd, F. Prêteux

Skeleton Point Trajectories for Human Daily Activity Recognition

Adrien Chan-Hon-Tong[1], Nicolas Ballas[1], Catherine Achard[2], Bertrand Delezoide[1], Laurent Lucat[1], Patrick Sayd[1] and Françoise Prêteux[3]

[1]*CEA, LIST, DIASI, Laboratoire Vision et Ingénierie des Contenus, Gif-sur-Yvette, France*

[2]*Institut des Systèmes Intelligents et Robotique, UPMC, Paris, France*

[3]*Mines ParisTech, Paris, France*

Keywords: Skeleton Trajectory, Human Activity Classification.

Abstract: Automatic human action annotation is a challenging problem, which overlaps with many computer vision fields such as video-surveillance, human-computer interaction or video mining. In this work, we offer a skeleton based algorithm to classify segmented human-action sequences. Our contribution is twofold. First, we offer and evaluate different trajectory descriptors on skeleton datasets. Six short term trajectory features based on position, speed or acceleration are first introduced. The last descriptor is the most original since it extends the well-known bag-of-words approach to the bag-of-gestures ones for 3D position of articulations. All these descriptors are evaluated on two public databases with state-of-the art machine learning algorithms. The second contribution is to measure the influence of missing data on algorithms based on skeleton. Indeed skeleton extraction algorithms commonly fail on real sequences, with side or back views and very complex postures. Thus on these real data, we offer to compare recognition methods based on image and those based on skeleton with many missing data.

1 INTRODUCTION

Human activity recognition is becoming a major research topic (see (Aggarwal and Ryoo, 2011) for a review). The ability to recognize human activities would enable the development of several applications. One is intelligent video surveillance in a medical context to monitor, at home, people with a limited autonomy (elderly or disabled person). Such systems could detect, in a non-invasive way, events affecting people safety such as falls or fainting and warn automatically the medical assistance. Human activity recognition could also lead to the construction of gesture-based human computer interface and vision-based intelligent environment.

During the last decade, the analyse of natural and unconstrained videos has known many improvements, such as (Laptev et al., 2008; Liu et al., 2009; Rodriguez et al., 2008). These improvements were driven by recent progresses in object recognition in static image. In this field, most of the state-of-the-art approaches are based on the standard bag-of-words pipeline (Sivic and Zisserman, 2003) that couples low-level features like (Heikkilä et al., 2009; Lazebnik et al., 2005; Lowe, 1999) with semantic understanding.

Besides, some works taking place in a video-surveillance setting, with a constrained environment, explore the characterization of human activities using 3D features (Li et al., 2010; Ni et al., 2011), or, middle-level information related to person poses. 3D features based methods mainly involve bag-of-words pipeline and extends low-level features by adding 3D information provided by depth-map.

Skeleton based methods rely on the extraction of the body-part positions in each frame of a video or motion capture sequence. Despite having a suitable accuracy when based on markers like in pioneer works (Campbell and Bobick, 1995), focused on specific body parts in (Just et al., 2004) or being adapted to dedicated applications such as human sign language recognition (Bashir et al., 2007), skeleton based methods were first not suited to the recognition of various natural actions. But, recent devices like low-cost accelerometers (Parsa et al., 2004) or KINECT allow the use of skeleton information to characterize general activities.

In this work, we take advantage of these recent

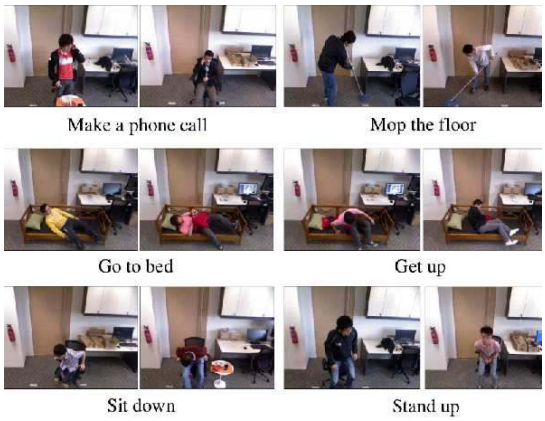


Figure 1: RGBD-HuDaAct dataset (Ni et al., 2011).

progresses and present an algorithm for human-action sequence recognition. This algorithm is based on skeleton data, and on trajectory descriptors. Trajectory of each skeleton articulation is associated with a descriptor. This descriptor is invariant to rotations (toward the vertical axis) and translations of the body, and, captures both trajectory short-term information (throw an extension to 3D of (Ballas et al., 2011) state-of-the-art descriptors) and trajectory middle-term information (throw elementary gesture recognition). These descriptors (one for each skeleton articulation) form the sequence signatures which are analysed by a Multiple-Kernel Learning (MKL).

