UNSUPERVISED DANCE FIGURE ANALYSIS FROM VIDEO FOR DANCING AVATAR ANIMATION*

F. Ofli, E. Erzin, Y. Yemez, A. M. Tekalp

College of Engineering, Koç University
34450 Sarıyer, İstanbul, Turkey

Ç. E. Erdem, A. T. Erdem, T. Abacı, M. K. Özkan

Momentum A. Ş.
Kocaeli, Turkey

ABSTRACT

This paper presents a framework for unsupervised video analysis in the context of dance performances, where gestures and 3D movements of a dancer are characterized by repetition of a set of unknown dance figures. The system is trained in an unsupervised manner using Hidden Markov Models (HMMs) to automatically segment multi-view video recordings of a dancer into recurring elementary temporal body motion patterns to identify the dance figures. That is, a parallel HMM structure is employed to automatically determine the number and the temporal boundaries of different dance figures in a given dance video. The success of the analysis framework has been evaluated by visualizing these dance figures on a dancing avatar animated by the computed 3D analysis parameters. Experimental results demonstrate that the proposed framework enables synthetic agents and/or robots to learn dance figures from video automatically.

Index Terms— unsupervised human body motion analysis, dance figure identification, dancing avatar animation

1. INTRODUCTION

State of the art human motion analysis research is devoted to detect, track and interpret human behaviors from image sequences. Comprehensive surveys of human motion analysis research can be found in [1, 2, 3]. Estimation of 3D body posture is a key problem in human body motion analysis, especially for interpreting human behaviors towards realistic body motion synthesis and animation. Some works rely on direct use of a motion capture process to estimate the set of body posture parameters without explicitly modeling the temporal dynamics of the body motion [4, 5, 6, 7]. They mostly aim at regenerating the original body posture parameters in a video and synthesizing articulated 3D human actions as human behaviors like walking, running, sitting down, standing up, climbing stairs, etc.

The analysis and synthesis of body movements in the context of a dance performance pose new challenges. In the first place, the body motion patterns, i.e, the dance figures, are very complex, but they usually follow certain syntactic rules and hierarchies. Furthermore, they are open to interpretation, and exhibit variations in time even for the same person. Mori and Hoshino propose an independent component analysis based technique to extract the motion signal corresponding to the perceptually meaningful motion features that characterize the complex human body movements with applications to dance scenarios [8]. Brand and Hertzmann introduce the idea of "style machines," where the same class of motions with different style (such as walking slowly, rapidly etc.) is described by HMMs and other stylistic motions can be generated by the analysis results of these HMMs [9]. Nakazawa et al. introduce the notion of dividing the human motion into some motion primitives that consist of a "basic motion" and a "motion style" [10, 11]. The basic motion is assumed to be common to all dancers, and the style represents their characteristics.

We address the body motion analysis and synthesis problem in the context of dance performances, considering a fully automatic system which can be then used to drive the movements of a dancing avatar. The main novelty of this paper is to perform an unsupervised analysis of dance figures from the dance video recordings where each dance figure is modeled and synthesized using a set of HMM structures. The HMM structure is an extension of that proposed by [12] to model head gestures by studying the correlation between head gestures and speech prosody. Our previous work on audio-visual dance analysis considered supervised HMM analysis of audio, where dance figures were manually marked [13, 14]. Our main goal in this paper is unsupervised analysis of the variations between dance figures of certain dancers and dances from video. In order to analyze these variations, syntactic modeling of dance figures in terms of elementary dance motion patterns is desired and this paper aims to determine elementary dance motion patterns by performing an unsupervised clustering of body posture parameters. By means of such modeling of elementary dance motion patterns and complete dance figures, we hope to arrive at a syntactic dance description language although this is beyond the scope of the current paper.

