MULTIMODAL CLASSIFICATION OF DANCE MOVEMENTS USING BODY JOINT TRAJECTORIES AND STEP SOUNDS

Aymeric Masurelle, Slim Essid and Gaël Richard

Institut Mines-Télécom/Télécom ParisTech, CNRS-LTCI, Paris, France

ABSTRACT

We present a multimodal approach to recognize isolated complex human body movements, namely Salsa dance steps. Our system exploits motion features extracted from 3D sub-trajectories of dancers' body-joints (deduced from Kinect depth-map sequences) using principal component analysis (PCA). These sub-trajectories are obtained thanks to a footstep impact detection module (from recordings of piezoelectric sensors installed on the dance floor). Two alternative classifiers are tested with the resulting PCA features, namely Gaussian mixture models and hidden Markov models (HMM). Our experiments on a multimodal Salsa dataset show that our approach is superior to a more traditionnal method. Using HMM classifiers with three hidden states, our system achieves a classification performance of 74% in F-measure when recognizing gestures among six possible classes, which outperforms the reference method by 11 percentage points.

1. INTRODUCTION

During the last few years, significant advances in human-computer interaction technologies have facilitated the multimodal recording of spatio-temporal features of human body with the advent of devices such as the Wiimote, the Kinect, the Leap, etc. In parallel, gesture recognition has been an important research area, mainly in the computer vision community [1]. On the one hand using RGB-cameras or depth sensors, different gesture recognition systems based on shape or appearance features have shown their efficiency [1]. Following the idea that human body movements can be evoked using the motion of bright spots placed at main body joints [2], Bashir et al propose an original approach for gesture recognition. They use a representation formed by a temporal-segmentation of 2D hand trajectories to perform a sign language classification task [3]. On the other hand a few works have followed multimodal approaches. For example, for the analysis of traditional Japanese dance performances, Shiratori et al [4] combine a segmentation deduced from music beat information and a heuristic method for detecting key poses from speed data of a performer's hand, feet and center of mass.
Our work builds upon ideas developed in those previous contributions [3, 4] as we address the recognition of Salsa dance steps. Our system relies on a novel multimodal representation exploiting trajectories of dancers' 3D body joint positions (estimated from depthmap sequences) and an original segmentation technique segmenting those trajectories into sub-trajectories. The sub-trajectories are defined on instants where dancers' footstep impacts happen on the floor (detected from signals of piezoelectric sensors installed on the floor [5]). Then the classification of dance gestures is achieved by hidden Markov models (HMM) classifier using principal component analysis (PCA) to represent efficiently those sub-trajectories [3].
The paper is organised as follows: a presentation of our Salsa dance step recognition process is given in Section 2. Then details and results from the evaluation stage are exposed and discussed in Section 3. Some conclusions are then suggested in Section 4.

2. METHOD

In this section, we first present the data acquisition (e.g. the acquisition of body joint trajectories and temporal locations of footstep impacts on the floor). The different stages of the motion feature extraction are then explained. Finally the model-based classification is briefly described.
This system is a direct adaptation of the model developped by Bashir et al. for sign language recognition [3] to the problem of dance gesture recognition and improves it by exploiting footstep impact detection in order to segment the global motion trajectories. This yields an enhanced motion representation amenable to an effective dynamic classification by HMM.
A global view of this approach has been summed up through a bloc diagram in Figure 1.

2.1. Input data

Our system takes as input trajectories of 3D body joints and step-impact temporal locations. In order to obtain the trajectories of a dancer's body joints, we extract his/her skeleton (15 body joints) from a depth-map sequence using *OpenNI*[1]. For detecting footstep impact instants, we re-use a technique

[1] http://www.openni.org/

proposed by Essid et al [6]. From the signals captured by four onfloor piezoelectric tranducers, onset detection functions are extracted. Then those extracted features feed a one-class support vector machine (SVM). Finally the SVM output values which are below a certain threshold are considered to be indicators of footstep impact instants.

