MULTICAMERA AUDIO-VISUAL ANALYSIS OF DANCE FIGURES

*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 an automated system for multicamera motion capture and audio-visual analysis of dance figures. The multiview video of a dancing actor is acquired using 8 synchronized cameras. The motion capture technique is based on 3D tracking of the markers attached to the person's body in the scene, using stereo color information without need for an explicit 3D model. 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 first stage of multimodal analysis, we perform Hidden Markov Model (HMM) based unsupervised temporal segmentation of the audio and body motion features, separately, to determine the recurrent elementary audio and body motion patterns. Then in the second stage, we investigate the correlation of body motion patterns with audio patterns, that can be used for estimation and synthesis of realistic audio-driven body animation.

1. INTRODUCTION

Motion has played an important role in computer vision research since the very beginning and is becoming more and more central as multiple view environments are being introduced into several areas of this research field. One of these areas is devoted to the study of humans, e.g., face and facial expression recognition, gesture recognition, whole-body tracking and gait recognition, or in the more general sense, complete analysis of human activities. Nevertheless, the study of human motion is of interest to a number of disciplines including psychology, kinesiology, choreography, computer graphics and human-computer interaction as well.

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). Aggarwal and Cai review the research progress on human motion analysis in [1] in detail and Gavrila provides an in-depth survey in [2].

Marker-based systems rely on the contrast of the markers with the background to capture their motion. One can use active capture systems, such as LED markers that pulse in synchronization with the cameras' digital shutters, or passive systems, such as using strongly retro-reflective markers along with an illumination source co-located with each camera. These methods however can not acquire and capture the shape and texture properties of the subject, which could also give supplementary information about location of feature points. Hence, [3] proposes a motion capture algorithm based on the use of simple color-markers, aiming at a visually guided and more controllable 3D animation system. On the other hand, in [4], a vision-based full-body estimation and interaction system that uses a marker-less method is presented. It first extracts 2D blob features, and then estimates the 3D full-body parameters. Ricquebourg and Bouthemy in [5] develop a method to track the apparent contours of a moving articulated structure, avoiding the use of 3D models.

In this work we present an automated system for multicamera motion capture and audio-visual analysis of dance figures. The proposed motion capture technique is based on 3D tracking of the markers attached to the person's body in the scene without need for an explicit 3D model. We fit a generic 3D skeleton model to detect and track markers. We make use of the multistereo correspondence information from multiple cameras to obtain 3D positions of the markers. This provides us with a set of 3D point locations over time that expresses the alignment of the markers in 3D world. We employ Kalman filtering for smoothing out the observations and predicting the next target locations of the points in that point cloud in a similar fashion explained in [6]. The resulting set of 3D points are then used to analyze the correlation between the audio patterns and body motion patterns according to [7].

2. MULTICAMERA MOTION CAPTURE

The motion capture process involves tracking a number of markers attached to the subject's body as observed from multiple cameras and extraction of the corresponding motion features. Figure 1 demonstrates our setting for this scenario. Markers in each video frame are tracked making use of their chrominance information. The 3D position of each marker at each frame is then determined via triangulation based on the observed projections of the markers on each camera's image plane.

2.1. Initialization

Markers on the subject are manually labeled in the first frame for all camera views. We change the color space from RGB to YCrCb which gives flexibility over intensity variations in the frames of a video as well as among the videos captured by cameras at different views. We assume that the distributions of Cr and Cb channel intensity values belonging to marker regions are Gaussian. Thus, we calculate the mean, μ, and the covariance, $\mathbf{\Sigma}$, over each marker region (a pixel neighborhood around the labeled point), where $\mu = [\mu_{Cr}, \mu_{Cb}]^T$ and $\mathbf{\Sigma} = (\mathbf{c} - \mu)(\mathbf{c} - \mu)^T$, $\mathbf{c}$ being $[c_{Cr}, c_{Cb}]^T$.