2. SYSTEM DESCRIPTION

The overall system comprises of two modules: dance figure analysis and dance figure animation. In the analysis block, multiview video sequences are analyzed in order to capture the time-varying posture of the dancer's body. The body posture parameters are then used to temporally segment the multiview videos into semantic recurring dance motion patterns by training a set of HMMs, each of them modeling a different dance figure. The animation module generates dance figures using the computed body motion parameters on a dancing avatar. Detailed descriptions of the analysis and animation modules are given in Section 3 and 5, respectively. Currently, our avatar has been trained to dance only three genres, *salsa*, *belly* and *folk*.

3. DANCE FIGURE ANALYSIS

In this research, HMMs are employed to model the dance figures, i.e., the elementary body motion patterns recurring in the dance performances. The HMMs are trained with the parameter set resulting

*THIS WORK HAS BEEN SUPPORTED BY TURKISH ACADEMY OF SCIENCES (TUBA) AND EUROPEAN COMMISSION WITHIN FP6 UNDER GRANT 511568 WITH THE ACRONYM 3DTV.

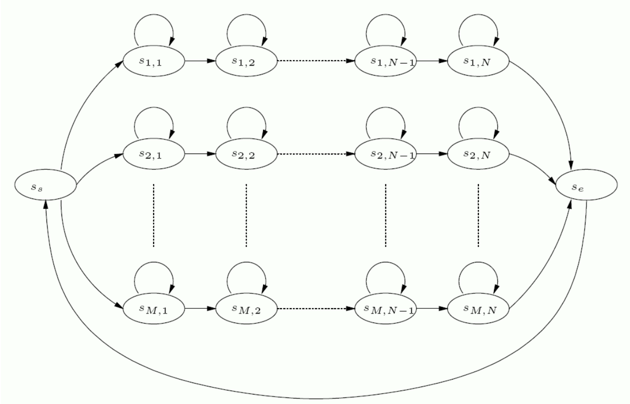


Fig. 1. Parallel HMM structure.

from the motion capture process [14], that includes the 3D joint positions. For each dance figure, two separate HMMs are employed to better capture the dynamic behavior of the dancing body, one for the upper part and one for the lower part 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. The temporal segmentation of the body motion parameters is performed using Viterbi decoding to maximize the probability of model match, which is the probability of body motion parameters given the trained HMM. This rather simple scheme leads to satisfactory results without need for more complex HMM configurations.

The relatively crucial task is to automatically determine the number of body motion patterns, i.e., the number of different dance figures, for the upper and lower parts of the body in the given video recordings. Since different figures may have different representation complexities and the duration of a figure may vary from one dance figure to another, one should also consider optimizing the number of states in the HMM structures that correspond to different dance figures. Consequently, there are two important quantities to be determined before training the HMMs that model the dance figures: the number of different dance motion patterns in a given video recording and the number of states necessary for each of these different dance motion patterns.

Parallel HMM structures are employed to determine the two important parameters before training of the complete model (see Figure 1). This HMM structure has M parallel branches and N states which are to be optimized jointly. An iterative approach is used for selection of M and N. For varying values of M and N, two fitness measures are checked. The first fitness measure is the average logarithmic probability of model match, which increases with the increasing number of temporal dance motion patterns. Consequently, the second fitness measure, which is the average statistical separation between two similar temporal dance motion patterns, increases with the decreasing number of temporal dance motion patterns. Each branch in these parallel HMM structures corresponds to different dance motion pattern that exists in the dance video recordings. Therefore, the parallel HMM structure with the optimum number of branches can be interpreted as the collection of models for different dance figures.

