GESTURE RECOGNITION USING A NMF-BASED REPRESENTATION OF MOTION-TRACES EXTRACTED FROM DEPTH SILHOUETTES

Aymeric Masurelle, Slim Essid and Gaël Richard

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

ABSTRACT

We present a novel approach that classifies full-body human gestures using original spatio-temporal features obtained by applying non-negative matrix factorisation (NMF) to an extended depth silhouette representation. This extended representation, the *motion-trace* representation, incorporates temporal dimensions as it is built by superimposition of consecutive depth silhouettes. From this representation, a dictionary of local motion features is learned using NMF. Thus the projection of these local motion feature components on the incoming motion-traces results in a compact spatio-temporal feature representation. Those new features are then exploited using hidden Markov models for gesture recognition. Our experiments on a gesture dataset show that our approach outperforms more traditional methods that use pose features or decomposition techniques such as principal component analysis.

Index Terms— Gesture recognition, Depth-silhouette, Motion-trace, Non-negative matrix factorisation , Hidden Markov models.

1. INTRODUCTION

Applications dealing with Human-Computer Interaction (HCI) have become omnipresent in our daily lives: video games, tablet computer or smart phone usages are just a few examples of this reality. One important aspect of those interactions relies on the analysis and recognition of human body movements such as hand movements, head gestures, body language, etc. For the last few years, significant advances in HCI technologies have facilitated the recording of spatio-temporal features of human body. In parallel, gesture recognition has been an important research area, mainly in the computer vision community. Numerous surveys on this topic have been written [1, 2]. Usually related works focus on methods using data deduced from visual sensors: RGB-cameras or depth sensors.

Using this kind of data, different types of data representations have been considered for gesture recognition task. Several approaches, very suitable for the action recognition task, use the pixel locations of spatio-temporal interest points to compute local spatio-temporal descriptors [3, 4]. Other systems make efficiently use of the 3D human body joint representation either using it directly [5, 6], or using angles formed by the obtained joints [7, 8]. Also, several works have reported successful gesture recognition using global human shape representations. For example, quantized representations of binary silhouettes have been successfully exploited by Yamato et al. [9] to learn several tennis actions using hidden Markov models (HMM). The concept of binary silhouettes has been extended by Bobick and Davis [10] by incorporating temporal dimensions using two different types of temporal integration over sequences of binary silhouettes. An improvement using this global representation has been made by deducing both local and global descriptors [11]. Although such representations and variants thereof have been exploited to represent a wide variety of body gestures with success, they tend to suffer from ambiguities in the body configurations. For instance, identical binary silhouette representations can be obtained with different arm or hand gestures when they are performed in front of the performer's torso. To solve ambiguities, an extension of the binary silhouette representation has been proposed by Muñoz-Salinas et al. [12]: the *depth silhouette* representation. This representation is obtained by filling the pixels of the silhouette with the corresponding depth value. Using this representation in a gesture recognition task, satisfactory results have been obtained through the combined use of principal component analysis (PCA) and a set of HMM. Through their system, Muñoz-Salinas et al. suppose that a gesture is a temporal juxtaposition of poses, where those poses are expressed as a combination of "eigen" poses.

In our work, we further develop this concept of gesture by considering it as a concatenation of atomic sequences of motions and poses of different body parts. Thus our gesture classification system relies on local spatio-temporal features exploiting the depth silhouette representation. First we extend the depth silhouette representation incorporating local dynamics using a temporal integration process which results in a global motion representation: the *motion-trace*. Then to spatially decompose this global motion representation, a non-negative matrix factorisation (NMF) is carried out. Indeed Lee and Seung [13] have successfully shown the efficiency of NMF in learning the parts of face images in comparison

This research was supported by the European Commission under contract "FP7-287723 REVERIE".

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

to PCA or vector quantization techniques. Thus, in our case, NMF is used to learn a dictionary of redundant local motion patterns in order to express the global motion-trace representation as a composition of local motion patterns. Finally, the classification of gestures is achieved by a HMM classifier.
The paper is organised as follows: a presentation of our human body gesture 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

The general architecture of our approach is summed up in Figure 1. The input of our system is the depth-map sequences captured by a monocular depth sensor placed in front of the performer. First the depth silhouette representation is obtained using a background subtraction technique followed by a cropping and resizing procedure. A temporal integration reveals local dynamics through the motion-trace representation. Then this global representation is decomposed in a dictionary of local spatio-temporal patterns using non-negative matrix factorization. Thus to perform the recognition task, the sequence of dictionary component activations representing a specific motion-trace sequence feeds HMM classifiers.