In addition, we study the effect of missing skeleton on our pipeline: despite large academic efforts to deal with general pose estimation (Baak et al., 2011; Girshick et al., 2011; Shotton et al., 2011), local failures in skeleton extraction are common in real-life color-depth-videos (RGBD-video). Hence, we extend works of (Yao et al., 2011) where effects of noise on joint positions on detection results is studied, by focusing on our classification results according to the degree of failures in skeleton extraction on the daily-life sequences provided in the RGBD-HuDaAct dataset (figure 1).

In this paper, related works are first reviewed in section 2. Studied trajectory descriptors are presented in section 3 and tested in section 4 on public datasets (figure 2) where complete skeleton data are provided (as part of the dataset). Then, impact of non-artificial failures in skeleton extraction on the classification performance of our algorithm is presented in section 5.

Figure 2: CUHA dataset (first column-(Sung et al., 2011)) and TUM dataset (second column-(Tenorth et al., 2009)).

2 RELATED WORKS

Trajectory features have been introduced in the context of human action recognition to capture video motion patterns through long-term analysis. In (Matikainen et al., 2010), trajectory motion vectors form directly non-fixed-length trajectory descriptors. In (Raptis and Soatto, 2010), the average of descriptors of points composing the track over time is taken as the trajectory descriptor, resulting in a fixed-size descriptor. However, this approach discards the temporal information of the trajectory. To tackle this issue, a Markov process is used in (Messing et al., 2009; Sun et al., 2009): elementary motions are quantized and transitions between motions words are modelled through a Markov model to represent trajectory. In (Mezaris et al., 2010), Multiple Haar filters extract motion information at different time-scales, and, in (Ballas et al., 2011), motion and velocity information are combined to form trajectory descriptor.

After the descriptors extraction, an aggregation scheme transforms the local descriptors into a global video signature. In (Ballas et al., 2011; Raptis and Soatto, 2010; Sun et al., 2009), this aggregation relies on the bag-of-words model, while in (Messing et al., 2009), the aggregation is obtained through a Gaussian

Mixture Model.

All these approaches rely on numerous salient patches with low semantic meaning to construct the video signature. However, it has been noticed by the well-known moving light experiment (Johansson, 1973) that a human is able to recognize a human action only from the set of skeleton articulations. This experiment, coupled with the recent success in skeleton extraction based on depth map provided by active captor, invites to explore the idea of human action analysis based on skeleton data.

Previous works have explored human skeleton-based features for human action recognition. In (Yao et al., 2011), distances between skeleton articulation are used as weak features and Hough-forests framework as detector. They prove that Hough-forests using both skeleton and image features achieve better performance than forests using only image features. In (Raptis et al., 2011), a system based on skeleton pose estimation is presented. The system considers 120 frames-long skeleton trajectories and applies the maximum normalized cross correlation framework to recognize dance gestures. In (Sung et al., 2011) a two-layered maximum entropy Markov model (MEMM) is built on pose based features, while in (Tenorth et al., 2009), pose based features are considered in a conditional random field (CFR) context. Finally in (Barnachon et al., 2012), activities are represented as words on a gesture alphabet, this allows to use an automaton to recognize activity languages. Those previous works have proven the relevancy of the skeleton-based features for human activity recognition.

In this work, we apply trajectory descriptors to skeleton data to form sequence signatures, invariant both in rotation and translation. These signatures are then used to classify sequences throw a machine learning algorithm. The offered algorithm is tested on both CUHA dataset (Sung et al., 2011) and TUM dataset (Tenorth et al., 2009).

In addition, we study the impact of skeleton extraction failure on our classification results. In every previous works on human action recognition based on skeleton data, skeleton extraction failure is rarely considered. In (Yao et al., 2011), the correct-detection rate is evaluated when adding noise on joint positions. The offered algorithm provides a Gaussian noise robustness on detection performance up to a standard deviation of $0.27m$. Following this work, we present the performances of our system when skeleton extraction is only intermittent as frequently observed in natural sequences.

3 SKELETON DATA BASED SEQUENCE SIGNATURE

The computation of our sequence-signature based on skeleton data is divided into several steps. First, trajectories are normalized in order to be invariant according to rotation around the vertical axis and translations. Then, one descriptor capturing the trajectory information is computed for each articulation. These descriptors handle intermittent data to manage cases where skeleton extraction failed. Finally, this set of trajectory descriptors forms the final sequence signature.

3.1 Pose Extraction and Normalization

In order to be independent of the skeleton data across datasets and existing systems (such as NITE), we choose to use only trajectories of feet, knees, hips, shoulders, head, elbows and hands articulations as input trajectories.

