15th European Signal Processing Conference (EUSIPCO 2007), Poznan, Poland, September 3-7, 2007, copyright by EURASIP

MULTICAMERA AUDIO-VISUAL ANALYSIS OF DANCE FIGURES USING SEGMENTED BODY MODEL

*F. Ofli, Y. Demir, E. Erzin, Y. Yemez, and A. M. Tekalp**

Multimedia, Vision and Graphics Laboratory
Koç University,
Sarıyer, Istanbul, 34450, Turkey
{fofli,ydemir,eerzin,yyemez,mtekalp}@ku.edu.tr

ABSTRACT

We present a multi-camera system for audio-visual analysis of dance figures. The multi-view video of a dancing actor is acquired using 8 synchronized cameras. The motion capture technique of the proposed system is based on 3D tracking of the markers attached to the person's body in the scene. The resulting set of 3D points is then used to extract the body motion features as 3D displacement vectors whereas MFC coefficients serve as the audio features. In the multi-modal analysis phase, we perform Hidden Markov Model (HMM) based unsupervised temporal segmentation of the audio and body motion features such as legs and arms, separately, to determine the recurrent elementary audio and body motion patterns in the first stage. Then in the second stage, we investigate the correlation of body motion patterns with audio patterns that can be used towards estimation and synthesis of realistic audio-driven body animation.

1. INTRODUCTION

Human body motion analysis has been an interesting research topic in computer vision due to its various applications, such as animation, athlete training, medical diagnostics, virtual reality, and human-machine interfaces. In the analysis of human body motion, three tasks are involved: tracking and estimating the motion parameters, analyzing the human body structure, and recognizing the motion activities. For animation, detailed skeletal body models are commonly applied.

Motion capture systems have continuously been evolving and there exist already various techniques and approaches in the literature, that can be distinguished mainly based on whether they make use of markers (active or passive), or fully rely on image features, and the type of motion analysis they employ (model-based or not). The simultaneous recovery of pose and body shape from video streams has been considered [1]. Optical flow and probabilistic body part models were used to fit a hierarchical skeleton to walking sequences [2].

Much previous work has been done in modeling complex human motion model and they can be largely categorized into two classes. The first class is by supervised learning. Mixture motion model is used for tracking in [3]. But the primitives are pre-defined and segmented manually for training. The second class of approach, unsupervised or semi-unsupervised human motion modeling, avoids such tedious and error prone process of manual segmentation. In [4], HMM(hidden Markov model) is learnt for human locomotion (walking, running). But the topology of the HMM is given and it is difficult to extend it to more complex motion.In [5] HMM is used to analyze dance figures of a dancing person. In [6], each primitive follows a different dynamic law (acceleration) which can be used to differentiate each other. Variable length Markov models (VLMM) [7] were learnt to model human behavior. However, simple heuristics such as low velocity points at the boundary of two primitives was employed for segmentation. SLDS (switching linear dynamic systems) are learnt in [8] for classifying human motion.

In this work audio-visual analysis of dance figures is presented. 3D world points related to 16 human body joints are used to analyze the correlation between the audio patterns and body motion patterns according to [9, 5].

2. MULTICAMERA MOTION CAPTURE

Our motion capture technique employs an optical flow method to record subject's motion where a set of markers are attached to the subject and then observed by a number of cameras. These markers are located at 16 different points on the body as can be seen in Figure 1. Markers in each video frame are detected by applying thresholds over their chrominance information. In this setting, the motion capture system determines the 3D position of each marker at each frame by triangulation based on the observed projections of the markers onto each camera's image plane. The 3D positions of the markers are tracked over the frames by Kalman filtering where the filter states correspond to 3D position and velocity of each marker. The list of 3D points obtained by back-projection of 2D points in respective camera image planes constitute the observations for this filter. The list of 3D marker positions over frames is our body model features that will be used in the analysis and animation process.

3. AUDIO-VISUAL DANCE ANALYSIS

In this section, a two-step analysis framework based on unsupervised temporal segmentation is considered. The first stage analysis aims to extract elementary audio patterns and body motion patterns separately as left leg, left arm, right arm and right leg. The correlation between these parts are determined by the co-occurrence matrices. In the second stage analysis, the correlation between audio patterns and body motion patterns is investigated.