2.2. Motion feature extraction

The motion feature extraction is done through three steps: trajectory segmentation, sub-trajectory representation and PCA-based sub-trajectory representation, which are further detailed below.

Trajectory segmentation

For segmenting 3D body joint trajectories into sub-trajectories, we simply use the footstep impact locations obtained as segmentation points for all considered body joint trajectories. We aim at segmenting these trajectories into elementary ones to be associated to primitive motions from which local features are to be extracted. As explained by Shiratori et al [4], music beat information is an efficient cue for segmenting dance gestures into "primitive motions". However, a dancer, especially if he/she is not an experimented dancer, will not be accurately synchronized in time with the music. Hence a segmentation based on music beat locations cannot be expected to be fully reliable for segmenting the global motion trajectories into primitive motions. In our case, a Salsa dance step is usually contained into a 4-beat music measure where the second, third and fourth beats are marked by footstep impacts. Thus we use footstep impact temporal locations as trajectory segmentation points which are inherent in dancers' gestures instead of using the music beat locations.

Sub-trajectory representation

In order to encourage a scale- and translation-invariant sub-trajectory representation, all considered sub-trajectory coordinates (x,y,z) are centered and normalized as follows:

$$x'_n = \frac{x_n - x_{min}}{x_{max} - x_{min}} \qquad y'_n = \frac{y_n - y_{min}}{y_{max} - y_{min}}$$
$$z'_n = \frac{z_n - z_{min}}{z_{max} - z_{min}} \tag{1}$$

where (x_n,y_n,z_n) and (x'_n,y'_n,z'_n) are respectively the original and normalized versions of a 3D body joint coordinates at the n^{th} instant of time. The maximum and minimum values of those original coordinates are deduced over all sub-trajectories of the current training set.

PCA-based motion feature

Following [3], a PCA-based representation is used to extract motion features from the resulting sub-trajectories. Firstly all those sub-trajectories are re-sampled to the median segment size of all training sub-trajectories. Secondly all considered joint coordinates from each re-sampled sub-trajectories are concatenated to form raw data vectors. Then from these sub-trajectory vectors the motion features are obtained by projecting them on a set of eigen vectors derived from a PCA.

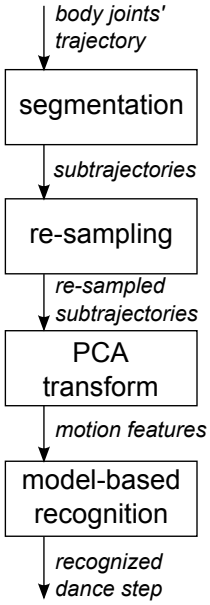

Fig. 1. Schematic illustration of our Salsa dance step recognition approach.

2.3. Model-based classification

To perform the gesture recognition task, two kinds of model-based classifiers are investigated: Gaussian mixture models (GMM) which are commonly used to estimate static probability models and ergodic continuous density HMM which extend GMM to model temporal variations [7].
A model-based classifier is associated to each dance gesture class. Its hyper-parameters are learnt using the EM algorithm. Then test trajectories are categorized into one of the considered gesture classes using maximum likelihood decision.
The implementation of the considered algorithms has been done using a machine learning toolbox called *Scikit-learn* [8].

3. EXPERIMENTAL EVALUATION

In this section, we describe the dataset used, the evaluation protocol and the results obtained compared to the reference system proposed by Bashir et al. [3].