In these trajectories, the 3D coordinates of the skeleton articulations are expressed in a system of coordinates which is linked to the camera position. To recognize activity under various viewpoints, a normalization scheme has to be applied resulting in articulation coordinates invariant to geometric transformations. For instance, translation, and rotation invariance is achieved in (Raptis et al., 2011; Yao et al., 2011). In (Yao et al., 2011), skeleton information is reduced to a set of features based on distances between articulations. In (Raptis et al., 2011), the skeleton articulation coordinates are expressed in a new system of coordinates defined by the principal axis of torso points. However, with complete rotation invariance, some activities such as *lie-down* and *standing* become indistinguishable. To ensure the discriminative power of our final descriptor, we choose to be invariant only toward skeleton rotations around the vertical-axis and to skeleton global translations. To reach this invariance, articulation coordinates are expressed in a new system of coordinates (O,ux,uy,uz) where the origin is the center of the shoulders, u_z is an estimation of the vertical vector (built using the video vertical vector), u_y is the orthogonal projection of the vector connecting the left shoulder to the right shoulder on the horizontal plane determined by u_z, and u_x is set so that (u_x, u_y, u_z) defines an orthonormal system (figure 3).

Using these new coordinates, we define 6 short-term descriptors which model one frame information and a middle-term descriptor which models elementary gestures.

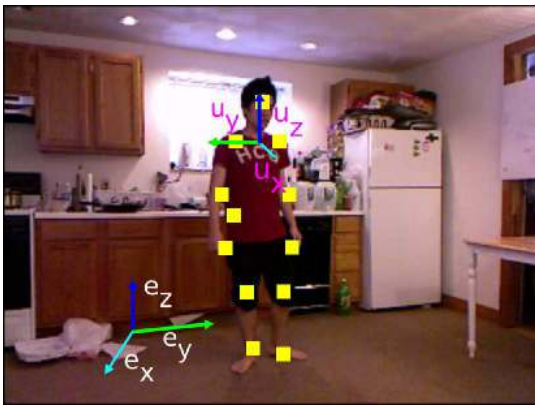


Figure 3: Schema of skeleton normalization: $u_z = e_z$, u_y is the vector between the two shoulders projected on the plane $z = 0$ and the 0 is the center of the two shoulders.

3.2 Short-term Descriptors

Short-term modelling focuses on capturing the distribution of instantaneous information in the sequence represented as $(v_1, ..., v_I)$: v_i ($i \in \{1, ..., I\}$ where I is the length of the sequence) is a vector of position, velocity (difference of position between two consecutive frames) or acceleration (difference of velocity between two consecutive frames). To achieve scale invariance, vectors are normalized according to the trajectory maximum vector magnitude $max_{i=1..I}||v_i||_2$. Then, the vectors v_i are quantified to estimate their distributions. We propose a 3D extension of the quantification from (Sun et al., 2009) where polar grid is used to quantize both vector direction and vector magnitude. As final classification performances are empirically stable relatively to this quantification, we introduce the least change from (Sun et al., 2009) quantification. Hence, we use spherical coordinates with 3 equal bins for vector magnitude and respectively 2 and 4 uniform bins for inclination and polar angles. An additional bin is added to represent the null vector resulting in a 25 bins quantification as in (Sun et al., 2009).

Two kinds of models are considered for the trajectories. First, a simple histogram is estimated resulting in a 25 dimensional vector.

As an histogram is an orderless representation, it does not take into account the temporal relation between the elementary vectors. To complete the previous representation, we also consider the Markov Stationary Features (Ni et al., 2009; Sun et al., 2009) that enforces the temporal consistency by considering the temporal co-occurrence statistics. The stochastic matrix counting the co-occurrences of successive bins is computed. As this matrix belongs to a high dimensional space (25×25), it can not be directly used as descriptor and the stationary distribution associated to the Markov process (Breiman, 1992) is used instead.

Both representations are L_1 normalized in order to be invariant to the action duration.

It results in 6 different descriptors for each articulation: Short Term Position, Short Term Motion, Short Term Acceleration, Short Term Markov Position, Short Term Markov Motion, Short Term Markov Acceleration. In addition to these descriptors, a descriptor based on middle-term modelling is also introduced.

3.3 Gesture based Descriptor

An activity can be described as a succession of elementary gestures, defined by several frames of the sequence. Contrary to (Barnachon et al., 2012) where a rough segmentation is performed based on trajectory discontinuities to extract non-overlapping gestures, we choose to adopt a dense and overlapping approach to be independent from ad-hoc gesture-segmentation algorithm, and, to ensure that all discriminative gestures are extracted (figure 4.a). We consider each window (with different allowed sizes) of each trajectory as a gesture. In this work, the set of allowed sizes is $\{0.2, 0.4, 0.6, 0.8\}$ second. Hence, let $(v_1, ..., v_I)$ be the sequence of positions measured at frequency f, then for all size $s \in \{0.2, 0.4, 0.6, 0.8\}$ and for all offset $i \in \{1, ..., I - sf\}$ the sub-trajectory $v_i, v_{i+1}, ..., v_{i+sf}$ is a gesture (when skeleton is available in all frames of the window).

