Adaptation Procedure for HMM-Based Sensor-Dependent Gesture Recognition

Sohaib Laraba*, Joëlle Tilmanne†, Thierry Dutoit‡
TCTS Lab, Numediart Institute, University of Mons, Belgium

Abstract

In this paper, we address the problem of sensor-dependent gesture recognition thanks to adaptation procedure. Capturing human movements by a motion capture (MoCap) system provides very accurate data. Unfortunately, such systems are very expensive, unlike recent depth sensors, like Microsoft Kinect, which are much cheaper, but provide lower data quality. Hidden Markov Models (HMMs) are widely used in gesture recognition to learn the dynamics of each gesture class. However, models trained on one type of data can only be used on data of the same type. For this reason, we propose to adapt HMMs trained on Mocap data to a small set of Kinect data using Maximum Likelihood Linear Regression (MLLR) to recognize gestures captured by a Kinect. Results show that using this method, we can achieve a recognition average accuracy of 84.48% using a small set of adaptation data while, using the same set to create new models, we obtain only 72.41% of accuracy.

CR Categories: I.3.3 [Computer Graphics]: Three-Dimensional Graphics and Realism—Display Algorithms

Keywords: Motion Capture, Gesture Recognition, HMMs, MLLR Adaptation

1 Introduction

Traditional dances are one of the Intangible Cultural Heritage (ICH) forms that are at risk of being forgotten with time, and thus of being lost if they are not safeguarded and transmitted to next generations. The advance of motion capture technology can play an important role for the safeguarding of this form of ICH allowing the capture and analysis of expert movements. The dance sequence recorded can eventually be used in serious games to allow a student to learn the dance by imitation and comparing his performance to the expert. Gesture recognition strongly depends on data quality and on the features extracted from the data. Sophisticated Motion Capture (Mocap) systems like Qualisys[1] and Vicon[2] provide very accurate 3D skeletal data. However, these systems are marker-based and very expensive. The recent introduction of low-cost depth sensors, particularly the Microsoft Kinect [3] allow tracking of 3D joints positions in real time at a lower price, but this comes at the expense of data quality.

The present work is inscribed within the framework of the i-Treasures European project [4] which addresses the use of new technologies for the preservation of ICH using Information and Communication Technology (ICT), and more specifically of the dance use case. One of the problems when considering gesture data capture and analysis is the availability of the expert who, generally, is only available for limited time, where at the same time, accurate data is important to better analyze the gestures and can be obtained using high precision motion capture systems. On the other hand, the user that will try to learn these dance steps by imitating the expert will be captured using less accurate motion capture system because of their lower price. In this case, data configuration and quality is different. In addition, using different sensors at the same time for capturing can create interference and add noise to captured data. Our aim is to take benefit of high precision motion capture system, when available, to capture expert gestures with high precision, then with a small amount of expert data recorded using a low-cost sensor (it can be few gestures, recorded at the end of the capturing session), we can create a system able to recognize gestures recorded by this low-cost sensor. We extract first meaningful features from the skeletal data instead of using only 3D locations or angles of the joints in order to improve the accuracy of the gesture recognizer, and then we use the features, extracted from the precise data to train HMMs. Finally, we adapt these models using a small amount of data captured by a different sensor using Maximum Likelihood Linear Regression (MLLR) technique. To test our system, a database has been recorded using both Qualisys Mocap system and a Kinect V2 to record an expert of traditional dance from the Walloon region (Belgium). The rest of the paper is organized as follows: Section 2 briefly reviews existing literature. In Section 3 we present the recorded database. Section 4 describes our feature vector extraction and Section 5 presents some details about HMM modeling and MLLR adaptation techniques used. Experimental evaluation of our system is presented in Section 6 followed by conclusions and future works in Section 7.

2 Related works

To our knowledge, model adaptation for sensor-dependent gesture recognition is not found in literature. For this reason, we will hence present here in brief, some existing works about skeleton-based gesture recognition in general.