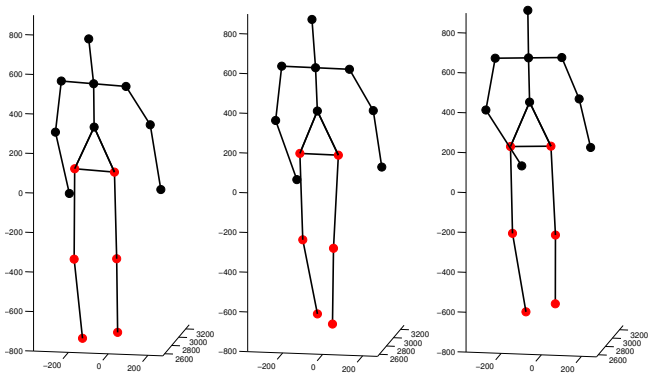

Fig. 2. 15 body joint skeleton representations obtained with *OpenNI* at three different time instants of a choreography performed by one dancer from the *3DLife* dance database [5]. (The joints in red color are the ones we use in our system.)

3.1. Dataset

To evaluate our system, a subset of the *3DLife* dance database [5] is used. This database is composed of multimodal recordings of Salsa choreographies performed by several dancers.In addition to these recordings complementary files including synchronization delays for each media, and dance step annotations related to each music choreography are attached.
For classifier training and evaluation purposes, a ground-truth annotation of dance movements executed at every time instant has been performed, with a temporal precision of one musical beat (the beat information being given along with the annotations provided with the *3DLife* dataset).
In Table 1, the statistics of the considered dataset we use are detailed.

# of dancers	8
# of different steps	6
# of occurrences of the different dance steps:	
backward-basic-step	144
forward-basic-step	123
suzie-q-left	18
suzie-q-right	18
double-cross-left	18
double-cross-right	18
median # of skeleton trajectory samples over all considered steps	31

Table 1. Statistics of the considered dataset to evaluate the performance of the presented algorithms.

The selected Salsa steps for this evaluation are essentially characterized by legs movements. Thus to reduce the problem dimensionality, only six joints are used (ankle left/right, knee left/right and hip left/right), see Figure 2.

More details about the entire *3DLife* dance database can be found in [5].

3.2. Dance step classification

3.2.1. Evaluation procedure

Our evaluation is performed using cross-validation. To avoid using dance steps performed by a particular dancer both in the test and train partition through the same fold, the train and test partitions for each cross-validation fold are formed in a leave-one-dancer-out fashion. Thus the number of folds is equal to the number of dancers, that is eight.
For the GMM, the effect of the number of Gaussian mixtures has been investigated ($M = \{2, 3, 4, 5, 6, 7, 8\}$). In the HMM implementation, we use one Gaussian probability density function per hidden state and different hidden state numbers are tested ($Q = \{2, 3, 4, 5, 6, 7, 8\}$). To deal with the dimension of the feature vector, Bashir et al propose to choose the set of the most significant PCA components representing 95% of the variance-information rate. Using this information rate, we observed that in most cases only 4 or 5 (depending on the segmentation technique employed) of the 84 principal components are kept. We also test the values of the following set for the PCA-feature space dimension, d: $\{10, 20, 40, 60, 80\}$.

3.2.2. Classification results

In Table 2, we present the best results in F-measure obtained over all model parameters and feature space dimensions tested for our system compared to a reference one that consists in a direct adaptation of [3].

	GMM	HMM
Bashir	61% (M=6, d=80)	63% (Q=4, d=40)
Proposed	68% (M=4, d=40)	74% (Q=3, d=60)

Table 2. Best classification results (with 6 body joints) in F-measure on 6 Salsa steps with GMM and HMM classifiers.

The results of Table 2 show that our method obtains superior performance compared to the reference system for both GMM and HMM classifiers. The improvement becomes very clear when using HMM.
When we compare the results obtained with GMM and HMM, an increase in the performance is observed for both methods separately. Actually for the approach proposed by Bashir et al [3], the increase is about 2 percentage points whereas for ours, the improvement is much bigger, about 6 percentage points. As we considered a trajectory as a sequence of sub-trajectories represented by motion features, the need for modeling temporal dependence among those features using HMM is confirmed, especially for our method.

