AUDIO-DRIVEN HUMAN BODY MOTION ANALYSIS AND SYNTHESIS

F.Ofli[a], C. Canton-Ferrer[b], J. Tilmanne[c], Y. Demir[a], E. Bozkurt[d], Y. Yemez[a], E. Erzin[a], A. M. Tekalp[a]

[a]Koç University, Istanbul, Turkey
[b]Technical University of Catalonia, Barcelona, Spain
[c]Faculty of Engineering of Mons, Mons, Belgium
[d]Momentum Technologies, Turkey

ABSTRACT

This paper presents a framework for audio-driven human body motion analysis and synthesis. We address the problem in the context of a dance performance, where gestures and movements of the dancer are mainly driven by a musical piece and characterized by the repetition of a set of dance figures. The system is trained in a supervised manner using the multiview video recordings of the dancer. The human body posture is extracted from multiview video information without any human intervention using a novel marker-based algorithm based on annealing particle filtering. Audio is analyzed to extract beat and tempo information. The joint analysis of audio and motion features provides a correlation model that is then used to animate a dancing avatar when driven with any musical piece of the same genre. Results are provided showing the effectiveness of the proposed algorithm.

Index Terms— multicamera motion capture, audio-driven body motion synthesis, dancing avatar animation

1. INTRODUCTION

In a typical dance performance, the body movements of the dancer are primarily driven by, and hence, highly correlated with the musical audio signal. The work presented in this paper can be thought of as a first attempt to model this correlation towards the goal of automatic synthesis of a dancing avatar driven by musical audio.

In the signal processing literature, there exists little research that addresses the problem of audio-driven human body motion synthesis. The most relevant literature is on speech-driven lip animation [1]. Since lip movement is physiologically tightly coupled with acoustic speech, it is relatively an easy task to find a mapping between the phonemes of speech and the visemes of lip movement. Many schemes exist to find such audio-to-visual mappings among which the HMM (Hidden Markov Model)-based techniques are the most common as they yield smooth animations exploiting temporal dynamics of speech. Some of these works also incorporates synthesis of facial expressions along with the lip movements to make animated faces look more natural [2, 3, 4]. The recent works that study the correlation between head gestures and speech prosody [5] or between hand gestures and speech content [6] towards the goal of more realistic speaker animation can also be considered in the same context.

The analysis and synthesis of body movements driven by musical audio pose more difficult challenges as compared to the speaker animation problem. In the first place, the body motion patterns, i.e, the dance figures, are usually very complicated in structure, having certain syntactic rules and hierarchies of figures. They are open to interpretation, and exhibit variations in time even for the same person. Secondly, the characteristic features of a musical audio signal, such as beat, tempo and tune, that are important in driving the dance performance are not well defined and hence need to be studied from the signal processing perspective.

We address audio-driven body motion analysis and synthesis problem considering a simple but illustrative scenario in a dance performance. The dancing avatar automatically classifies the genre of a given musical piece and associates with it a single dance figure that has been learnt from manually segmented multiview video sequences of the dancer. Each dance figure is modeled and synthesized using an HMM structure, and synchronized with the musical audio signal using the beat information. A crucial task during avatar training is capturing the motion of the dancer, and one major contribution of this work is a marker-based algorithm based on annealing particle filtering, that can automatically extract the human posture from multiview video without any human intervention.

2. SYSTEM OVERVIEW

The overall system, as depicted in Fig.1, comprises three modules: multimodal analysis (training), audio-driven body motion synthesis and animation. In the analysis block, multiview video sequences are analyzed in order to capture the time-varying posture of the dancer's body while audio is processed to extract beat information. The multiview videos are manually segmented into semantic recurring motion patterns: the dance figures. The corresponding body posture parameters are then used to train a set of HMMs, each of them modeling a different dance figure. Since the audio and video sequences are synchronized, each repetition of a dance figure determines a time segment from which the beat frequency associated with the figure can be estimated.

In the synthesis module, a given musical audio signal is first classified, within a time interval, into one of the genres that have been learnt in the analysis part. For genre classification, we rely on Mel Frequency Cepstral Coefficients (MFCC) and employ the HMM-based classification technique described in [7]. Beat information is then extracted in order to decide on the dance figure to synthesize and its duration. Afterwards, the system generates the body motion parameters associated with the chosen dance figures by using the corresponding HMMs, in synchrony with the beat information. Finally, the motion parameters, i.e., the angles of the body posture, are animated on a 3D stick body model.