In skeleton-based gesture recognition, three main issues are to be addressed. Data capture, extraction and representation of meaningful features and modeling and learning of different gesture classes. Generally, having precise 3D joint positions is important for gesture recognition. Sophisticated motion capture (Mocap) systems like Qualisys provide very accurate 3D joint positions. Such systems are marker-based and very expensive. Recently, cost-effective motion capture systems and particularly the Kinect have been used for motion capturing and produced reasonable results despite the noisy data provided.

For feature extraction, 3D single joints or combinations of joints have been used in [Hussein et al. 2013] and [Lv and Nevatia 2006] for training, while in [Raptis et al. 2011], authors have used a joint angles representation of the skeleton to recognize dance gestures. In [Pazhoumand-Dar et al. 2015], in addition to sequences of joint

*e-mail: sohaib.laraba@umons.ac.be
†e-mail: joelle.tilmanne@umons.ac.be
‡e-mail: thierry.dutoit@umons.ac.be

MIG '15, November 16 – 18, 2015, Paris, France.
© 2015 ACM. ISBN 978-1-4503-3991-9/15/11...$15.00
DOI: http://dx.doi.org/10.1145/2822013.2822032

[1] Qualisys Motion Capture System: http://www.qualisys.com/
[2] Vicon Motion Capture System: http://www.vicon.com/
[3] Kinect for Windows - Microsoft: https://www.microsoft.com/en-us/kinectforwindows/
[4] The European Project i-Treasures: http://i-treasures.eu/

angles, relative positions of joint pairs have been used for action recognition. In [Müller et al. 2009], the authors have addressed the problem of action recognition using motion templates. Patterns from a sequence of animation have been extracted to recognize actions. [Vemulapalli et al. 2014] have used instead, a body part-based skeletal representation for action recognition. Relationships between different body parts are explicitly modeled using rotations and translations in 3D space.
In the training stage, many algorithms have been used. Trajectories of 3D locations were modeled in [Hussein et al. 2013] using temporal hierarchy of covariance descriptors. Dynamic Time Warping (DTW) has been used for sequence alignment and a threshold approach for classification of Macedonian folk dance gestures in [Pohl and Hadjakos 2010]. DTW has also been used in [Vemulapalli et al. 2014], combined with Fourier Temporal Pyramid representation and linear SVM. [Pazhoumand-Dar et al. 2015] classified extracted features using a similarity function based on the non-metric, Longest Common Subsequences (LCSS) algorithm.
Up to now, gesture recognition methods are designed only for MoCap data alone or low cost sensor data alone and models from one system's data cannot be used for another system because of the difference of data quality, and also, representation of this data.
We find in literature a similar problem addressed in speech recognition. Models being trained on a single speaker's provide poor recognition accuracy for any other user. To address this problem, some transformation-based adaptation techniques have been used to help reducing acoustic mismatch between training and testing conditions. A method of speaker-dependent adaptation for continuous density Hidden Markov Models (HMMs) named Maximum Likelihood Linear Regression (MLLR) has been used in [Leggetter and Woodland 1995a] and [Leggetter and Woodland 1995b] to improve the modeling of a new speaker by updating the parameters of well trained HMM. [Acero et al. 2000] used a method based on truncated Vector Taylor Series to estimate the parameters of HMM matched to a noisy environment given a HMM trained with a clean data. This can be projected on gesture recognition problem where in some cases, we have data provided by a high precision Mocap system for training, but for recognition in the final application, a different sensor will be used. This is also the case when interesting databases are available, recorded using a sensor that is not available for tests.

