Learn2Dance: Learning Statistical Music-to-Dance Mappings for Choreography Synthesis

Ferda Ofli, *Member, IEEE*, Engin Erzin, *Senior Member, IEEE*, Yücel Yemez, *Member, IEEE*, and A. Murat Tekalp, *Fellow, IEEE*

***Abstract*—We propose a novel framework for learning many-to-many statistical mappings from musical measures to dance figures towards generating plausible music-driven dance choreographies. We obtain music-to-dance mappings through use of four statistical models: 1) musical measure models, representing a many-to-one relation, each of which associates different melody patterns to a given dance figure via a hidden Markov model (HMM); 2) exchangeable figures model, which captures the diversity in a dance performance through a one-to-many relation, extracted by unsupervised clustering of musical measure segments based on melodic similarity; 3) figure transition model, which captures the intrinsic dependencies of dance figure sequences via an n-gram model; 4) dance figure models, capturing the variations in the way particular dance figures are performed, by modeling the motion trajectory of each dance figure via an HMM. Based on the first three of these statistical mappings, we define a discrete HMM and synthesize alternative dance figure sequences by employing a modified Viterbi algorithm. The motion parameters of the dance figures in the synthesized choreography are then computed using the dance figure models. Finally, the generated motion parameters are animated synchronously with the musical audio using a 3-D character model. Objective and subjective evaluation results demonstrate that the proposed framework is able to produce compelling music-driven choreographies.**

***Index Terms*—Automatic dance choreography creation, multimodal dance modeling, music-driven dance performance synthesis and animation, music-to-dance mapping, musical measure clustering.**

I. Introduction

CHOREOGRAPHY is the art of arranging dance movements for performance. Choreographers tailor sequences of body movements to music in order to embody or express ideas and emotions in the form of a dance performance. Therefore, dance is closely bound to music in its structural course, artistic expression, and interpretation. Specifically, the rhythm and expression of body movements in a dance performance are in synchrony with those of the music, and hence, the metric orders in the course of music and dance structure coincide, as Reynolds states in [1]. In order to successfully establish the contextual bond as well as the structural synchrony between dance motion and the accompanying music, choreographers tend to thoughtfully design dance motion sequences for a given piece of music by utilizing a repertoire of choreographies. Based on this common practice of choreographers, our goal in this study is to build a framework for automatic creation of dance choreographies in synchrony with the accompanying music; as if they were arranged by a choreographer, through learning many-to-many statistical mappings from music to dance. We note that the term *choreography* generally refers to spatial formation (circle, line, square, couples, etc.), plastic aspects of movement (types of steps, gestures, posture, grasps, etc.), and progression in space (floor patterns), whereas in this study, we use the term *choreography* in the sense of composition, i.e., the arrangement of the dance motion sequence.

Manuscript received December 07, 2010; revised August 25, 2011 and November 17, 2011; accepted December 12, 2011. Date of publication December 23, 2011; date of current version May 11, 2012. This work was supported by TUBITAK under project EEEAG-106E201 and COST2102 action. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Daniel Gatica-Perez.

F. Ofli is with the Tele-Immersion Group, Electrical Engineering and Computer Sciences Department, College of Engineering, University of California at Berkeley, Berkeley, CA 94720 USA (e-mail: fofli@eecs.berkeley.edu).

E. Erzin, Y. Yemez, and A. M. Tekalp are with the Multimedia, Vision and Graphics Laboratory, College of Engineering, Koç University, Sariyer, Istanbul 34450, Turkey (e-mail: eerzin@ku.edu.tr; yyemez@ku.edu.tr; mtekalp@ku.edu.tr).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2011.2181492

A. Related Work

Music-driven dance animation schemes require, as a first step, structural analysis of the accompanying music signal, which includes beat and tempo tracking, measure analysis, and rhythm and melody detection. There exists extensive research in the literature on structural music analysis. Gao and Lee [2] for instance propose an adaptive learning approach to analyze music tempo and beat based on maximum a posteriori (MAP) estimation. Ellis [3] describes a dynamic programming solution for beat tracking by finding the best-scoring set of beat times that reflect the estimated global tempo of music. An extensive evaluation of audio beat tracking and music tempo extraction algorithms, which were included in MIREX'06, can be found in [4]. There are also some recent studies on the open problem of automatic musical meter detection [5], [6]. In the last decade, chromatic scale features have become popular in musical audio analysis, especially in music information retrieval, since introduced by Fujishima [7]. Lee and Slaney [8] describe a method for automatic chord recognition from audio using hidden Markov models (HMMs) through supervised learning over chroma features. Ellis and Poliner [9] propose a cross-correlation based cover song identification system with chroma features and dynamic programming beat tracking. In a very recent work, Kim *et al.* [10] calculate the second order statistics to form dynamic chroma feature vectors in modeling harmony structures for classical music opus identification.