Currently, our avatar has been trained to classify and dance only two genres, *salsa* and *belly*, and it is capable of making a single dance figure associated to each genre.

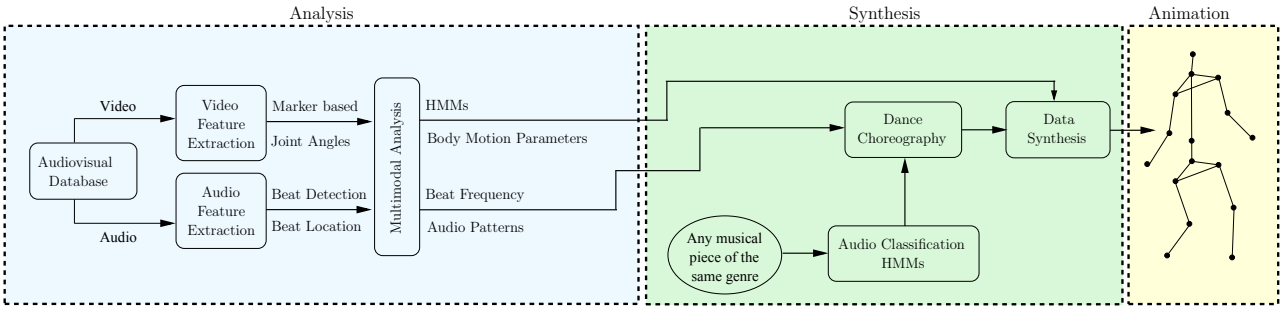

Fig. 1. Block diagram of the complete analysis-synthesis system.

3. MULTICAMERA BODY MOTION TRACKING

The motion capture process involves retrieving the body configuration in terms of its defining parameters, namely $\mathbf{\Theta}_t = \{\theta_0, ..., \theta_{M-1}\}_t$, from the multiple video streams at a given time t. These set of parameters include the articulation angles along with torso rotation and translation. A marker based approach is employed where a set of distinguishable color markers are attached to the joints of the dancer. There exist a number of marker-based commercial systems as evaluated in [8] for human motion capture but most of them rely on a high number of cameras to avoid occlusions, high frame rates or expensive hardware. In this work, we describe a low-cost method for multicamera marker-based body motion capture, that is accurate enough to train our dancing avatar.

For a given frame in the video sequence, a set of N images are obtained from the N cameras. Each camera is modeled using a pinhole camera model based on perspective projection. Accurate calibration information is available. In order to estimate the positions of the markers attached to the body of the dancer, the original images are processed in the CbCr color space where the marker color is efficiently detected and their locations are accurately estimated. The number of detected markers in every image may vary due to occlusions or low performance of the detection technique. However, tracking information and redundancy among views would allow overcoming this problem.

In order to analyze the incoming data, an articulated body model is employed. This body model allows exploiting the underlying anthropomorphic structure of the data [9]. The employed model is formed by a set of joints and links representing the limbs, head and torso of the human body and a given number of degrees of freedom (DoF) are assigned to each articulation (joint). Particularly, our model has 22 DoFs to properly capture all possible movements of the body (see an example of this in Fig.2).

We track the body angles $\mathbf{\Theta}_t$ along time using an Annealing Particle Filtering strategy [10]. This technique is employed to tackle estimation problems involving a high dimensional state space such as in this articulated human body tracking task. Two major issues must be addressed when employing particle filtering: likelihood evaluation and propagation model. The first establishes the observation model, that is, how a given configuration of the body matches the incoming data. For a given particle, we compute the 3D positions of the articulations by means of exponential maps [9] and then project them onto the N incoming images. In order to compute the likelihood of the detected markers against the projected position of the joints, we employ the robust symmetric epipolar distance introduced in [11]. This distance measures the closeness of a set of 2D points observed as the projections of the same 3D location from different views, exploiting the redundancy among cameras.

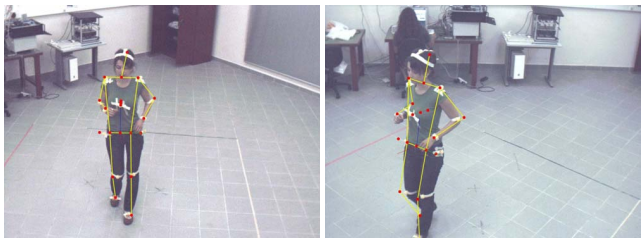

Fig. 2. An instance of the marker-based human body motion tracking process from two camera views. The articulated body model with 22 DoFs is represented as a stick model on the dancer's body.