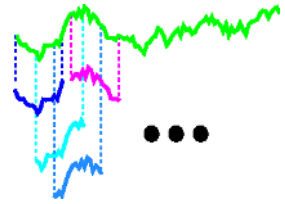
A bag-of-words model (Sivic and Zisserman, 2003) is then used to capture the gesture distribution contained in a trajectory. A vocabulary of gestures is learned through an unsupervised clustering performed by the K-means algorithm (figure 4.b). K-means is done for each gesture size and articulation separately. Then, each articulation trajectory is described with the corresponding histogram (figure 4.c) which forms our bag-of-gestures descriptor (BOG).

3.4 Classification based Multiple Kernel Learning

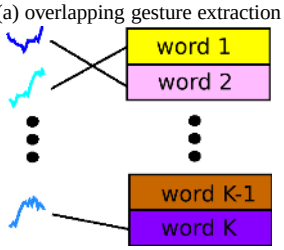
Our sequence signatures are composed by 13 trajectory descriptors - one for each skeleton articulation. To exploit this multiple channels representation, a Multiple-Kernels support machine is used along with a χ^2 kernel. The χ^2 distance between two vectors u, v of size N is given by

$$D_N(u,v) = \frac{1}{2}\sum_{n=1}^{N}\frac{(u_n - v_n)^2}{u_n + v_n}$$

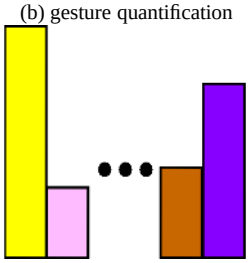
Let A and B be two sequence signatures and let H be



(a) overlapping gesture extraction



(b) gesture quantification



(c) BOG trajectory descriptor

Figure 4: Bag-of-gesture descriptor for a trajectory.

the size of the corresponding trajectory descriptor, then the kernel value for A, B is

$$K(A,B) = exp(-\sum_{s=1}^{13} \beta_s D_H(A_s, B_s))$$

where β_s is the weight associated to the channel s (a specific articulation here).

Optimal β_s are automatically determined using training samples through MKL implemented by SHOGUN library (Sonnenburg et al., 2010).

4 EVALUATION ON TUM AND CUHA DATASETS

We evaluate all introduced trajectory-based sequences signatures on two public datasets using the same MKL algorithm for classification. In addition, we also evaluate combination of trajectory descriptors.

4.1 Datasets

The CUHA dataset (for Cornell University Human Activity) is presented in (Sung et al., 2011). It deals with 14 classes of daily-life activities performed by four people (figure 2). It contains 68 sequences around 30 seconds each. The evaluation process consists in a leave-one-subject-out cross validation in precision-recall terms.

The TUM dataset is presented in (Tenorth et al., 2009). It deals with 10 action classes occurring when setting a table (figure 2). It contains 19 realistic sequences around 2 minutes each, performed by 5 people. Evaluation from (Yao et al., 2011) consists to split data between training and testing, and to output result in correct-classification rate term. In this dataset, each frame is associated to one action label. As our system expects segmented data (contrary to those from (Tenorth et al., 2009; Yao et al., 2011)), we split each sequence each time the action label changes. this gives around 1000 sub-sequences with homogeneous label which are used as input.

4.2 Results

The evaluation of our trajectory signatures are presented in table 1 (in percentage). These results are completed by the results of 3 well-known methods of the literature:

- The first one (MEMM) proposed by (Sung et al., 2011) uses a two-layered maximum entropy Markov Model on pose-based features.
- The second one (Yao et al., 2010; Yao et al., 2011) uses the Hough forests to classify actions associated to weak classifiers based on distances between skeleton or visual feature.
- Finally, a Dynamic Time Wrapping (DTW) algorithm combined with the nearest neighbour classification approach (Fengjun et al., 2005; Müller and Röder, 2006) has also been evaluated on these datasets.

Only the four best short-term descriptors are presented (Short Term Position, Short Term Motion, Short Term Acceleration, Short Term Markov Motion). The number of centroids for K-means algorithm which provides the best trade-off on all experiments is 60 for BOG. As we do not see any improvement by using multiple windows in the BOG descriptor, we use only 0.4 second windows for gesture extraction.

One result of this experiment is that the introduced descriptors provide complementary results depending on the databases. Short Term Position has the best performances among other Short Term descriptors on the CUHA dataset which deals with several static actions (*standing still, relaxing on a couch*) whereas Short Term Motion performs better on the TUM dataset, which contains dynamic actions (*opening a door, closing a door*). Bag-of-gestures (BOG),

Table 1: Results on CUHA and TUM datasets.