3 Database

To investigate the effect of the quality of the recorded data on the process of recognition and the possibility of adaptation, a database have been recorded under the framework of the i-Treasures project. In this database, we have captured the gestures of an expert of a traditional dance from the Walloon region of Belgium. This dance is a bit complicated because it contains different styles that are not easy to distinguish. The data was recorded using Qualisys Mocap system including eleven high-speed infrared cameras, capturing 68 markers placed on the body (Figure 3.b) at a frame rate of 177 fps and the second version of Microsoft Kinect, which provides a skeleton of 25 joints (Figure 3.c) with a frame rate of 30 fps. The database contains also a video, synchronized with Qualisys and Kinect data, which will be used as a reference. Different types of steps have been recorded. In the present study, we have focused on four basic steps: Maclotte Base (MB), repeated 42 time, Passe-Pied Base (PB), repeated 30 times, Passe-Pied Fleuret (PF) repeated 20 times and the Backward step which accompanies all other steps, repeated 92 times in the database. Each gesture has a duration of 1.3 to 1.7 seconds. The database contains also non-expert gestures: 22 occurrences of Maclotte Base and 30 occurrences of Passe-Pied Base. The database has been manually annotated using MotionMachine [Tillmanne and d'Alessandro 2015], a framework that allows fast prototyping of motion features based on skeletal data.

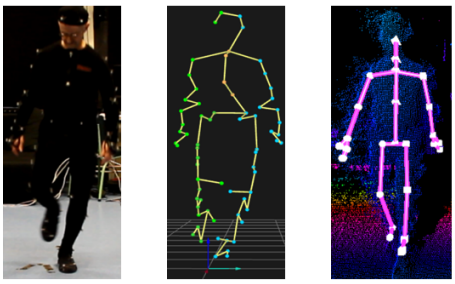

Figure 1: *a) The expert performing the gesture. b) The skeleton reconstructed from Qualisys Data. c) The Skeleton reconstructed from the Kinect data*

4 Feature Vector Extraction

In this section, we investigate different features that are meaningful in the case of Walloon dance.
The general scheme proposed in this paper is illustrated in Figure 2. Instead of using only 3D positions of joints to train our models, other features have been extracted with the help of MotionMachine.

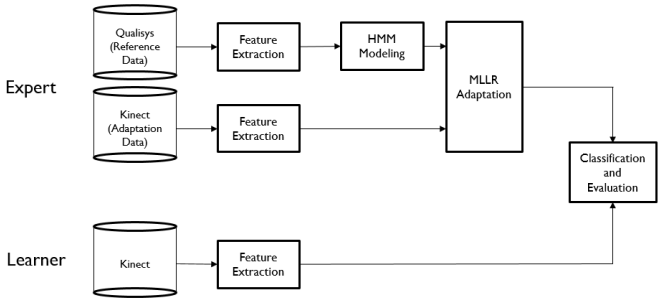

Figure 2: *general scheme of the proposed method*

4.1 Skeletal Representation

Before proceeding to feature extraction, a preprocessing step is needed. A skeleton of twenty joints is formed from different markers positions (Figure 3, left). In Walloon dance steps, arms movements are not important so we decided to exclude their corresponding joints and we keep only twelve joints of head, torso and legs (In red color in Figure 3, right).

Once 3D positions of joints are captured, a skeleton normalization is performed where we take one of the skeletons in a sequence as a reference and normalize all other skeletons in that sequence such as part lengths are equal to the corresponding lengths of the reference skeleton (Figure 3). This normalization makes the skeleton scale invariant.

To make the skeletal data invariant to absolute locations of the person in the scene, we center the skeleton using the pelvis joint as the origin (Figure 4).