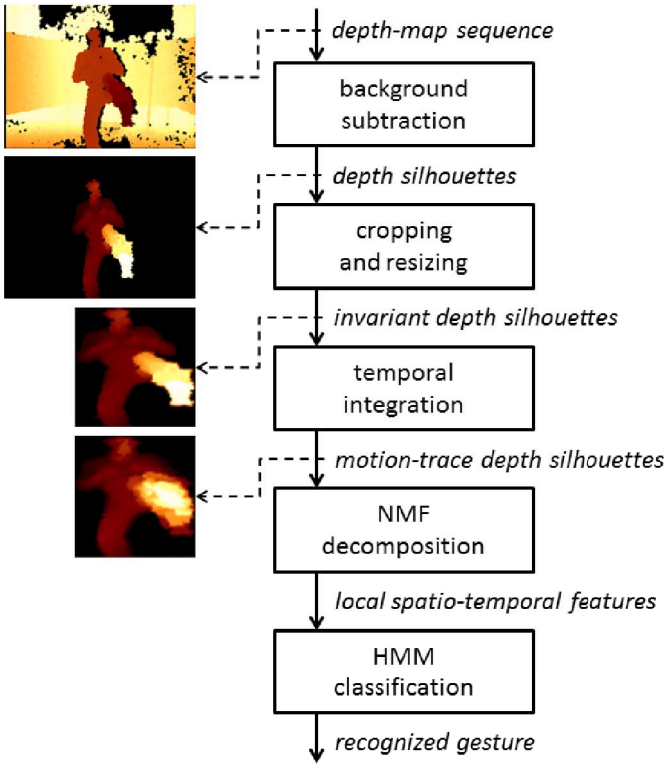

Fig. 1. Schematic illustration of our gesture recognition approach.

2.1. Depth silhouette extraction

2.1.1. Background subtraction

As we are only focusing on human full-body gestures, a background segmentation process is first used to separate the performer from his/her background. To process this background segmentation on a depth-map stream, a constant depth threshold, τ, is applied over all the frames of the sequence:

$$I'[j,k,t] = \begin{cases} I[j,k,t] & \text{if} \quad I[j,k,t] \leq \tau \\ 1 & \text{if} \quad I[j,k,t] > \tau \end{cases} \qquad (1)$$

where $I'[j,k,t]$ is the resulting value placed at the pixel index (j,k) of the depth-map frame at time index t and $I[j,k,t]$ is the corresponding original depth-map frame.

2.1.2. Silhouette bounding-box cropping

For encouraging translation and scale invariance, the following three steps are accomplished. First, all frames of the resulting depth-map sequence are cropped with respect to the bounding-box within which the performer lies. This allows horizontal and vertical translation invariances. Then, for the depth translation invariance, the depth values belonging to the performer's pixels are normalized between 0 (the furthest point) and 1 (the closest point). And thus the scale-invariance is obtained by resizing the cropped frames to a fixed frame resolution, in our case 64x64.

Now that we have extracted translation- and scale-invariant depth silhouettes, we have an efficient spatial feature of the performer's body presence for each depth frame.

2.2. Feature representation

In this section, we further describe our feature representation which aims at representing a gesture as a juxtaposition of local spatio-temporal pattern combinations.

2.2.1. Motion-trace representation

Inspired by the temporal templates introduced by Bobick and Davis [10], we have extended the concept of depth silhouette to incorporate motion dynamics. A temporal integration is thus applied on consecutive depth silhouettes, S^d, to obtain a motion-trace, S^{mt}. This temporal integration is done using a sliding window whose length is set to a fixed number of frames, n_τ, with an overlap of half the window size.

$$S^{mt}[j,k,t] = \sum_{i=0}^{n_\tau - 1} \frac{S^{mt}[j,k,t-i]}{n_\tau} \qquad (2)$$

This integration improves the robustness of the feature

and achieves a reduction of the image rate for the rest of the process.

2.2.2. NMF motion-trace depth silhouettes representation

The motion-trace representation can be seen as a global representation of a short time instant of a full body gesture. With respect to our gesture consideration, we need to decompose motion-trace representation into local spatio-temporal patterns. NMF is known to be appropriate for representing an image as a linear combination of images parts which may be previously learned [13]. In our work a NMF is carried out on motion-traces in order to represent this global representation as a combination of local spatio-temporal features.
Let $\mathbf{X}$ be a non-negative matrix, the problem of NMF is to find two matrices $\mathbf{W}$ and $\mathbf{H}$ such that:

$$\begin{cases} \mathbf{X} \simeq \mathbf{WH} \\ \mathbf{W} \geq 0, \mathbf{H} \geq 0 \end{cases} \tag{3}$$

where $\mathbf{W}$ is a dictionary of local components and $\mathbf{H}$ represents the activations of the dictionary components. The above factorization is achieved by solving the following problem:

$$\min_{\mathbf{W},\mathbf{H}\geq 0} D(\mathbf{X}|\mathbf{WH}) \tag{4}$$

where $D(\mathbf{X}|\mathbf{WH})$ is a cost function which can be expressed as:

$$D(\mathbf{X}|\mathbf{WH}) = \sum_{n=1}^{N}\sum_{f=1}^{F} d([\mathbf{X}]_{fn}|[\mathbf{WH}]_{fn}) \tag{5}$$

where $d(x|y)$ is the chosen scalar divergence function.