	CUHA Dataset			TUM Dataset
	Precision	Recall	$F_{0.5}$	correct-classification rate
MEMM (Sung et al., 2011)	69	57.3	64.2	
Hough forest based visual features (Yao et al., 2010)				69,5
Hough forest based skeleton features (Yao et al., 2011)				81,5
DTW baseline	83.7	74.5	77.2	76.3
Short Term Position	80.2	86.3	81.4	76.6
Short Term Motion	72.2	75.6	72.9	84.4
Short Term Acceleration	67.9	74.2	69.1	68.35
Short Term Markov	79.4	84.5	80.4	81.7
Bag-of-Gestures	**90.1**	**84.5**	**88.9**	84.5
Short Term Combination	70.5	71.4	70.7	**90.8**

which combines pose and movement on a middle-term duration, has better results than any individual short term descriptor on CUHA and TUM datasets. It can also be noted that both short-term descriptors and bag-of-gestures lead to results similar or better than those of the literature on the two databases.

The size of short-term descriptor and bag-of-gesture descriptor are really different since the first vector size is $25 \times 13 = 325$ while the second one is $60 \times 13 = 780$. Moreover, as the best short-term descriptor depends on the database, we propose to combine these short-term descriptors into a single one with size $325 \times 4 = 1300$. As shown in Figure 1, this new descriptor called short-term combination leads to better results on the TUM database and is less effective on the CUHA database. As its size is much larger, a good compromise is to use the bag-of-gesture descriptor which leads to better results than the state-of-the-art methods, regardless of the test database.

As it has been noticed by previous work, skeleton based approaches provided high performances. In this work, we show that on datasets where skeleton is provided (CUHA, TUM), our approach leads to high results and outperforms the state-of-the-art. However if skeleton based approaches seem to be robust, one may wonder what happens on real sequences where the skeleton is obtained only intermittently.

Indeed, some algorithms such as DTW combined with the nearest neighbour classification framework which relies directly on comparison between skeleton articulations positions can not deal with missing observation. In the last section, we describe the behaviour of our algorithm on intermittent data. More precisely, we show that the BOG descriptor is flexible enough to adapt to missing data and provides results comparable to state-of-the-art ones.

5 BEHAVIOUR OF THE METHOD ON INTERMITTENT DATA

In order to evaluate the performance of our system in presence of missing skeleton data, incomplete sequences could be generated from complete ones by removing data. However real-life failure can be correlated with specific part of the action and in the worst cases, with the most discriminative parts of the action. Hence, we decide to study the performance of our system on real intermittent data instead of artificial ones. In that purpose, we use the NITE software suite to extract skeleton from RGBD-HuDaAct dataset. As this database is not designed for skeleton extraction, some problems occur with side or back views of people, or when the subject has very specific postures that does not allow skeleton extraction. This provides incomplete sequences, and, we evaluate our algorithm according to various percentages of missing data.

5.1 Dataset

The RGBD-HuDaAct dataset is presented in (Ni et al., 2011). It deals with 12 classes (+ one random class) of daily-life activities performed by 30 people. It is composed by 1189 RGBD-video sequences, around 1 minute each. Contrary to other datasets like (Tenorth et al., 2009; Sung et al., 2011), RGBD-HuDaAct is not designed for skeleton extraction. In TUM dataset (Tenorth et al., 2009), skeleton stream is provided by a body-part tracker which is manually helped when tracking failures occur. In CUHA dataset (Sung et al., 2011), skeleton stream is provided by the general public NITE software suite, but there is no skeleton extraction failures, maybe be-

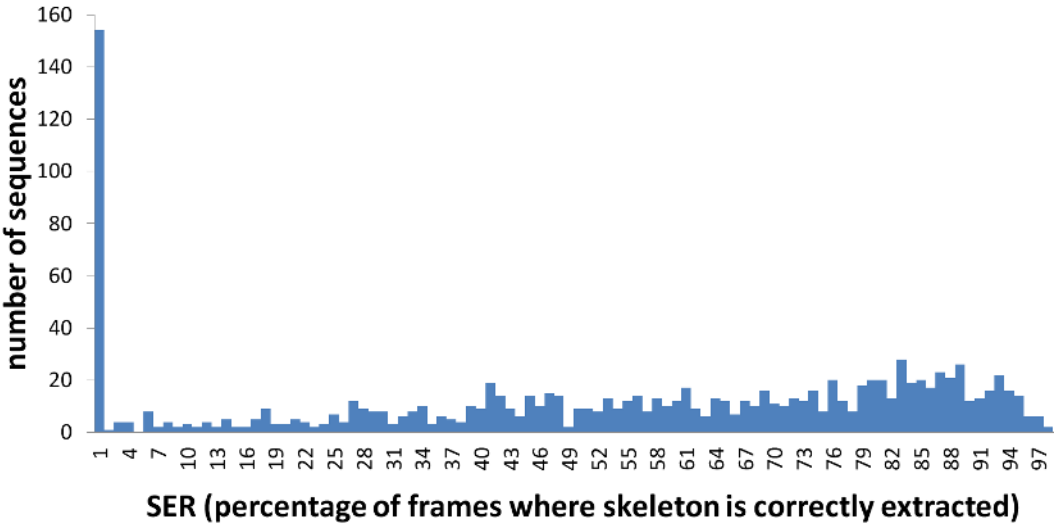


Figure 5: Number of sequences with a given SER.