Human body motion analysis/synthesis, as a unimodal problem, has also been extensively studied in the literature in many different contexts. Bregler *et al.* [11] for example describe

a body motion recognition approach that incorporates low-level probabilistic constraints extracted from image sequences of articulated gestures into high-level manifold and HMM-based representations. In order to synthesize data-driven body motion, Arikan and Forsyth [12], and Kovar *et al.* [13] propose motion graphs representing allowable transitions between poses, to identify a sequence of smoothly transiting motion segments. Li *et al.* [14] segment body motions into textons, each of which is modeled by a linear dynamical system, in order to synthesize human body motion in a manner statistically similar to the original motion capture data by considering the likelihood of switching from one texton to the next. Brand and Hertzmann [15] study motion "style" transfer problem, which involves intensive motion feature analysis and learning motion patterns via HMMs from a highly varied set of motion capture sequences. Min *et al.* [16] present a generative human motion model for synthesis of personalized human motion styles by constructing a multilinear motion model that provides explicit parametrized representation of human motion in terms of "style" and "identity" factors. Ruiz and Vachon [17] specifically work on dance body motion, and perform analysis of dance figures in a chain of simple steps using HMMs to perform automatic recognition of basic movements in the contemporary dance.

A parallel track of literature can be found in the domain of speech-driven gesture synthesis and animation. The earliest works in this domain study speaker lip animation [18], [19]. In [18], Bregler *et al.* use morphing of mouth regions to re-sync the existing footage to a new soundtrack. Chen shows in [19] that lip reading a speaker yields higher speech recognition rates and provides better synchronization of speech with lip movements for more natural lip animations. Later a large body of work extends in the direction of synthesizing facial expressions along with lip movements to create more natural face animations [20]–[22]. Most of these studies adopt variations of hidden Markov models to represent the relationship between speech and facial gestures. The most recent studies in this domain aim at synthesizing not only facial gestures but also head, hand, and other body gestures for creating more realistic speaker animations [23]–[25]. For instance, Sargin *et al.* develop a framework for joint analysis of prosody and head gestures using parallel branch HMM structures to synthesize prosody-driven head gesture animations [23]. Levine *et al.* introduce *gesture controllers* for animating the body language of avatars controlled online with the prosody of the input speech by training a specialized conditional random field [25].

Designing a music-driven automatic dance animation system, on the other hand, is a relatively more recent problem involving several open research challenges. There is actually little work in the literature on multimodal dance analysis and synthesis, and most of the existing studies focus solely on the aspect of synchronization between a musical piece and the corresponding dance animation. Cardle *et al.* [26] for instance synchronize motion to music by locally modifying motion parameters using perceptual music cues, whereas Lee and Lee [27] employ dynamic programming to modify timing of both music and motion via time-scaling the music and time-warping the motion. Synchronization-based methods cannot however (actually do not aim to) generate new dance motion sequences. In this sense, the works in [28]–[31] present more elaborate dance analysis and synthesis schemes, all of which follow basically the same similarity-based framework: They first investigate rhythmical and/or emotional similarities between the audio segments of a given input music signal and the available dance motion segments, and then, based on these similarities, synthesize an optimal motion sequence on a motion transition graph using dynamic programming.

In our earlier work [32], we have addressed the statistical learning problem in a multimodal dance analysis scheme by building a correlation model between music and dance. The correlation model was based upon the confusion matrix of co-occurring motion and music patterns extracted via unsupervised temporal segmentation, and hence, was not complex enough to handle realistic scenarios. Later in [33], we have described an automatic music-driven dance animation scheme based on supervised modeling of music and dance figures. However the considered dance scenario was very simplistic, where a dance performance was assumed to have only a single dance figure to be synchronized with the musical beat. In this current paper, we propose a complete framework, based on [34], for modeling, analysis, annotation, and synthesis of multimodal dance performances, which can handle complex and realistic scenarios. Specifically, we focus on learning statistical mappings, which are in general many-to-many, between musical measure patterns and dance figure patterns for music-driven dance choreography animation.