The propagation model is adopted to add a drift to the parameters of the particles in order to progressively sample the state space in the following iterations [12]. Moreover, an underlying motion pattern is employed in order to efficiently sample the state space, thus reducing the number of particles required. This motion pattern is represented by the kinematic constrains and physical limits of the joints of the human body.

The described technique has produced satisfactory results with $N = 6$ cameras at 30 fps. The use of a body model together with a particle filtering strategy has proved to be robust against severe occlusions and poor measurements. An instance of the tracking process is shown in Fig.2.

4. MULTIMODAL ANALYSIS

4.1. Body Motion Analysis

In this research, HMMs are employed to model the dance figures, i.e., the body motion patterns recurring in the dance performance. The HMMs are trained with the parameter set resulting from the motion capture process, that includes the joint angles as well as the rotation and translation of the torso. For each dance figure, three separate HMMs are employed to better capture the dynamic behavior of the dancing body, one for the torso and two for the upper and lower parts of the body. The HMM structure for the upper part of the body models basically the movement of the arms while the one for the lower part models the movement of the legs.

A typical dance figure contains a well-defined sequence of movements, hence we employ a left-to-right HMM structure to model each figure. Each body posture parameter is represented by a single Gaussian function and one full covariance matrix is computed for each HMM model. This rather simple scheme leads

to satisfactory results without need for more complicated HMM configurations.

4.2. Audio Analysis

Among various features that characterize a musical audio signal, such as tonality, harmony or melody, tempo is the one that primarily drives and synchronizes the dancing act. Hence we have employed tempo and the relevant beat information as the audio features that drive our dancing avatar. We estimate the tempo in terms of beats per minute (BPM) using the algorithm suggested in [13]. Tempo estimation involves three basic tasks: onset detection, periodicity estimation and beat location estimation. Onset detection aims to point out where musical notes start, and tempo is established by the periodicity of the detected onsets. Beat location is computed directly from periodicity estimation.

First, onsets are detected based on the spectral energy flux of the input audio signal, that signifies one of the most salient features. Onset detection is determining, since beat tends to occur at onsets. Next, the periodicity is estimated from the detected onsets using an autocorrelation based method. Once the periodicity is determined, the tempo can be calculated in terms of BPM. Finally, beat locations are estimated by generating an artificial pulse train with the estimated periodicity and by cross-correlating it with the onset sequence. Maximum values of this function marks the starting of a beat location. See Fig.3 for an example of this process.

Beat information allows estimating the tempo for each dance figure, typically ranging between 60 and 200 BMP. Analysis results of our experiments show that the average tempo is 185 BMP for *salsa* and 134 BMP for *belly*. We have also observed that a *salsa* dance figure in our training video comprises 8 beats whereas a *belly* dance figure corresponds to 3 beats. We make use of this information in the synthesis step to determine the beginning and ending frames of a dance figure.

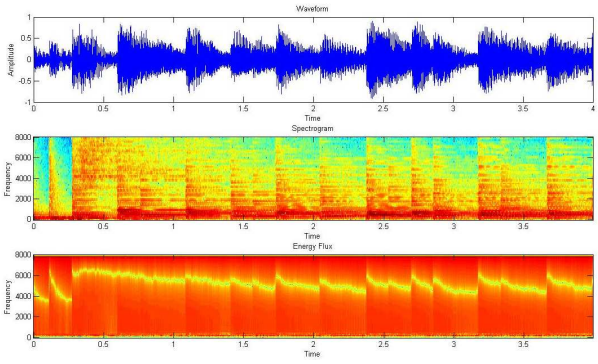

Fig. 3. Beat detection example: time waveform, spectrogram and spectral energy flux of 4 seconds of *salsa* type music computed with a 50% overlap analysis window.

5. SYNTHESIS

The goal of the synthesis stage is to generate the corresponding body posture parameters synchronized with a test musical audio signal. The first task is to classify the audio signal with respect to its genre (salsa or belly in our case) over sliding windows. For this, we use MFCCs and employ the HMM-based classification technique described in [7]. The classified audio tracks are then analyzed to extract the beat and tempo information via the method explained in Section 4.2. The genre of the audio track determines the dance figure to be synthesized (recall that in our training video there is only one single figure associated with each genre) whereas the beat locations and the tempo information determine the duration and location of the figure. We note that the beat frequency for the same dance figure may vary within a musical audio signal or from one piece to another.