cause actions are fronto-parallel to the camera and relatively constrained. In RGBD-HuDaAct, sequences seem more natural (actions are not fronto-parallel to the camera...) and skeleton extraction based on NITE software suite suffers from many failures.

NITE software provides for each skeleton articulation, a boolean describing if the position is considered as reliable by the system. Hence, we consider that the skeleton is correctly estimated in a frame, if and only if, one skeleton only is extracted from the frame and for all body-parts of interest (feet, knees, hips, shoulders, head, elbows and hands) the corresponding booleans are true.

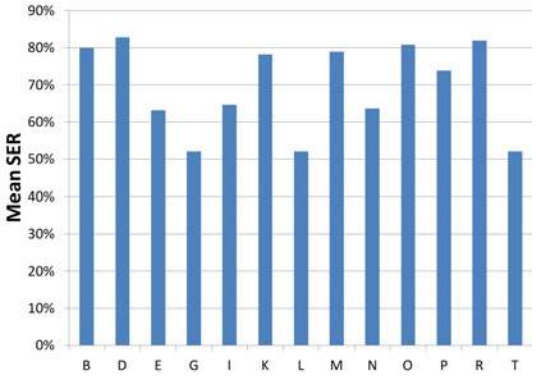
In order to measure the intensity of failures in skeleton extraction, the ratio between the number of frames where skeleton is correctly estimated and the total number of frames is computed for each sequence of the dataset. This ratio expressed in percentage is called, in this work, SER for Skeleton Extraction Ratio, and is used as a measure of the sequence validity.

The histogram corresponding to the number of sequences versus SER is presented in figure 5.

We can already notice that for more than ten percents of the sequences, the skeleton is never extracted during the entire sequence.

Figure 6 shows the average SER for each action of the database. It varies between 51.3% for *stand up (T)* and 82.3% for *put on jacket (D)*. It can be noticed that the sequences where the skeleton is not well extracted do not rely on the performed actions.

The large range of SER presented in figure 5 allows us to evaluate our action classification system with different levels of skeleton extraction failures. For this purpose, we extract all sequences with a SER greater than λ% (where λ is an integer varying from 0 to 100) and evaluate our algorithm on this subset using the leave-one-subject-out scheme as in (Ni et al., 2011).



Class labels are designed by a code letter:
go to bed (B), put on jacket (D), exit the room (E), get up from bed (G), sit down (I), drink water (K), enter room (L), eat meal (M), take off jacket (N), mop floor (O), make a phone call (P), background activities (R), stand up (T).

Figure 6: SER averages for each action class.

5.2 Classification

As the number of sequences vary with the level of SER, we have to deal sometimes with classes having a small number of sequences. Hence, we use a multiple-C-SVM framework (as suggested in (He and Ghodsi, 2010)) instead of MKL to perform classification.

Basically, in binary C-SVM, data are mapped to point in some vector space and a plane is designed to minimize the number of misclassification between

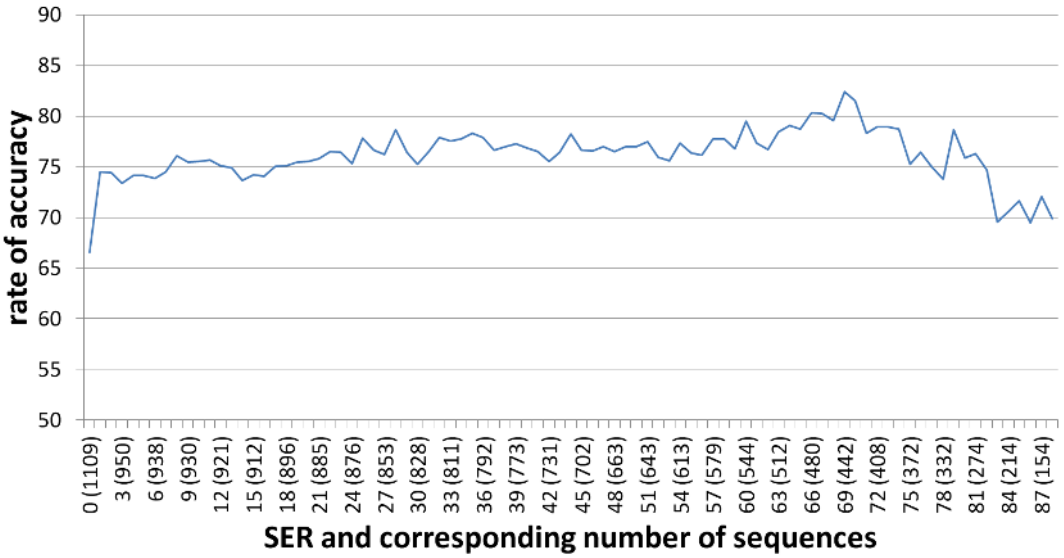


Figure 7: Evolution of accuracy versus minimal SER.