In our work, the observation matrix $\mathbf{X}$ is formed by a concatenation of vectorized version of motion-traces. $\mathbf{W}$ is a dictionary of local spatio-temporal features and $\mathbf{H}$ represents the activations of each local motion feature for each respective motion-traces. $\mathbf{X}$, $\mathbf{W}$ and $\mathbf{H}$ have respectively the dimensions $F \times N$, $F \times K$ and $K \times N$ with F the dimension of the vectorized motion-trace representation, K the number of dictionary components and N the number of motion-trace examples.
The divergence used in this work is the common squared error (eq.6).

$$d_{SE}(x|y) = \frac{1}{2}(x-y)^2 \tag{6}$$

We choose the use of the multiplicative update rules to solve the minimization problem (4) under the latter two objective functions [14].
To achieve the proposed NMF representation process, we have to perform two consecutive steps: first the formation of the local spatio-temporal feature dictionary through the learning step and then the representation one.

- learning step : in this step, we iteratively learn $\mathbf{W}$ and $\mathbf{H}$ (randomly initialized) using the entire training set concatenated in a single observation matrix $\mathbf{X}$. This results in a dictionary of K local spatio-temporal features that is going to be used to decompose the incoming motion-traces.
- representation step : once the dictionary of local motion features is learned, each incoming motion-trace is projected on these dictionary components to obtain the corresponding vector of their decomposition factors. Thus those activation vectors are used as motion features.

As a result of this NMF decomposition process, a compact and semantically meaningful motion representation is obtained.

2.3. HMM-based classification

Ergodic continuous density hidden Markov models are used to perform the gesture recognition task [15].
A HMM classifier is associated to each gesture class. Its hyper-parameters are learned using the Baum-Welch algorithm. Then test data 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* [16].

3. EXPERIMENTAL EVALUATION

In this section, we describe the dataset used, the evaluation protocol and the results obtained.

3.1. Dataset

To evaluate our system, a subset of the *Huawei/3DLife 3D human reconstruction and action recognition Grand Challenge* database[1] is used: a set of the 8 dynamic actions ('Golf drive', 'Lunges', 'Squats', 'Jumping Jacks', 'Tennis backhand','Walking on the treadmill', 'Punching and kicking', 'Hand Waving') performed by 8 participants (S01 $\rightarrow$ $S08$) from the session 1 of the dataset 1. This database is composed of multimodal recordings (RGB and depth-map video streams, audio streams and inertial data streams) of multiple gestures performed several times (at least 5 times by each individual) by each participant. In addition, gesture annotations related to each performance are attached.

3.2. Reference systems

We have reimplemented three existing gesture recognition systems as faithfully as possible in order to compare their performance with our proposed approach. We have chosen these systems in the context that is relevant to our work, that is gesture recognition for HCI applications.

[1] http://mmv.eecs.qmul.ac.uk/mmgc2013/

One of them is the approach introduced in [12] which directly uses the depth-map stream to extract the corresponding depth silhouette sequence and then uses a principal component analysis to acquire the features that will feed the HMM classifiers. In addition, we add a temporal integration step (c.f. Section 2.2.1) in between the extraction of depth silhouettes and the PCA process to compress the data and to make it more robust. The two others references approaches are described in [8, 6]. They both use the 3D positions of the participants' main body joints as input. Those position are extracted from a depth-map sequence using the OpenNi[2]. To feed the HMM classifiers, the approach of Papadopoulos et al.[8] uses spherical angles between selected body joints and their respective angular velocities expressed in a torso-centered coordinate system. Also, our previous method [6] exploits motion features extracted from 3D sub-trajectories of participants' body joint positions using PCA. The original implementation of this approach segments the body joint trajectories where participants' footstep impacts happen on the floor using signals of onfloor piezoelectric sensors. In this study, no piezoelectric sensors are available, then we have segmented the trajectories using a fixed size window.