The body posture parameters corresponding to each dance figure are generated using the associated HMM structures learnt at the motion analysis stage (see Section 4.1). For each dance figure, we construct a single HMM structure by coupling the individual HMM models that are trained separately for the torso and the upper and lower parts of the body. The states of each such coupled HMM structure correspond to the motion patterns that form the dance figure.

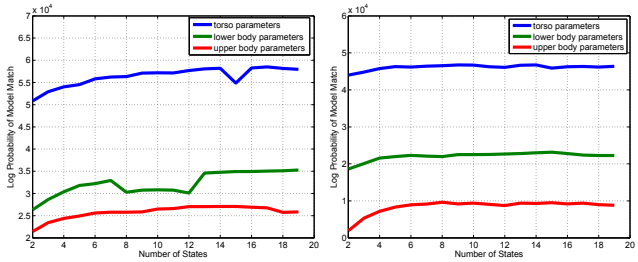

Fig. 4. Evolution of the logarithmic probability of the model match with varying number of states for the 6 HMM structures (three for *salsa* on the left and three for *belly* on the right).

6. EXPERIMENTS AND RESULTS

Our training dataset includes multiview video recordings of two dance performances, one for *salsa* and one for *belly*, each with a duration of approximately 5 minutes. The performances are recorded synchronously from 6 cameras at 30 fps. Each video recording consists of one single dance figure repeated successively during the whole performance.

For motion analysis, we manually label the start and end frames of each dance figure throughout the entire dance recordings. The three HMM models of each dance figure, for the torso and the upper and lower parts of the body, are trained in a supervised manner with the body posture parameters captured from these manually labeled segments.

In order to determine the optimal number of states for each of the HMMs, we train each HMM with different number of states (varying from 2 to 19). By computing the average logarithmic probability of the model match for each value, we examine the progression of the learning process and the accuracy of the trained model. The evolution of this parameter for the totality of the 6 HMM structures that we trained is displayed in Figure 4. We observe that the optimal number of states is related to the complexity of the dance figure. In the case of the salsa figure, which is more complicated than the belly, the optimal numbers are around 13, 12 and 14 for lower body, upper body and torso, respectively, whereas these numbers are around 6, 8 and 5 for the belly figure. To determine the optimal number of states, we basically search for the point where the plots start to saturate since we also want to keep the number of states, and hence the model complexity, as low as possible.

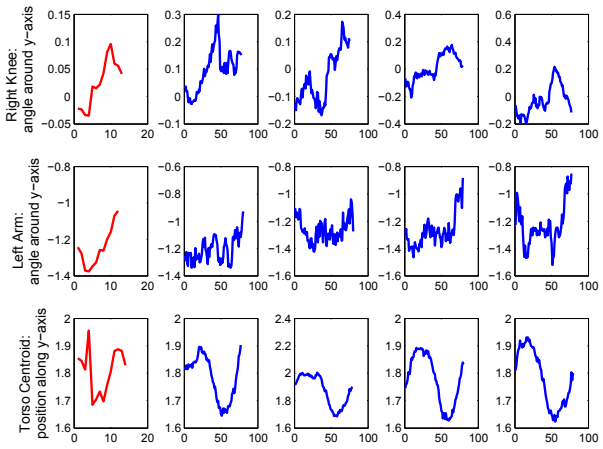

Fig. 5. For the salsa figure, variation of the means of three parameters over the HMM states (plotted in red) and evolution of the same three parameters during four different realizations sampled from the training video (plotted in blue).

In order to verify that the posture parameters are correctly modeled with the resulting HMMs, in Fig. 5, we compare, for some parameters, the evolution of the means of their Gaussian distributions over the HMM states with the evolution of the same parameters through the realizations of the corresponding dance figures in the training data set. The shapes of the evolution are clearly observed to be similar, even for the parameters which show significant variations from one realization to another in the training set and are thus difficult to model.

The musical audio signals are recorded at 16 kHz as 16 bit mono PCM wavefiles. The signals are analyzed over a 25 ms Hamming window at every 10 ms. The set of 13 MFCCs along with their first and second derivatives, adding up to a total of 39 features, forms the audio feature vector for the genre classification task. Use of MFCCs as the only audio feature set becomes sufficient for the classification problem in our case, since we have only two types of musical audio, *salsa* or *belly*.