3.1. Body Motion Patterns

Body motion patterns are extracted from 3D displacement vectors of 16 points located on the joints of the person's body. The displace-

*This work has been supported by the European FP6 Network of Excellence SIMILAR.

Fig. 1. Dance scene captured by the 8-camera system available at Koç University. Markers are attached at or around the joints of the body.

ment vectors are calculated relative to the reference frame after subtracting the rotational and translational motions which can be represented as a transformation matrix for the body as a whole. This transformation matrix is calculated using the torso which is composed of four points located on the hips, chest and back of the subject. Points are defined in homogenous coordinates such as $\vec{p} - 1 = [x_1 y_1 z_1 1]$. The transformation matrix is calculated relative to the first frame. Let $M = [\vec{p}_1 \vec{p}_2 \vec{p}_3 \vec{p}_4]$ be 4x4 invertible matrix composed of initial locations of each torso joint. The locations of these points in i^{th} can be given in a similar matrix format, $M^i = [\vec{p}_1' \vec{p}_2' \vec{p}_2' \vec{p}_3']$. The 4x4 transformation matrix M_{proj} is calculated as $M_{proj} = (M^i - \vec{m}) \times (M - \vec{m})^{-1}$ where $\vec{m}$ is the mean of the points located on hips and shoulders in the first frame. Each initial point in the first frame is projected to the current frame by multiplying with the transformation matrix M_{proj} and features are calculated as the differences of original point coordinates and the projected initial points, i.e., $\mathbf{F}^b = M_{proj} \times \vec{p}^0 - \vec{p}^i$ where $\vec{p}^i$ and $\vec{p}^0$ are the location of points in current and initial frames, respectively.

3.2. Audio Features

The act of dancing is the natural response of the body to the rhythm of the sound. MFCCs are good choices for representing the audio features in our scenario since they approximate the human auditory system's response to the sound. According to these responses the movements of the body is shaped and dance figures are generated that are correlated with the audio.

3.3. Unsupervised Temporal Segmentation

The HMM structure $\mathbf{\Lambda}$ has M parallel branches and N states. The parallel HMM $\mathbf{\Lambda}$ is composed of M parallel left-to-right HMMs, $\{\lambda_1, \lambda_2, \ldots, \lambda_M\}$, where each λ_m is composed of N states, $\{s_{m,1}, s_{m,2}, \ldots, s_{m,N}\}$. The state transition matrix $\mathbf{A}_{\lambda_m}$ of each λ_m is associated with a sub-diagonal matrix of $\mathbf{A}_\Lambda$. The feature stream is a sequence of feature vectors, $\mathbf{F} = \{\mathbf{f}_1, \mathbf{f}_2, \ldots, \mathbf{f}_T\}$, where $\mathbf{f}_t$ denotes the feature vector at frame t. Unsupervised temporal segmentation using HMM model $\mathbf{\Lambda}$ yields L number of segments $\varepsilon = \{\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_L\}$. The l^{th} temporal segment is associated with the following sequence of feature vectors,

$$\varepsilon_l = \{\mathbf{f}_{t_l}, \mathbf{f}_{t_l+1}, \ldots, \mathbf{f}_{t_{l+1}-1}\} \quad l = 1, 2, \ldots, L \tag{1}$$

where $\mathbf{f}_{t_1}$ is the first feature vector $\mathbf{f}_1$ and $\mathbf{f}_{t_{L+1}-1}$ is the last feature vector $\mathbf{f}_T$. The segmentation of the feature stream is performed using Viterbi decoding to maximize the probability of model match, which is the probability of feature sequence $\mathbf{F}$ given the trained parallel HMM $\mathbf{\Lambda}$,

$$\begin{aligned} \mathrm{P}(\mathbf{F}|\mathbf{\Lambda}) &= \max_{t_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\{\mathbf{f}_{t_l}, \mathbf{f}_{t_l+1}, \ldots, \mathbf{f}_{t_{l+1}-1}\}|\lambda_{m_l}) \\ &= \max_{\varepsilon_l, m_l} \prod_{l=1}^{L} \mathrm{P}(\varepsilon_l|\lambda_{m_l}) \end{aligned} \tag{2}$$

where ε_l is the l^{th} temporal segment, which is modeled by the m_t^{th} branch of the parallel HMM $\mathbf{\Lambda}$. One can show that λ_{m_l} is the best match for the feature sequence ε_l, that is,

$$m_l = \underset{m}{\operatorname{argmax}}\, \mathrm{P}(\varepsilon_l|\lambda_m) \tag{3}$$