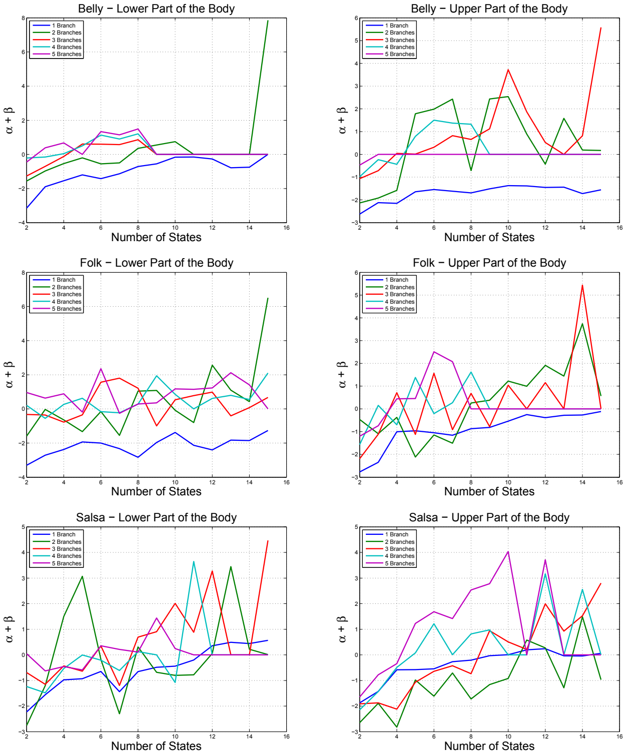


Fig. 2. Joint maximization of the logarithmic probability of the model match and average statistical separation between two similar temporal body motion patterns with varying number of states for the 6 HMM structures (two for *belly* in the first row, two for *folk* in the second row and two for *salsa* in the last row) Different lines represent different number of branches.

4. ANALYSIS EXPERIMENTS AND RESULTS

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

For body motion analysis, we automatically segment the figures from the video and determine the start and end frames of each dance figure throughout the entire dance recordings. The two HMM models of each dance figure, for the upper and lower parts of the body, are trained in an unsupervised manner with the body posture parameters captured from the multiview dance recordings.

In order to determine the optimal number of branches and states for each of the HMMs, we train each parallel HMM with different number of branches (varying from 1-5) and of states (varying from 2 to 15). By computing the average logarithmic probability of the model match and average statistical separation between two similar temporal dance motion patterns for each iteration, we examine the progression of the learning process and the accuracy of the trained model. The evolution of these parameters for the totality of the 6

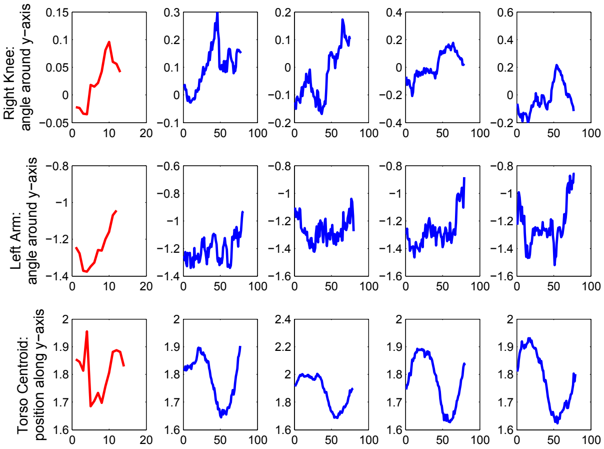


Fig. 3. For the salsa figure, variation of the means of three parameters over the HMM states (plotted in red) and evolution of the same three parameters during four different realizations sampled from the training video (plotted in blue)

HMM structures that we trained is displayed in Figure 2.

The optimal number of HMM branches and states deduced from Figure 2 is the point that maximizes the plot. We observe that the optimal numbers are related to the complexity of the dance figure. In the case of the *salsa* figure, which is more complicated than the *belly* and the *folk*, the optimal numbers are around 5 branches-10 states for the upper part of the body and 3 branches-15 states for the lower part of the body whereas these numbers are around 3 branches-15 states and 2 branches-15 states, for the upper and lower parts of the body, respectively, for the *belly* figure; and 3 branches-14 states and 2 branches-15 states, for the upper and lower parts of the body, respectively, for the *folk* figure.

In order to verify that the posture parameters are correctly modeled with the resulting HMMs, in Figure 3, 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.