3.3. Evaluation procedure

Our evaluation is performed using cross-validation. To avoid using gestures performed by a particular participant 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-participant-out fashion. Thus the number of folds is equal to the number of participants, that is eight.
Different window size are tested ($w = \{3, 4, 5, 6, 7\}$). For the NMF, the effect of the number of components has been investigated ($k = \{64, 128, 256\}$). We also test the values of the following set for the PCA-feature space dimension: $d = \{20, 30, 40, 50, 60\}$. 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 quantify action classification results we use the F-measure[3]. This quantity is first computed for each gesture in each fold, then the obtained F-measure are uniformly averaged over all gestures giving a F-measure value for each fold. Finally a global F-measure is obtained by averaging uniformly over all folds.

3.4. Classification results

In Table 1, we present the best results in F-measure obtained over all model parameters and feature space dimensions tested for our system and the reference systems that consist in an adaptation of [8, 6, 12] as described in Section 3.2.

[2] `http://www.openni.org/`

[3] F-measure = 2·(precision × recall) / (precision + recall)

Approaches			
[8]	[6]	[12]	proposed
78%	89%	89%	**91%**
(Q=6)	(w=6, d=30, Q=2)	(w=5, d=40, Q=2)	(w=6, k=128, Q=7)

Table 1. Best classification results in F-measure on 8 gesture classes with HMM classifiers.

The results of Table 1 show that our approach obtains superior performance compared to the reference systems. First we can observe that the method based only on global pose features [8] is outperformed by the other methods that incorporate motion dynamics and a decomposition process through their feature representation. However, our experiments have shown that the number of hidden states had only minor impact on the results whereas the window size and the number of NMF components are two critical parameters that have a strong influence on the results. As the performance of our system is superior to the other reference systems, the use of a NMF decomposition technique seems to be more suitable to efficiently represent spatio-temporal motion features compared to a PCA decomposition. This highlights the importance of considering a gesture as a concatenation of atomic sequences of motions and poses of different body parts.

4. CONCLUSION

Through this paper we have presented a new gesture classification system based on a compact and efficient NMF representation using local spatio-temporal features. First by incorporating motion dynamics to depth silhouette representation, then by representing it through a NMF decomposition using a dictionary of local spatio-temporal features, we obtain an efficient motion representation. Finally temporal dependencies between those motion features are modelled by HMM. Our system obtains better performance than reference systems based on PCA decomposition and/or 3D body joint positions. An important extension of our research would be to encourage complementary invariances and to include data from other media such as RGB cameras and accelerometers.

5. REFERENCES

[1] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," *IEEE Transactions on Systems, Man and Cybernetics, Part C (Applications and Reviews)*, vol. 37, no. 3, pp. 311–324, May 2007.

[2] 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.

[3] I. Laptev, "On space-time interest points," *Int. J. Comput. Vision*, vol. 64, no. 2-3, pp. 107–123, Sept. 2005.

[4] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action Recognition by Dense Trajectories," in *Proceedings of the IEEE Conference on Computer Vision & Pattern Recognition*.

[5] Zhe Wu, Xiong Li, Xu Zhao, and Yuncai Liu, "Hybrid generative-discriminative recognition of human action in 3d joint space," in *Proceedings of the 20th ACM international conference on Multimedia*, New York, NY, USA, 2012, MM '12, pp. 1081–1084, ACM.

[6] A. Masurelle, S. Essid, and G. Richard, "Multimodal classification of dance movements using body joint trajectories and step sounds," in *International Workshop on Image and Audio Analysis for Multimedia Interactive Services WIAMIS*, 2013.

[7] L. W. Campbell and A. F. Bobick, "Recognition of human body motion using phase space constraints," in *Proceedings of the Fifth International Conference on Computer Vision*, Washington, DC, USA, 1995, ICCV '95, pp. 624–630, IEEE Computer Society.

[8] G.Th. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using kinect data," in *Proceedings of the 20th International Conference on MultiMedia Modeling*, accepted for publication, MMM 2014.

[9] J. Yamato, J. Ohya, and K. Ishii, "Recognizing human action in time-sequential images using hidden Markov model," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1992, pp. 379–385.

[10] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 3, pp. 257–267, Mar. 2001.

[11] D. Wu and L. Shao, "Silhouette analysis based action recognition via exploiting human poses," *IEEE Transactions on Circuits and Systems for Video Technology*, 2013.

[12] R. Muñoz Salinas, R. Medina-Carnicer, F.J. Madrid-Cuevas, and A. Carmona-Poyato, "Depth silhouettes for gesture recognition," *Pattern Recogn. Lett.*, vol. 29, no. 3, pp. 319–329, Feb. 2008.

[13] D.D. Lee and H.S. Seung, "Learning the parts of objects by nonnegative matrix factorization," *Nature*, vol. 401, pp. 788–791, 1999.

[14] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*. 2001, pp. 556–562, MIT Press.

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

[16] 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.