Since the temporal segment ε_l from frame t_l to $(t_{l+1} - 1)$ is associated with segment label m_l, we define the sequence of frame labels based on this association as,

$$\ell_t = m_l \qquad \text{for } \; t = t_l, t_l + 1, \ldots, t_{l+1} - 1 \tag{4}$$

where ℓ_t is the label of the t^{th} frame and we have a label sequence $\ell = \{\ell_1, \ell_2, \ldots, \ell_T\}$ corresponding to the feature sequence $\mathbf{F}$. The first stage analysis extracts the frame label sequences ℓ^b and ℓ^a given the body motion and audio feature streams $\mathbf{F}^b$ and $\mathbf{F}^a$.

The parallel HMM structure has two important parameters to set before the training of the model $\mathbf{\Lambda}$. The first parameter is the number of states in each branch, N. It should be selected by considering the average duration of temporal patterns. N is selected to be $N_{\Lambda_b} = 10$, assuming minimum motion pattern duration is $\frac{1}{3}$ sec (10 frames). On the other hand, the number of temporal patterns for audio is set to $N_{\Lambda_a} = 5$ states in each branch of the audio HMM model $\mathbf{\Lambda}_a$ to model audio patterns.

The second parameter is the number of temporal patterns with the notation M. Finding an optimum value for M two fitness measures are checked where the first fitness measure, α, is the probability of model match and the second, β, is the average statistical

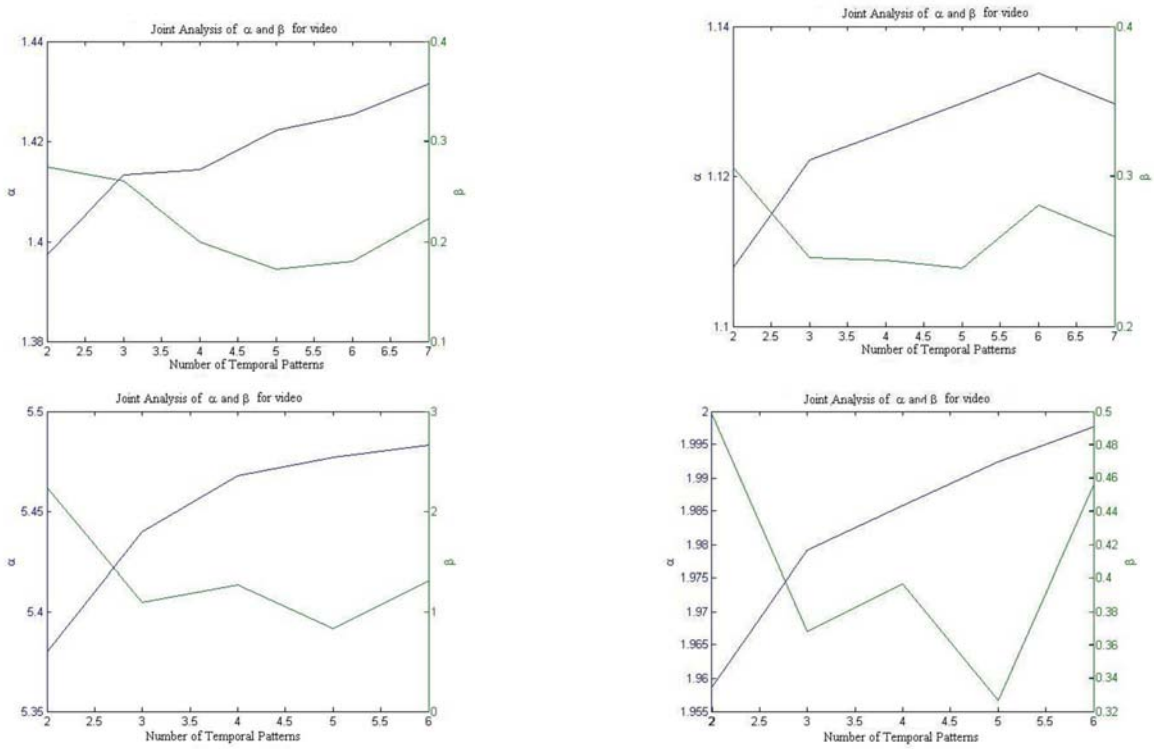

Fig. 2. Results of iterative approach for selection of M for the body motion patterns, upper left graphics is for for left leg and the upper right positioned graphics for right leg, left below graphics represents α and β measure for left arm and the graphics located right below represents for right arm.