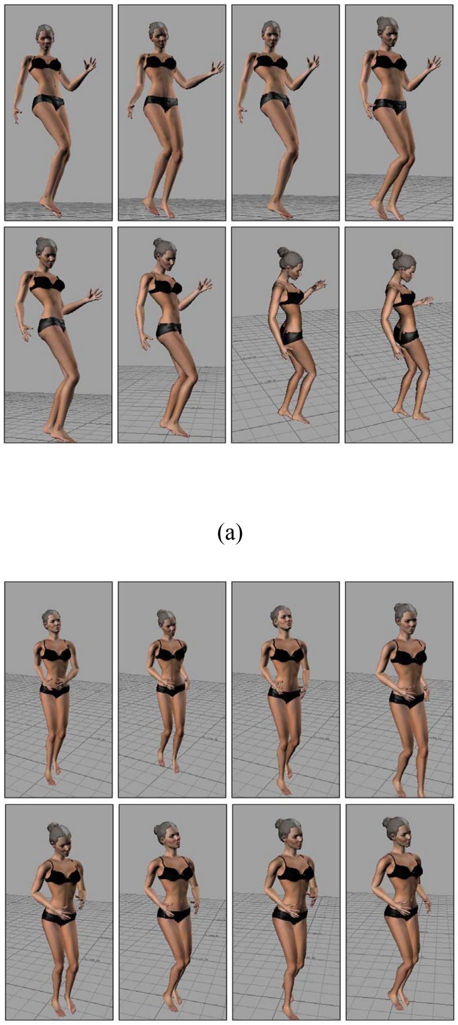
5. DANCE ANIMATION

In order to animate a virtual 3D character using the output of the analysis stage, we used a specialized commercial software package [15]. For our purposes, it makes sense to treat each set of body posture parameters segmented by the analysis stage as a new set of motion capture data (marker positions). Therefore, we have created the animation sequences in Figure 4 by importing the analysis results into the aforementioned software together with a pre-designed 3D human body.

The process of animating a virtual character is outlined in Figure 5. Usually, the process starts by manually fitting the character (actor) to a well-defined pose (ideally, a *T-pose*) to estimate dimensions at the motion capture stage. Our original motion capture data did not include a T-pose, but we were still able to obtain acceptable results by using a similar pose selected from one of the output sequences.

In order to determine how the motion capture data is to be interpreted, it is necessary to assign markers to *actor cells*. The set of assignments we have chosen to employ is depicted in Figure 6, where the circles represent the cells. Some cells required more than one marker to behave properly during the animation. Among these, the torso was somewhat problematic, which made it necessary for us to set a fixed orientation for the upper torso in the sequences for all genres (*salsa*, *belly* and *folk* dance). The most likely cause for this is the poor choice of marker positions in the torso area, especially for marker 9.

Once the body posture parameters are successfully imported as motion capture data, the virtual character is animated and rendered using standard matrix palette skinning techniques, as shown in Figure 4. Video sequences for the animated dance figures can be found at *http://mvgl.ku.edu.tr/bodymotionanalysis/icip08/*.



(a)

(b)

Fig. 4. (a) Animation results for automatic belly dance synthesis. (b) Animation results for automatic salsa dance synthesis.

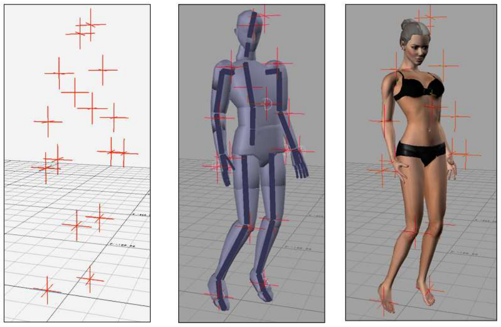


Fig. 5. Animation process outline.