B. Contributions

An open challenge in music-driven dance animation is due to the fact that, for most dance categories, such as ballroom and folk dances, the relationship between music and dance primitives usually exhibits a many-to-many mapping pattern. As discussed in the related work section, the previous methods proposed in the literature for music-driven dance animation, whether similarity-based [28]–[30] or synchronization-based [26], [27], do not address this challenge. They are all deterministic methods and do not involve any true dance learning process (other than building motion transition graphs), therefore cannot capture the many-to-many relationship existing between music and dance, producing always a single optimal motion sequence given the same input music signal. In this paper we address this open challenge by modeling the many-to-many characteristics via a statistical framework. In this respect, our primary contributions are 1) choreography analysis: automatic learning of many-to-many mapping patterns from a collection of dance performances in a multimodal statistical framework, and 2) choreography synthesis: automatic synthesis of alternative dance choreographies that are coherent to a given music signal, using these many-to-many mapping patterns.

For choreography analysis, we introduce two statistical models: one capturing a many-to-one and the other capturing a one-to-many mapping from musical primitives to dance primitives. The former model learns different melody patterns associated with each dance primitive. The latter model learns the group of candidate dance primitives that can be replaced with one another without causing an artifact in the choreography. To further consolidate the coherence and the quality of

the synthesized dance choreographies, we introduce a third model to capture the intrinsic dependencies of dance primitives and to preserve the implicit structure existing in the continuum of dance motion. Combining the aforementioned three models, we present a modified Viterbi algorithm to generate coherent and enriched sequences of dance primitives for plausible choreography synthesis.

The organization of the paper is as follows: Section II first gives an overview of our music-driven dance animation system, and then describes briefly the feature extraction modules. We present the proposed multimodal choreography analysis and synthesis framework, hence our primary contribution, in Section III. The problem of character animation for visualization of synthesized dance choreographies is addressed in Section IV. Section V presents the experiments and results, and finally, Section VI gives concluding remarks and discusses possible applications of the proposed framework.

II. System Overview and Feature Extraction

Our music-driven dance animation scheme is musical measure based, hence we regard musical *measures* as the music primitives. A *measure* is the smallest compositional unit of music that corresponds to a time segment, which is defined as the number of beats in a given duration. We define a *dance figure* as the dance motion trajectory corresponding to a single measure segment. *Dance figures* are taken as the dance primitives.

The overall system, as depicted in Fig. 1, comprises of three parts: analysis, synthesis, and animation. Audiovisual data preparation and feature extraction modules are common to both analysis and synthesis parts. An audiovisual dance database can be pictured as a collection of measures and dance figures aligned in two parallel streams: music stream and dance stream. Fig. 2 illustrates a sample music-dance stream extracted from our folk dance database. In the data preparation module, the input music stream is segmented by an expert into its units, i.e., musical measures. We use m_t to denote the measure segment at frame t. Measure segment boundaries are then used by the expert to define the motion units, i.e., dance figures. We use d_t to denote the dance figure segment corresponding to measure at frame t. The expert also assigns each dance figure d_t a figure label l_j to indicate the type of the dance motion. The collection of l_j forms the set of candidate dance figures, i.e., $\mathcal{L} = \{l_j | j = 1, \ldots, N\}$, where N is the number of distinct dance figure labels that exist in the audiovisual dance database. The resulting sequence of dance figure labels l_j is regarded as the original (reference) choreography, i.e., $\mathbf{r} = \{r_t\}_{t=1}^{t=T}$, where $r_t \in \mathcal{L}$ and T is the number of musical measure segments. The feature extraction modules compute the dance motion features $\mathbf{F}^{d_t}$ and music chroma features $\mathbf{F}^{m_t}$ for each d_t and m_t, respectively.

We assume that the relation between music and dance primitives in a dance performance has a many-to-many mapping pattern. That is, a particular dance primitive (dance figure) can be accompanied by different music primitives (measures) in a dance performance. Conversely, a particular musical measure can correspond to different dance figures. Our choreography analysis and synthesis framework respects the many-to-many

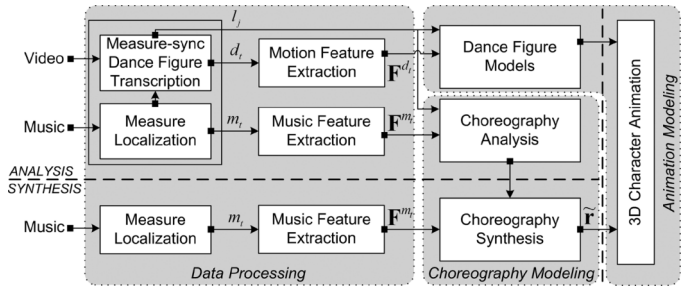

Fig. 1. Block diagram of the overall multimodal dance performance analysis-synthesis framework.