Let M be the number of markers on the subject and $\mathbf{W}$ be the

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

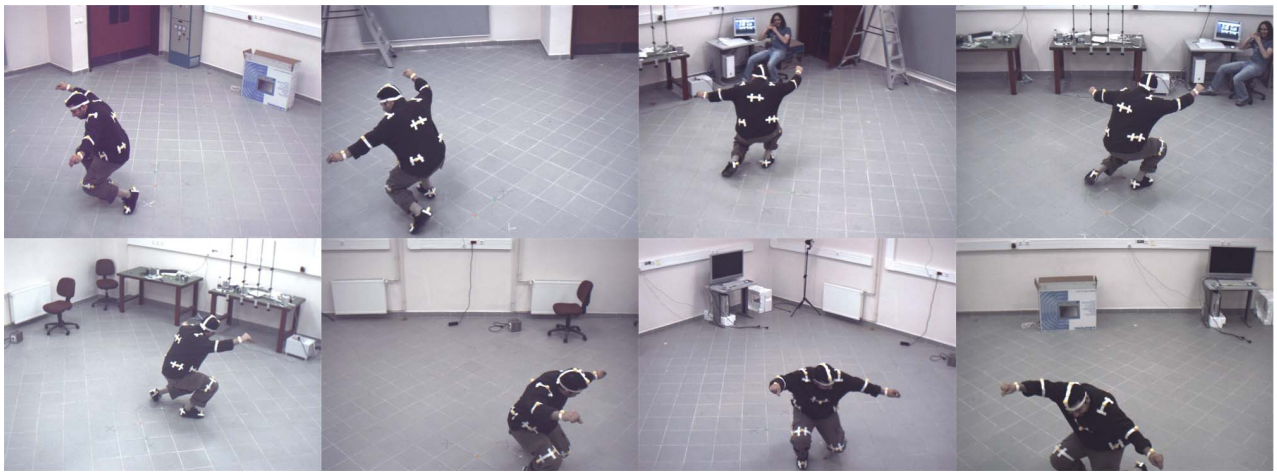

Fig. 1. Dance scene captured by the 8-camera system available at Koç University. Markers are attached at or around the joints of the body.

set of search windows, where $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_M]$ such that each window $\mathbf{w}_m$ is centered around the location, $[x_m, y_m]^T$, of the corresponding marker. The set $\mathbf{W}$ is used to track markers over frames. Thus the center of each search window, $\mathbf{w}_m$, is initialized as the point manually labeled in the first frame and specifies the current position of the marker.

2.2. Tracking

To track the marker positions through the incoming frames, we use the Mahalanobis distance from $\mathbf{c}$ to $(\mu, \mathbf{\Sigma})$ where $\mathbf{c}$ is a vector containing Cr and Cb channel intensity values $[c_{Cr}, c_{Cb}]^T$ of a point $\mathbf{x}_n \in \mathbf{w}_m$. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$ be the set of candidate pixels for which the chrominance distance is less than a certain threshold. If the number of these candidate pixels, N, is larger than a predefined value, then we label that marker as visible in the current camera view and update its position as the mean of the points in $\mathbf{X}$ for the current camera view. The same process is repeated for all marker points in all camera views. Hence, we have the visibility information of each marker from each camera, and for those that are visible, we have the list of 2D positions of the markers on that specific camera image plane.

Once we scan the current scene from all cameras and obtain the visibility information for all markers, we start calculating the 3D positions of the markers by back-projecting the set of 2D points which are visible in respective cameras, using triangulation method. Theoretically, it is sufficient to see a marker at least from two cameras to be able to compute its position in 3D world. If a marker is not visible at least from two cameras, then its current 3D position is estimated from the information in the previous frame.

The 3D positions of markers are tracked over 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 visible 2D points in respective camera image planes constitute the observations for this filter. This filtering operation has two purposes:

- to smooth out the measurements for marker locations in the current frame,
- to estimate the location of each marker in the next frame and to update the positioning of each search window, $\mathbf{w}_m$, on the corresponding image plane accordingly.

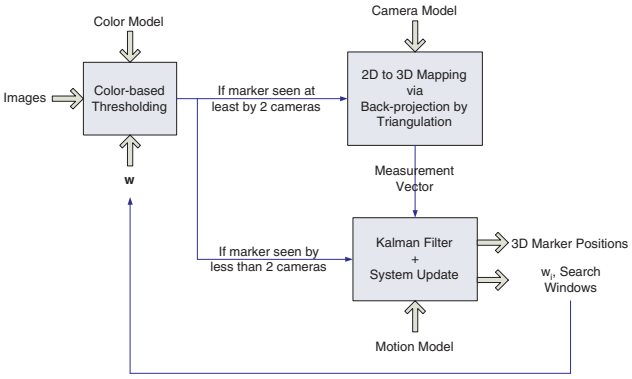

Fig. 2. Block diagram of the proposed tracking system.