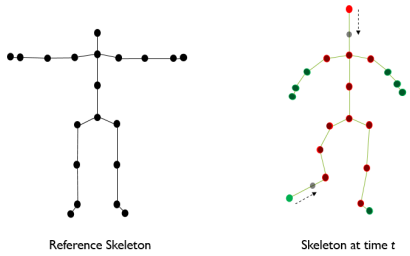

Figure 3: *Normalization of a Skeleton at time t relative to reference skeleton*

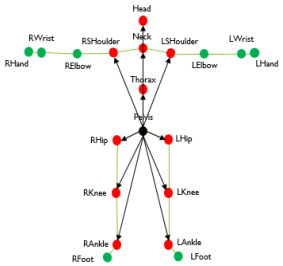

Figure 4: *Centering skeleton to joint Pelvis*

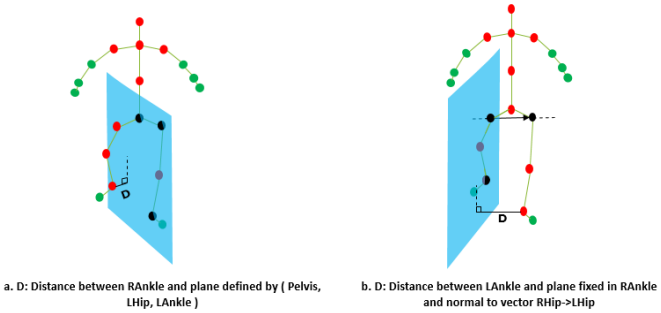

Figure 5: *Relational Features RF1 and RF2*

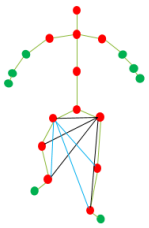

Figure 6: *Relative Motion between joints*

4.2 Feature Extraction

In addition to centered 3D coordinates of the selected joints, two sets of features are extracted. The first set is named relational features. It is a subset of features proposed by Meinard Müller [Müller 2007]. Relational features represent geometric relations between different body joints. The extracted relational features are:

- $RF1$ distance between the right ankle joint and the plane defined by the pelvis, left hip and left ankle joints (Figure 5.a)
- $RF2$ distance between the left ankle joint and the plane fixed in the right ankle and normal to the vector (right hip, left hip) (Figure 5.b)
- $RF3$ angle between the vectors (Right Knee, Right Hip) and (Right Knee, Right Ankle).
- $RF4$ angle between the vectors (Left knee, Left Hip) and (Left Knee, Left Ankle).
- $RF5$ angle between the vectors (Neck, Pelvis) and (Right Hip, Right Knee). (Angle between right leg and body spine)
- $RF6$ angle between the vectors (Neck, Pelvis) and (Left Hip, Left Knee). (Angle between Left leg and body spine)
- $RF7$ angle between the vector (Neck, Pelvis) and the vector perpendicular to ground (verticalness of body spine)

The second set of features is named Relative Motion (RM) between joints [Pazhoumand-Dar et al. 2015]. If we consider two joints (j_x) and (j_y), it represents the sequence of distances between each other (Figure 6) during an action $(t = 1 : e)$.

$$RM(j_x, j_y) = \{eq(f^t_{j_x}, f^t_{j_y})\}^{t=e}_{t=1} \quad (1)$$

Where eq is the Euclidean distance, $f^t_{j_x}$ and $f^t_{j_y}$ are the positions of (j_x) and (j_y) at time t.

Only Relative Motion between lower body joints have been used in our case. The size of the final feature vector is 76 (30 centered 3D coordinates, 7 relational features and 39 relative motion features), which is the input to the learning algorithm.

Figure 7 illustrates the relative features $RF1$ and $RF2$ during two instances of a Maclotte-Base step using Qualisys data (green and red) and Kinect data (blue and black). One can observe the clear presence of noise in the Kinect data features.

5 HMM Modeling and Adaptation

Once features are extracted, we use them to train one HMM for each gesture class. Our approach for HMM Model training and adaptation is inspired by a procedure already developed for speech recognition and based on functions implemented for speech in the Hidden Markov Model Toolkit (HTK), publicly available on HTK website [HTK].

In our case, we consider the situation of having a large database from a high precision Mocap system (Qualisys) and a small set of Kinect data and where, for recognition, only Kinect data will be available.

The first stage of the procedure is to train models using features extracted from precise data captured by a Qualisys system, and in the second stage, these models will be adapted using only a small set of adaptation data provided by the Kinect.

5.1 Hidden Markov Model (HMM)

Hidden Markov Models (HMMs) are forms of statistical Markov Models. A HMM λ of N states and M observations is defined by three parameters. A transition matrix $A = a_{ij}$ where $\{a_{ij}\}$ is the probability of transition from state q_i to state q_j, Output probability Matrix $B = \{b_j(O)\}$ where $b_j(k)$ is the probability of q_j generating the observation o_k and an initial state distribution vector $\Pi = \{\pi_i\}$.