positive and negative points. But, it has been emphasized in (He and Ghodsi, 2010) that SVM classifiers efficiency decreases when positive and negative misclassification rates (number of misclassifications over number of points) are too different, which is common when data are not balanced. Hence, in binary 2C-SVM, a plane is designed to minimize both the number of misclassifications and the difference between positive and negative misclassification rates.

In practice, this is performed by dividing the misclassification cost of a point by the number of points of the corresponding class. The standard multiple-classes classifier built on the poll of one-vs-one binary 2C-SVM is then used to perform classification (Hsu and Lin, 2002). Linear-SVM implementation is provided by LIBSVM (Chang and Lin, 2011).

5.3 Results on Intermittent Data

The evolution of the correct-classification rate versus λ (minimal SER allowed) is presented in figure 7. We do not evaluate the system for λ greater than 88 as only 10% of sequences from RGBD-HuDaAct have a SER greater than 88%. The correct-classification rate is low both when there are too few data (λ close to 88) or when data contains some heavily corrupted samples (λ close to 0). However, for $\lambda = 69$ (corresponding to 442 sequences) a good compromise is found between the number of sequences and their quality: the algorithm has 82.41% of correct-classification rate.

In order to link these results to the state-of-the-art, let us remind that the best known results (Ni et al., 2011) on this dataset is a correct-classification rate of 81.5% on 59% of all sequences (655 sequences randomly sampled). Two multi-modal strategies, combining color and depth information, have been developed: spatio-temporal interest points (STIPs) and motion history images (MHIs). These two methods do not use skeleton and thus are not sensitive to the failure problems during skeleton extraction. It is not easy to predict their results on the complete dataset, or, on the 442 sequences leading to our 82.41% of correct-classification rate. Moreover, as the 655 used sequences were randomly selected, it is not possible to build their testing database in order to compare both results. However, our algorithm is competitive with their method as the recognition rates are similar. Hence, we can conclude that skeleton based approaches like our algorithm provide state-of-the-art results even on intermittent data and thus can be used for action recognition in real-life.

6 CONCLUSIONS

In this paper, we evaluate 7 trajectory descriptors in context of human action recognition based on skeleton trajectory. We first presented 6 short term trajectories descriptors based on position, speed or acceleration. The last descriptor is more original since it extends the well-known bag-of-words approach to the bag-of-gestures ones, defined only on 3D position of articulations. To our knowledge, this is the first time that bag-of-gestures are defined using 3D points.

Performances of each descriptor and combination of them associated with a same MKL classifier have been evaluated on the public CUHA and TUM

datasets, for which skeleton stream is provided. The main result is that the descriptor based on bag-of-gestures outperforms three very recent methods of the state-of-the-art on the two databases : 88.9% of $F_{0.5}$ measure on CUHA database and 84.5% of correct classification on TUM dataset.

We also study the proposed algorithm on a more difficult database not designed to extract skeleton: the public RGBD-HuDaAct database. During the skeleton estimation performed with NITE software, some problems occurred with side or back views of people, or when the person has very specific postures that do not allow skeleton extraction. Even if the RGBD-HuDaAct database is really challenging to perform action recognition based on skeletons, it has been considered here to represent some conditions of video surveillance system at home. The main result of these tests if that even in these conditions, and with a significant amount of missing data, our descriptor achieves the state-of-the-art performance of 82% of recognition rate. To our knowledge, it is the first time that failures in skeleton extraction are considered during action recognition.

In future works, we will continue to explore the links between high-level human actions and elementary gestures and design a framework to learn middle-semantic gestures which are the most relevant for human action recognition. Such framework will allow us to recognize an action from a small number of middle-semantic gestures whereas the algorithm presented in this work recognizes actions from the set of all gestures. Hence, such framework is likely to both speed up and increase accuracy of the system. In addition, we will extend current bag-of-gestures descriptor by taking into account the co-occurrence relation of different articulations and co-occurrence relation of pairs of successive gestures of a given articulation.

REFERENCES

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Comput. Surv.*

Baak, A., Muller, M., Bharaj, G., Seidel, H., and Theobalt, C. (2011). A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 1092–1099. IEEE.