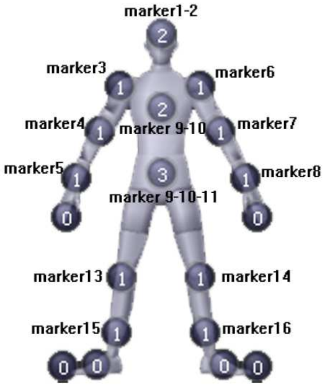


Fig. 6. Marker assignments.

6. CONCLUSIONS

The HMM based unsupervised segmentation scheme proves to be successful in determining the number of different dance figures in a given video recording and in optimizing the number of states in the corresponding HMMs to better capture the dynamics of the body motion patterns. Currently, our dancing avatar has been trained for *salsa*, *belly* and *folk*. The proposed framework can also be useful for the analysis of movement disabilities and monitoring of physical activity in the elderly.

It is important that we can generate various human body motion patterns easily and realistically to create attractive content for dance synthesis. The experiments show that the avatar can successfully synthesize the dance figures in a very realistic manner.

7. ACKNOWLEDGEMENT

We would like to thank Gökhan Uysal from Momentum A. Ş. for preparing the figures of animation.

8. REFERENCES

[1] J. J. Wang and S. Singh, “Video analysis of human dynamics - a survey,” *Real-Time Imaging*, vol. 9, no. 5, pp. 321–346, October 2003.

[2] L. Wang, W. Hu, and T. Tan, “Recent developments in human motion analysis,” *Pattern Recognition*, vol. 36, no. 3, pp. 585–601, March 2003.

[3] R. Poppe, “Vision-based human motion analysis: An overview,” *Computer Vision and Image Understanding*, vol. 108, no. 1-2, pp. 4–18, October-November 2007.

[4] J.-C. Cheng and J.M.F. Moura, “Capture and synthesis of human motion in video sequences,” *IEEE 2nd Workshop on Multimedia Signal Processing*, pp. 111–116, 7-9 December 1998.

[5] J. Lee, J. Chai, P. S. A. Reitsma, J. K. Hodgins, and N. S. Pollard, “Interactive control of avatars animated with human motion data,” *ACM Transactions on Graphics*, vol. 21, no. 3, pp. 491–500, 2002.

[6] S. Yonemoto and R. Taniguchi, “Human figure control software for real-virtual application,” 14-16 July 2004, pp. 858–862.

[7] F. Remondino and A. Roditakis, “Human motion reconstruction and animation from video sequences,” in *17th International Conference on Computer Animation and Social Agents*, July 2004, pp. 347–354.

[8] H. Mori and J. Hoshino, “Independent component analysis and synthesis of human motion,” 2002, vol. 4, pp. 3564–3567.

[9] M. Brand and A. Hertzmann, “Style machines,” in *SIGGRAPH’00: Proceedings of the 27th Annual Conf. on Computer Graphics and Interactive Techniques*, 2000, pp. 183–192.

[10] A. Nakazawa, S. Nakaoka, T. Shiratori, and K. Ikeuchi, “Analysis and synthesis of human dance motions,” 14-19 September 2003, vol. 3, pp. 3899–3904.

[11] A. Nakazawa, S. Nakaoka, and K. Ikeuchi, “Synthesize stylistic human motion from examples,” 14-19 September 2003, vol. 3, pp. 3899–3904.

[12] M. E. Sargin, E. Erzin, Y. Yemez, A. M. Tekalp, A. T. Erdem, C. Erdem, and M. Ozkan, “Prosody-driven head-gesture animation,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2007, vol. 2, pp. 677–680.

[13] F. Ofli, Y. Demir, E. Erzin, Y. Yemez, and A. M. Tekalp, “Multicamera audio-visual analysis of dance figures,” 2-5 July 2007, pp. 1703–1706.

[14] F. Ofli, C. Canton-Ferrer, J. Tilmanne, Y. Demir, E. Bozkurt, Y. Yemez, E. Erzin, and A. M. Tekalp, “Audio-driven human body motion analysis and synthesis,” in *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing*, 2008.

[15] “Autodesk motion builder,” www.autodesk.com/motionbuilder.