separation between two similar temporal patterns. The value determined for M would be helpful for modeling the body motion patterns. Therefore, the total number of temporal patterns, M, can be selected in the vicinity of the intersection of the normalized α and β measures.The definitions for these two measures are given below in equations.

$$\alpha = \frac{1}{T}\log(\mathrm{P}(\mathbf{F}|\mathbf{\Lambda})) \tag{5}$$

$$\beta = \frac{1}{T}\sum_{l=1}^{L}\log(\frac{\mathrm{P}(\varepsilon_l|\lambda_{m_l})}{\mathrm{P}(\varepsilon_l|\lambda_{m_l^*})}) \tag{6}$$

where $\lambda_{m_l^*}$ is the second best match for the temporal segment ε_l, that is given as,

$$m_l^* = \underset{\forall m \neq m_l}{\operatorname{argmax}}\,\mathrm{P}(\varepsilon_l|\lambda_m) \tag{7}$$

3.4. Multimodal Analysis

The first stage analysis defines elementary recurrent body motion patterns for separate body parts using unsupervised temporal clustering over individual feature streams. The body motion feature streams $\mathbf{F}^b$ are used to train HMM structure $\mathbf{\Lambda}_b$ that captures recurrent body motion patterns ε^b. Audio feature streams $\mathbf{F}^a$ are similarly used to train HMM structure $\mathbf{\Lambda}_a$ to capture recurrent audio patterns ε^a. For ease of notation, we use a generic notation to represent the HMM structure which is identical for body motion and audio streams.

In the second stage, we perform a joint analysis of body motion-audio patterns and extract recurrent co-occurring patterns. This joint correlation analysis will be based on the co-occurrence matrix obtained from the co-occurring body motion-audio events.

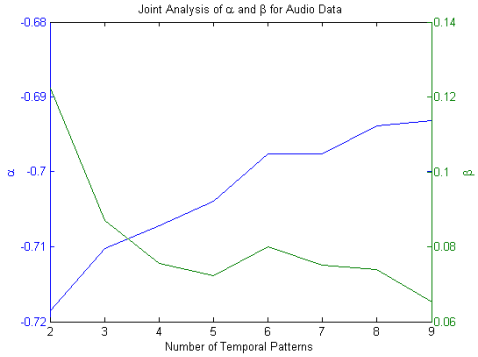

Fig. 3. Results of iterative approach for selection of M for the audio data.

4. RESULTS

Figure 2 shows the plots obtained for α and β measures of different body segments. For video, M is set as 3 which is in the vicinity of the intersection of the normalized α and β measures for separate body motion patterns. Hence, our HMMs for body motion pattern analysis consist of 3 branches each. On the other hand, Figure 3

Table 1. Co-occurrence matrix for Left Arm-Right Arm events in percentages.

	$LArm_a$	$LArm_b$	$LArm_c$
$RArm_a$	95.65	0	4.35
$RArm_b$	0	100	0
$RArm_c$	16.67	8.33	75

Table 2. Co-occurrence matrix for Left Leg-Right Leg events in percentages.

	$LLeg_a$	$LLeg_b$	$LLeg_c$
$RLeg_a$	100	0	0
$RLeg_b$	0	100	0
$RLeg_c$	0	0	100

Table 3. Co-occurrence matrix for Left Arm-Left Leg events in percentages.

	$LLeg_a$	$LLeg_b$	$LLeg_c$
$LArm_a$	94.6	2.7	2.7
$LArm_b$	0	100	0
$LArm_c$	0	0	100

Table 4. Co-occurrence matrix for Right Arm-Right Leg events in percentages.