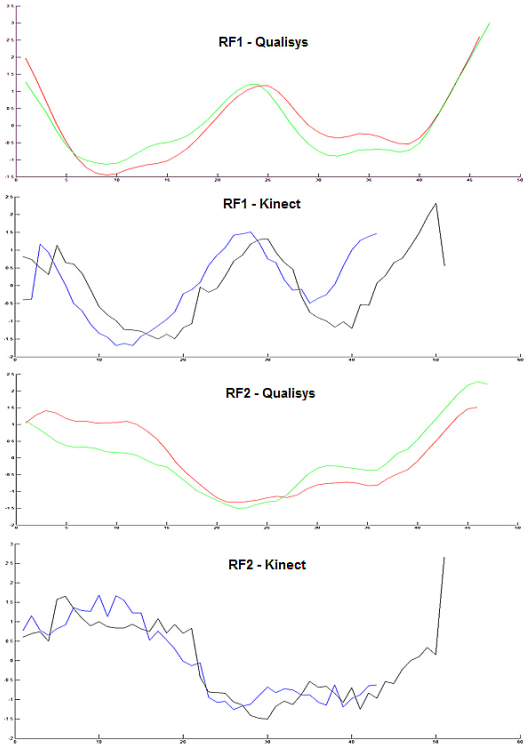

Figure 7: *RF1 and RF2 during two instances of a Maclotte Base step*

The Figure 8 shows a left-right HMM structure with no skips where the only possible state transitions at each time are either to stay in the same state or to go to the next state. A similar structure is used in our work.

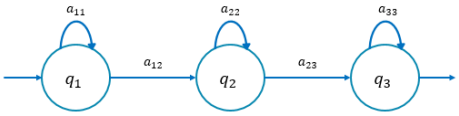

Figure 8: *A simple three-states left-right HMM with no skip*

To perform recognition using HMMs, two parts are needed: training a model and computing the probability that an observation sequence O is generated by the model λ. The objective of the training step is to optimize the parameters (A, B, π) of a HMM λ and can be achieved using the standard Baum-Welch algorithm.

More information about Hidden Markov Models can be found in [Rabiner 1989].

5.2 MLLR Adaptation for Sensor-Dependent gesture recognition

The approach we follow in order to be able to use models initially trained on sufficient amount of data from a given Mocap system to recognize gestures recorded with another Mocap system consists on training sensor-dependent HMMs thanks to an adaptation procedure. This approach is inspired by speech recognition where a speaker-independent system is adapted to improve the modeling of a new speaker by updating the HMM parameters using Maximum Likelihood Linear Regression (MLLR) algorithm. Statistics from the available adaptation data is used to compute a linear regression based transformation for the mean vectors. The transformation matrices are computed to maximize the likelihood of the adaptation data. In other words, MLLR estimates a set of linear transformations W for the means component of a HMM so that these transformations shift the means component in the initial system in a way that each state in the HMM system is more likely to generate the adaptation data. More details about MLLR approach can be found in [Leggetter and Woodland 1995a]. The functions necessary for MLLR adaptation are implemented in the HTK Toolkit.

6 Experimental results

In this section, we present and evaluate our sensor-dependent gesture recognition system. The idea is to create a HMM for each of the gestures that we want to recognize using Qualisys data, to adapt these models using the available Kinect data and finally to recognize gestures captured by the Kinect.
The number of HMM states was empirically determined by using a given set of features to train models using Qualisys and then Kinect data, then perform recognition using data from the same sensor, changing the number of states in each case from 3 to 15. We found that models with 11 states give the best recognition rate.
In subsection 6.1, we test the recognition on the selected features and compare them to the use of normalized 3D positions alone and in subsection 6.2, we present the use of model adaptation when only little amount of training data is available to create new models while there is enough training data from a different sensor.
In our case, we train a HMM for each of the four captured gestures. Each gesture has a duration of 1300 to 1700 ms.