Figure 2 summarizes the overall system. Having updated the list of 3D marker positions for the current frame and estimated the location of the search windows for the next frame, we move on to the next frame and search the marker positions within the new search windows. This algorithm is repeated for the whole video. The list of 3D marker positions over frames constitutes the body motion feature vector that will be used later in the animation process.

3. AUDIO-VISUAL DANCE ANALYSIS

In this section, we describe a two-step analysis framework based on unsupervised temporal segmentation. The first step aims to separately extract elementary body motion and audio patterns, and the second step determines a correlation model between these body motion and audio patterns.

3.1. Audio Features

One can consider the act of dancing as 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, which eventually shapes the movements of the body while dancing. Hence, our

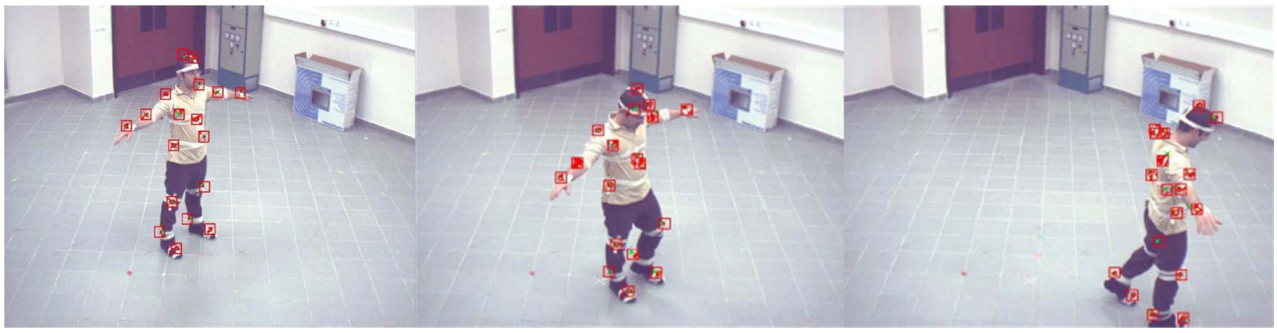

Fig. 3. Tracking of markers for a single view.

audio features are composed of MFCCs.

3.2. Multimodal Analysis

The first stage analysis defines recurrent elementary body motion and audio patterns separately using unsupervised temporal clustering over individual feature streams. The body motion and audio feature streams $\mathbf{F}^b$ and $\mathbf{F}^a$ are separately used to train two HMM structures $\mathbf{\Lambda}_b$ and $\mathbf{\Lambda}_a$, which capture recurrent body motion segments ε^b and audio segments 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. The HMM structure $\mathbf{\Lambda}$, which is used for unsupervised temporal segmentation, has M parallel branches and N states. The states labeled as s_s and s_e are non-emitting start and end states of the parallel HMM structure. 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 $\mathbf{F}^a$. While mapping the body motion and audio features to discrete frame labels, the mismatch between the frame rates of body motion and audio is eliminated by downsampling the frame rate of audio label stream to the rate of body motion label stream.

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

4. RESULTS

We have conducted experiments on a synchronously captured audio-visual data of a dancing person. The dance video is 3 minutes and 15 seconds long with a rate of 30 frames per second. We calculate the mean values and covariance matrices of Cr and Cb channels to build a Gaussian model for each marker and center our search windows around the manually labeled points in the first frame. Figure 3 demonstrates the performance of our tracking scheme after initialization in the first frame.

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. Selecting a small N may hamper modeling long term statistics for each branch of the parallel HMM. The extreme case $N = 1$ reduces to K-Means unsupervised clustering. The number of states in each branch of the body motion HMM model $\mathbf{\Lambda}_b$ is selected to be $N_{\Lambda_b} = 10$, assuming that minimum motion pattern duration is $\frac{1}{3}$ sec (10 frames). Note that, body motion patterns longer than 10 frames can be modeled with the self-state transitions in the HMM structure. 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, M, the number of temporal patterns. Since the number of body motion and audio patterns is dancer and database dependent, we propose an iterative approach for selection of M. For varying values of M, we check two fitness measures. The first fitness measure is the probability of model match, which

increases with the increasing number of temporal patterns. Consequently, the second fitness measure, which is the average statistical separation between two similar temporal patterns, increases with the decreasing number of temporal patterns.