Ballas, N., Delezoide, B., and Prêteux, F. (2011). Trajectories based descriptor for dynamic events annotation. In *Proceedings of the 2011 joint ACM workshop on Modeling and representing events*, pages 13–18. ACM.

Barnachon, M., Bouakaz, S., Guillou, E., and Boufama, B. (2012). Interprétation de mouvements temps réel. In *RFIA*.

Bashir, F., Khokhar, A., and Schonfeld, D. (2007). Object trajectory-based activity classification and recognition using hidden markov models. *Image Processing, IEEE Transactions on*, 16(7):1912–1919.

Breiman, L. (1992). *Probability*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Campbell, L. and Bobick, A. (1995). Recognition of human body motion using phase space constraints. In *Computer Vision, 1995. Proceedings., Fifth International Conference on*, pages 624–630. IEEE.

Chang, C. and Lin, C. (2011). Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27.

Fengjun, L., Nevatia, R., and Lee, M. W. (2005). 3d human action recognition using spatio-temporal motion templates. ICCV'05.

Girshick, R., Shotton, J., Kohli, P., Criminisi, A., and Fitzgibbon, A. (2011). Efficient regression of general-activity human poses from depth images. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 415–422. IEEE.

He, H. and Ghodsi, A. (2010). Rare class classification by support vector machine. In *Pattern Recognition (ICPR), 2010 20th International Conference on*, pages 548–551. IEEE.

Heikkilä, M., Pietikäinen, M., and Schmid, C. (2009). Description of interest regions with local binary patterns. *Pattern recognition*, 42(3):425–436.

Hsu, C. and Lin, C. (2002). A comparison of methods for multiclass support vector machines. *Neural Networks, IEEE Transactions on*, 13(2):415–425.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Attention, Perception, & Psychophysics*, 14(2):201–211.

Just, A., Marcel, S., and Bernier, O. (2004). Hmm and iohmm for the recognition of mono-and bi-manual 3d hand gestures. In *ICPR workshop on Visual Observation of Deictic Gestures (POINTING04)*.

Laptev, I., Marszalek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Lazebnik, S., Schmid, C., and Ponce, J. (2005). A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1265–1278.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14. IEEE.

Liu, J., Luo, J., and Shah, M. (2009). Recognizing realistic actions from videos 'in the wild'. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1996–2003. IEEE.

Lowe, D. (1999). Object recognition from local scale-invariant features. In *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, volume 2, pages 1150–1157. Ieee.

Matikainen, P., Hebert, M., and Sukthankar, R. (2010). Trajectons: Action recognition through the motion analysis of tracked features. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*.

Messing, R., Pal, C., and Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 104–111. IEEE.

Mezaris, V., Dimou, A., and Kompatsiaris, I. (2010). Local invariant feature tracks for high-level video feature extraction. In *Image Analysis for Multimedia Interactive Services (WIAMIS), 2010 11th International Workshop on*, pages 1–4. IEEE.

Müller, M. and Röder, T. (2006). Motion templates for automatic classification and retrieval of motion capture data. In *Proc. of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 137–146.

Ni, B., Wang, G., and Moulin, P. (2011). Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 1147–1153. IEEE.

Ni, B., Yan, S., and Kassim, A. (2009). Contextualizing histogram. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1682–1689. Ieee.

Parsa, K., Angeles, J., and Misra, A. (2004). Rigid-body pose and twist estimation using an accelerometer array. *Archive of Applied Mechanics*, 74(3):223–236.

Raptis, M., Kirovski, D., and Hoppe, H. (2011). Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 10th Annual ACM SIGGRAPH/Eurographics Symposium on Computer Animation, SCA 2011*, pages 147–156.

Raptis, M. and Soatto, S. (2010). Tracklet Descriptors for Action Modeling and Video Analysis. *Computer Vision–ECCV 2010*, pages 577–590.

Rodriguez, M. D., Ahmed, J., and Shah, M. (2008). Action mach: a spatio-temporal maximum average correlation height filter for action recognition. In *In Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition*.

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 *CVPR*, volume 2, page 7.

Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. Ieee.

Sonnenburg, S., Ratsh, G., Henschel, S., and C., W. (2010). The shogun machine learning toolbox. *The Journal of Machine Learning Research*, 99:1799–1802.

Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., and Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2004–2011. Ieee.

Sung, J., Ponce, C., Selman, B., and Saxena, A. (2011). Human activity detection from rgbd images. In *AAAI workshop on Pattern, Activity and Intent Recognition (PAIR)*.

Tenorth, M., Bandouch, J., and Beetz, M. (2009). The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 1089–1096. IEEE.

Yao, A., Gall, J., Fanelli, G., and Van Gool., L. (2011). Does human action recognition benefit from pose estimation? In *BMVC*.

Yao, A., Gall, J., and Van Gool, L. (2010). A hough transform-based voting framework for action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2061–2068. IEEE.