We have considered several animation scenarios for demonstration of our dancing avatar. In the first scenario, we mix two audio tracks of different genres, *salsa* and *belly*, and use this mixed track as the animation audio to show that the avatar can successfully recognize the changing audio and synthesize the correct dance figures. In the second scenario, we first slow down and then speed up the audio track to demonstrate that the avatar can keep track of the changing beat information and adjust the speed of the dance movements accordingly. In the final scenario, we take an arbitrary audio which is neither salsa nor belly to see how the avatar adapts itself to a different genre that it has not been trained for. Demo videos of these scenarios are available online at *http://mvgl.ku.edu.tr/bodymotionanalysis/icassp08/*.

7. CONCLUSIONS

This paper presents a novel framework for audio-driven human body motion analysis and synthesis. We have addressed the problem in the context of dance performance and considered a simple scenario possible in which only a single dance figure is associated with each musical genre. Currently, our dancing avatar has been trained for *salsa* and *belly*. The experiments show that the avatar can successfully recognize the genre changes in a given audio track and synthesize the correct dance figures in a very realistic manner. The avatar can also keep track of the changing beat information and adjust the speed of the dance movements accordingly.

A crucial task during avatar training is to capture the motion of the dancer in an accurate manner. For this, we have developed a marker-based algorithm based on annealing particle filtering, that can automatically extract the human posture from multiview video without any human intervention.

Future research within this topic involve unsupervised training of the dancing avatar for different musical genres in more complicated scenarios in which the dance figures are more sophisticated in structure, having certain syntactic rules and hierarchies of figures. To achieve this, we will also need to consider various musical audio features other than beat and tempo, such as tonality, harmony and melody.

8. REFERENCES

[1] T. Chen, "Audiovisual speech processing," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 9–21, 2001.

[2] M. Brand, "Voice puppetry," in *Computer graphics and Interactive Techniques (SIGGRAPH), Proc. Int. Conf. on*, New York, NY, USA, 1999, pp. 21–28.

[3] Y. Li and H.Y. Shum, "Learning dynamic audio-visual mapping with input-output hidden markov models," *Multimedia, IEEE Trans. on*, vol. 8, no. 3, pp. 542–549, 2006.

[4] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp, "Estimation and analysis of facial animation parameter patterns," in *Image Processing, IEEE Int. Conf. on*, 2007.

[5] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, and M. Ozkan, "Prosody-driven head-gesture animation," in *Acoustics, Speech and Signal Processing, IEEE Int. Conf. on*, 2007, vol. 2, pp. 677–680.

[6] M. E. Sargin, O. Aran, A. Karpov, F. Ofli, Y. Yasinnik, S. Wilson, E. Erzin, Y. Yemez, and A. M. Tekalp, "Combined gesture - speech analysis and speech driven gesture synthesis," in *Multimedia and Expo, IEEE Int. Conf. on*, 2006, pp. 893–896.

[7] U. Bagci and E. Erzin, "Automatic classification of musical genres using inter-genre similarity," *IEEE Signal Processing Letters*, vol. 14, no. 8, pp. 521–524, August 2007.

[8] Y. Ehara, H. Fujimoto, S. Miyazaki, S. Tanaka, and S. Yamamoto, "Comparison of the performance of 3d camera systems," *Gait and Posture*, vol. 3, no. 3, pp. 166–169, Sep. 1995.

[9] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *Computer Vision and Pattern Recognition, Proc. IEEE Int. Conf. on*, 1998.

[10] J. Deutscher and I. Reid, "Articulated body motion capture by stochastic search," *Int. Journal of Computer Vision*, vol. 61, no. 2, pp. 185–205, Feb. 2005.

[11] C. Canton-Ferrer, J. R. Casas, and M. Pardàs, "Towards a Bayesian approach to robust finding correspondences in multiple view geometry environments," in *Lecture Notes on Computer Science*, 2005, vol. 3515, pp. 281–289.

[12] M.S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for online nonlinear/non-Gaussian Bayesian tracking," *Signal Processing, IEEE Tran. on*, vol. 50, no. 2, pp. 174–188, 2002.

[13] M. Alonso, B. David, and G. Richard, "Tempo and beat estimation of music signals," in *Music Information Retrieval, Proc. Int. Conf. on*, 2004.