6.1 Evaluation of the selected features

To evaluate our features, we compare the results of classification accuracy of each feature set (Normalized 3D positions - P, Relational Feature – RF, that represents geometric relations between different body joints and Relative Motion – RM, that represents the distances between different joints during an action) using the Qualisys and the Kinect datasets.

For each dataset, we train HMMs on 70% of data use the remaining data as test set. Accuracy results are shown in Figure 9. We notice first that classification using Qualisys data is better in all cases, which shows the importance of data quality for recognition. Also, the combination of the presented feature sets shows better results than using Centered 3D positions alone.

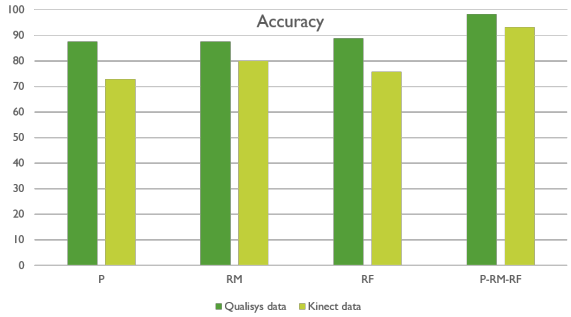

Figure 9: *Comparison of accuracy between different feature sets*

The Passe-Pied Base (PB) and the Passe-Pied Fleuret (PF) steps are very similar since they are variations of the same step, and analyzing the confusion matrices (Figure 10) of the last case (combination

of all features), we see that the classifier fails sometimes to distinguish between them. If we consider both steps as one, the classifier succeeds to get 100% of accuracy using Qualisys data and 98.28% using Kinect data.

At the end, we decided to use for rest of experiments, the combination of these features instead of using normalized 3D positions alone.

	ar	mb	pb	pf	Del
ar	28	0	0	0	0
mb	0	14	0	0	0
pb	0	0	8	0	0
pf	0	0	1	7	0
Ins	0	0	0	0	

a.

	ar	mb	pb	pf	Del
ar	28	0	0	0	0
mb	1	13	0	0	0
pb	0	0	6	2	0
pf	0	0	1	7	0
Ins	0	0	0	0	

b.

Figure 10: *Confusion matrices. a) using Qualisys data. b) using Kinect data.*

6.2 Adaptation effect on gesture classification

In this section, we study the MLLR adaptation of models initially trained on a sufficient amount of Qualisys data to recognize gestures using a Kinect, where only a small amount of Kinect data is available, but not enough to create new models. To verify if adaptation is more is better that training new models from a small amount of data, we have trained HMM Models using reduced amount of Kinect data (4 repetitions of Maclotte Base, 4 repetitions of Passe-Pied Base, 4 repetitions of Passe-Pied Fleuret and 12 repetitions of the Backward step), then we used the same data to adapt models trained previously using the Qualisys data. We tested recognition of gestures recorded by a Kinect using both models, before and after adaptation Figure 11 shows first, that recognition using trained models from Qualisys data with no adaptation has an average accuracy of 75.86% and is even better than using Kinect data alone for training (72.41%). Second, when we adapt Qualisys data models using a small set of Kinect data, the average accuracy has improved to 84.48%.

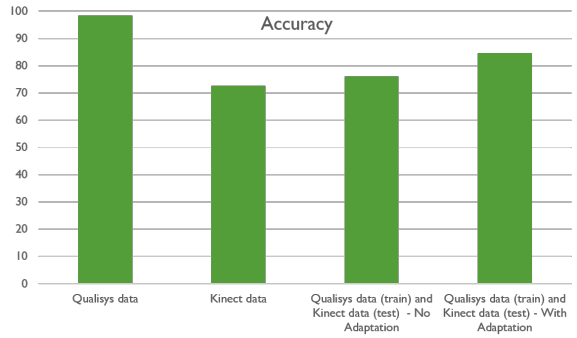

Figure 11: *Comparison of recognition accuracy before and after adaptation*

Figure 12 shows the confusion matrices for the case of using a small set of the Kinect data for training and using the same set for adaptation. For the first case, there were many confusions between MB and AR and between PB and PF. 7 MB steps out of 14 were recognized as AR steps and 5 PB steps out of 8 were recognized as PF steps. However, we have less confusions when we use the adapted models, 4 MB steps are still considered as AR, and only one PB is recognized as PF.