The first fitness measure α, which is inversely related to in-class variance, is defined as the frame average of the log-probability of model match,

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

The α measure is expected to saturate with increasing number of parallel branches in $\mathbf{\Lambda}$, since the training database is expected to contain limited number of temporal patterns. However, small variations within temporal patterns are also expected, hence the number of branches M, which saturates α measure, can be more than the actual number of temporal patterns in the training corpus. In order to make a better estimate of M, the second fitness measure β is considered as the average statistical separation between two similar temporal patterns, and it is defined as,

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

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

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

While M is increasing, the HMM branch models λ_{m_l} and $\lambda_{m_l^*}$ are expected to be similar, which decreases the β measure. Therefore, the total number of temporal patterns, M, can be selected by jointly maximizing the α and β measures.

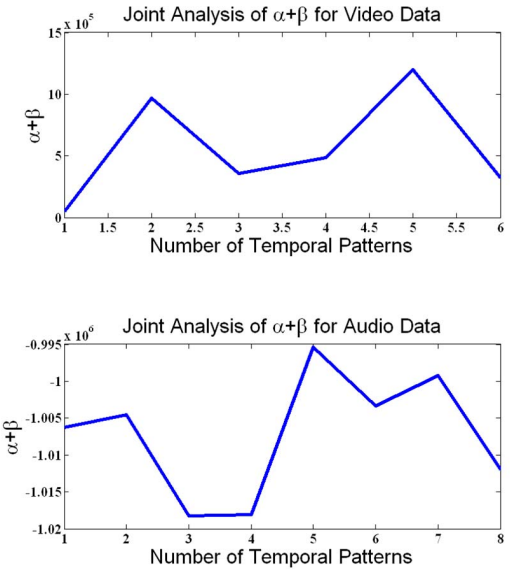

Fig. 4. Results of iterative approach for selection of M.

Figure 4 shows us that $M = 5$ maximizes α and β measures jointly. Hence, our HMM models for body and audio pattern analysis consists of 5 branches each.

Table 1 demonstrates the co-occurrence relation between the body motion and audio patterns obtained as a result of our first stage analysis. Each row in the table displays the co-occurrence percentages of different audio patterns with body motion patterns over the whole video. According to this co-occurrence matrix, the body motion pattern V_e is the most repetitive one in our audio-visual data. Nevertheless, when we look at the co-occurrence relation of the first audio pattern, i.e. A_a, we see that it is also highly correlated with the body motion patterns V_a. On the other hand, A_a never co-occurs with the body motion patterns V_c and V_d. The audio-visual sequences for each body motion patterns are available online [8].

Table 1. Co-occurrence matrix for body motion-audio events.

	V_a	V_b	V_c	V_d	V_e
A_a	40.43	8.51	0.00	0.00	51.06
A_b	5.49	12.09	13.19	6.59	62.63
A_c	10.99	2.20	0.00	4.95	81.86
A_d	0.00	2.94	0.00	2.94	94.12
A_e	22.22	8.55	28.21	4.27	36.75

5. CONCLUSIONS

Results of our analysis indicate that certain motion patterns are highly correlated with the audio channel. The temporal patterns of correlated visual motion and audio should prove useful for synthetic agents and/or robots to learn dance figures from audio.

6. REFERENCES

[1] J. K. Aggarwal and Q. Cai, "Human motion analysis: A review," *Computer Vision and Image Understanding: CVIU*, vol. 73, no. 3, pp. 428–440, 1999.

[2] D. M. Gavrila, "The visual analysis of human movement: A survey," *Computer Vision and Image Understanding: CVIU*, vol. 73, no. 1, pp. 82–98, 1999.

[3] S. Yonemoto, A. Matsumoto, D. Arita, and R.-I. Taniguchi, "A real-time motion capture system with multiple camera fusion," in *Proc. IEEE Int. Conf. on Image Analysis and Processing: ICIAP*, 1999, pp. 600–605.

[4] C.R. Wren, A. Azarbayejani, T. Darrell, and A.P. Pentland, "Pfinder: Real-time tracking of the human body," *IEEE Trans. on Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 780–785, 1997.

[5] Y. Ricquebourg and P. Bouthemy, "Real-time tracking of moving persons by exploiting spatio-temporal image slices," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 797–808, 2000.

[6] D. Comaniciu and V. Ramesh, "Mean shift and optimal prediction for efficient object tracking," in *Proc. IEEE Int. Conf. on Image Processing*, 2000, vol. 3, pp. 70–73.

[7] 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 (accepted to be published)*.

[8] "Audio-visual sequences for body motion patterns," are available at `http://mvgl.ku.edu.tr/audiovisual/`.