	$RLeg_a$	$RLeg_b$	$RLeg_c$
$RArm_a$	93.33	3.335	3.335
$RArm_b$	0	100	0
$RArm_c$	0	0	100

shows us that $M = 6$ in the vicinity of the intersection of the normalized α and β measures for the analysis of audio data.

Table 1 demonstrates the co-occurrence percentages between the left arm and the right arm motion patterns obtained as a result of our first stage analysis. Each row in the table displays the co-occurrence rates of different left arm motion patterns with right arm motion patterns over the whole video. According to this co-occurrence matrix, the left arm motion pattern L_a, L_b and L_c highly co-occurs with R_a, R_b and R_c, respectively. The dance figures related with both arm are labeled with same labels for similar figures where label a represents raising the arms up and then lowering them down, b occurs as holding the arms above the shoulder and c is observed as swinging arms forward and backward below shoulder.

Table 2 demonstrates the co-occurrence percentages between the left leg and right leg motion patterns obtained as a result of our first stage analysis. Similarly we can see that left and right arm are highly correlated and labels for similar figures are the same. Label a represents the act of standing at the same place with little bumps of legs, b occurs as pulling the legs up with big steps and c is observed as walking slowly. We can see from Table 3 that left leg and left arm has highly correlated patterns that co-occurs frequently. Nevertheless, we observe in Table 4 that right leg and right arm has highly correlated patterns that co-occurs frequently.

As a result of second stage analysis we investigated the correlation between body motion patterns and audio patterns. Table 5 gives the co-occurrence percentages of right leg and audio data patterns. Some motion patterns are highly correlated with audio patterns for instance $RArm_c$ highly co-occurs with audio pattern A_a where A_f is co-occurred with a small percentages with the same pattern.

5. CONCLUSIONS AND FUTURE WORK

The co-occurrence tables tells us that arms are jointly correlated, legs are jointly correlated and arms and legs are correlated jointly, as well. The temporal patterns of correlated visual motion and audio should prove useful for synthetic agents and/or robots to learn dance figures from audio.

For the future work, the set of Euler angles for each joint can be used as the feature set instead of the displacements, which will provide more robustness in calculation of torso rotation and translation compensation. In addition to MFCCs, other spectral properties such as ralloff, spectral centroid, spectral flux and zero crossing can be used to investigate separate beats or rhythm information of the audio data.

6. REFERENCES

[1] R. Plankers and P. Fua, "Tracking and modeling people in video sequences," *Computer Vision and Image Understanding*, vol. 81, no. 3, March 2001.

[2] C. Bregler and J. Malik, "Tracking people with twists and exponential maps," in *CVPR '98: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC, USA, 1998, p. 8, IEEE Computer Society.

[3] Michael Isard and Andrew Blake, "A mixed-state condensation tracker with automatic model-switching," in *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, Washington, DC, USA, 1998, p. 107, IEEE Computer Society.

[4] C. Kit and Y. Wilks, "Unsupervised learning of word boundary with description length gain," 1999.

[5] F. Ofli, Y. Demir, Y. Yemez, E. Erzin, and M.T. Tekalp, "Multicamera audio-visual analysis of dance figures," .

[6] A. Blake, B. North, and M. Isard, "Learning multi-class dynamics," 1998.

[7] Aphrodite Galata, Neil Johnson, and David Hogg, "Learning variable length markov models of behaviour," 2001.

[8] Tian-Shu Wang, Nan-Ning Zheng, Yan Li, Ying-Qing Xu, and Heung-Yung Shum, "Learning kernel-based hmms for dynamic

Table 5. Co-occurrence matrix for Left-Arm and audio patterns in percentages.

	A_a	A_b	A_c	A_d	A_e	A_f
$RArm_a$	10.64	25.53	19.86	12.06	9.22	26.69
$RArm_b$	21.13	19.01	24.29	11.97	6.69	16.90
$RArm_c$	38.71	10.11	2.81	4.93	8.45	0.35

sequence synthesis," *Graph. Models*, vol. 65, no. 4, pp. 206–221, 2003.

[9] M.E. Sargin, E. Erzin, Y. Yemez, A.M. Tekalp, A.T. Erdem, C. Erdem, and M. Ozkan, "Prosody-driven head-gesture animation," in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing: ICASSP 2007*.