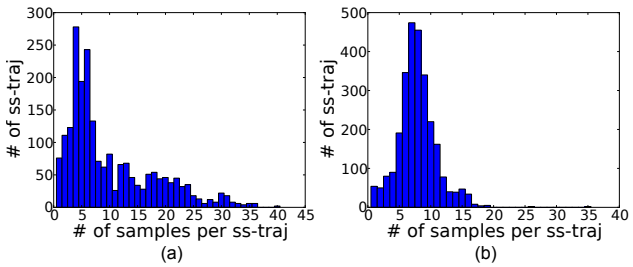

Fig. 3. Sub-trajectory size (in sample number) histograms using Bashir et al's segmentation technique, (a), and ours , (b), for the six different Salsa dance steps.

Comparing the difference between the performance increase for the two considered approaches, our segmentation technique is more appropriate. To study this point, histograms of the sub-trajectory sizes in sample number are given in Figure 3. These histograms show that sub-trajectories obtained with our segmentation method have a length distribution much more compact than using the technique proposed by Bashir et al [3]. This means that our segmentation technique allows for a better consistency of the sub-trajectory sizes.
Thus using our technique to segment trajectories, HMM classifiers manage to take advantage of the temporal dependencies between sub-trajectories which justifies the use of our segmentation approach.

4. CONCLUSION

Through this paper we have presented a new multimodal dance gesture classification system. Our approach takes advantage of an original temporal-segmentation method of 3D body joint trajectories based on footstep impact detections which allows an efficient representation of motion features. The efficiency of our segmentation technique is highlighted when the temporal dependencies between adjacent sub-trajectories are modeled by HMM. Using those latters, our approach outperforms by 6 percentage points systems based on GMM classifiers or/and using a segmentation technique based on sudden changes in trajectory curvatures. An important extension of our research would be to include complementary features and data from other media such as accelerometers and microphones.

5. ACKNOWLEDGEMENT

This research was supported by the European Commission under contract "FP7-287723 REVERIE". Thanks to Angélique Drémeau for having been very helpful during the annotation of the Salsa dance step.

6. REFERENCES

[1] J.K. Aggarwal and M.S. Ryoo, "Human activity analysis: A review," *ACM Comput. Surv.*, vol. 43, no. 3, pp. 16:1–16:43, Apr. 2011.

[2] G. Johansson, "Visual perception of biological motion and a model for its analysis," *Attention Perception Psychophysics*, vol. 14, no. 2, pp. 201–211, 1973.

[3] F.I. Bashir, A.A. Khokhar, and D. Schonfeld, "Object trajectory-based activity classification and recognition using hidden markov models.," *IEEE Transactions on Image Processing*, vol. 16, no. 7, pp. 1912–1919, 2007.

[4] T. Shiratori, A. Nakazawa, and K. Ikeuchi, "Detecting dance motion structure through music analysis," in *Proceedings of the Sixth IEEE international conference on automatic Face and Gesture Recognition*, Washington, DC, USA, 2004, FGR' 04, pp. 857–862, IEEE Computer Society.

[5] S. Essid, X. Lin, M. Gowing, G. Kordelas, A. Aksay, P. Kelly, T. Fillon, Q. Zhang, A. Dielmann, V. Kitanovski, R. Tournemenne, A. Masurelle, E. Izquierdo, N.E. O'Connor, P. Daras, and G. Richard, "A multimodal dance corpus for research into interaction between humans in virtual environments," *Journal on Multimodal User Interfaces*, 2012.

[6] S. Essid, Y. Grenier, M. Maazaoui, G. Richard, and R. Tournemenne, "An audio-driven virtual dance-teaching assistant," in *Proceedings of the 19th ACM international conference on Multimedia*, New York, NY, USA, 2011, MM '11, pp. 675–678, ACM.

[7] C.M. Bishop, *Pattern Recognition and Machine Learning*, Springer, Secaucus, NJ, USA, 2006.

[8] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, Oct 2011.