	ar	mb	pb	pf	Del
ar	28	0	0	0	0
mb	7	4	0	1	2
pb	0	0	3	5	0
pf	0	0	1	7	0
Ins	0	0	0	0	

a.

	ar	mb	pb	pf	Del
ar	27	0	1	0	0
mb	4	9	0	0	1
pb	0	0	7	1	0
pf	0	0	2	6	0
Ins	0	0	0	0	

b.

Figure 12: *Confusion matrices. a) training and training using Kinect data. b) use of adaptation technique*

7 Conclusion

In this paper, we have presented a set of features to be used in step recognition in dance from the Walloon region and an adaptation method of HMM models for sensor-dependent gesture recognition system using a small set of adaptation data. We have first represented a human skeleton by different features extracted from 3D locations of the joints. Using this representation we have modeled expert gestures, captured by a high precision Mocap system, using Hidden Markov Models and then adapted these models to a small set of data captured using a Kinect sensor, using Maximum Likelihood Linear Regression (MLLR) technique. We showed that when a small set of adaptation data is available, adaptation results are better than creating new models from this set of data.
As perspectives, we plan to test our method on different databases, we plan also to extract more meaningful features and automatically select them for each use case. We are also exploring different adaptation methods to improve our models, and automatic sensor-to-sensor adaptation applied to new gestures.

Acknowledgements

This work has been supported by the European Union (FP7-IC7-2011-9) under grant agreement n 600676 (i-Treasures project).

References

ACERO, A., DENG, L., KRISTJANSSON, T. T., AND ZHANG, J. 2000. Hmm adaptation using vector taylor series for noisy speech recognition. In *INTERSPEECH*, 869–872.

HTK, W. G. Hidden Markov Model Toolkit (HTK). `http://htk.eng.cam.ac.uk/`. [Online; accessed 2015].

HUSSEIN, M. E., TORKI, M., GOWAYYED, M. A., AND EL-SABAN, M. 2013. Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, AAAI Press, 2466–2472.

LEGGETTER, C. J., AND WOODLAND, P. C. 1995. Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models. *Computer Speech & Language 9*, 2, 171–185.

LEGGETTER, C., AND WOODLAND, P. 1995. Flexible speaker adaptation using maximum likelihood linear regression. In *Proc. ARPA Spoken Language Technology Workshop*, vol. 9, Citeseer, 110–115.

LV, F., AND NEVATIA, R. 2006. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. In *Computer Vision–ECCV 2006*. Springer, 359–372.

MÜLLER, M., BAAK, A., AND SEIDEL, H.-P. 2009. Efficient and robust annotation of motion capture data. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ACM, 17–26.

MÜLLER, M. 2007. *Information retrieval for music and motion*, vol. 2. Springer.

PAZHOUMAND-DAR, H., LAM, C.-P., AND MASEK, M. 2015. Joint movement similarities for robust 3d action recognition using skeletal data. *Journal of Visual Communication and Image Representation 30*, 10–21.

POHL, H., AND HADJAKOS, A. 2010. Dance pattern recognition using dynamic time warping. *Sound and Music Computing 2010*.

RABINER, L. R. 1989. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE 77*, 2, 257–286.

RAPTIS, M., KIROVSKI, D., AND HOPPE, H. 2011. Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation*, ACM, 147–156.

TILLMANNE, J., AND D'ALESSANDRO, N. 2015. Motionmachine: A new framework for motion capture signal feature prototyping. *Proceedings of EUSIPCO 2015, To appear*.

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*, IEEE, 588–595.