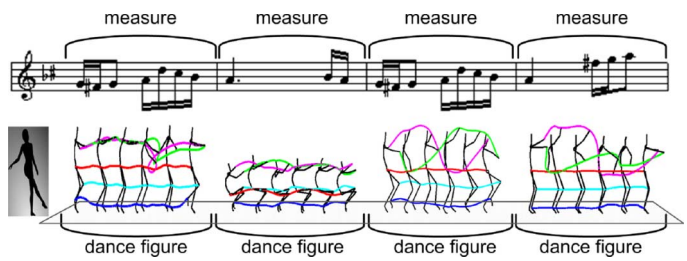

Fig. 2. Audiovisual dance database is a collection of dance figure-musical measure pairs. Recall that a *measure* is the smallest compositional unit of music that corresponds to a time segment, which is defined as the number of beats in a given duration. On the other hand, we define a *dance figure* as the dance motion trajectory corresponding to a single measure segment. Hence, by definition, the boundaries of the dance figure segments coincide with the boundaries of the musical measure segments, which is in conformity with Reynolds' work [1].

nature of the relationship between music and dance primitives by learning two separate statistical models: one capturing a many-to-one and the other capturing a one-to-many mapping from musical measures to dance figures. The former model learns different melody patterns associated with each dance figure. For this purpose, music chroma features $\mathbf{F}^{m_t}$ are used to train a hidden Markov model h_j^m for each dance figure label l_j to create the set of *musical measure models* $\mathcal{H}^m$. The latter model, i.e., the model capturing a one-to-many relation from musical measures to dance figures, learns the group of candidate dance figures that can be replaced with one another without causing an artifact in the dance performance (choreography). We call such a model as *exchangeable figures model* $\mathcal{X}$. Music chroma features $\mathbf{F}^{m_t}$ are used to cluster measure segments m_t according to the harmonic similarity between different measure segments. Based on these measure clusters, we determine the group of dance figures that are accompanied by the musical measures with similar harmonic content. We then create the exchangeable figures model $\mathcal{X}$ based on such dance figure groups. While the former model is designed to keep the underlying correlations between musical measures and dance figures as intact as possible, the latter model is useful for allowing acceptable (or desirable) variations in the dance choreography by offering various possibilities in the choice of dance figures that reflect the diversity in a dance performance (choreography). To further consolidate the coherence and quality of the synthesized dance choreography, we introduce a third model, i.e., the *figure transition model* $\mathcal{F}$, to capture the intrinsic dependencies of the dance figures and to preserve the implicit structure existing in the continuum of dance motion. The figure transition model

basically appraises the figure-to-figure transition relations by computing n-gram probabilities from the audiovisual dance database. The choreography synthesis makes use of these three models, namely, $\mathcal{H}^m$, $\mathcal{F}$, and $\mathcal{X}$, to determine the output dance figure sequence $\tilde{\mathbf{r}}$ (i.e., choreography), taking music chroma features as input, which are extracted from a test music signal. Here, $\tilde{\mathbf{r}} = \{\tilde{r}_t\}_{t=1}^{t=T}$, where $\tilde{r}_t \in \mathcal{L}$ and T is the number of musical measure segments. Specifically, the choreography synthesis module employs a modified Viterbi decoding on a discrete HMM, which is constructed by the musical measure models, $\mathcal{H}^m$, and the figure transition model, $\mathcal{F}$, to determine the sequence of dance figures $\tilde{\mathbf{r}}$ subject to the exchangeable figures model $\mathcal{X}$.

In the analysis part of the animation model, the dance motion features $\mathbf{F}^{d_t}$ are used to train a hidden Markov model h_j^d for each dance figure label l_j to construct the set of dance figure models $\mathcal{H}^d$. Eventually, the body posture parameters corresponding to each dance figure in the synthesized choreography $\tilde{\mathbf{r}}$ are generated using the dance figure models $\mathcal{H}^d$ to animate a 3-D character.

A. Music Feature Extraction

Unlike speech, music consists of a sequence of tones whose frequencies are defined. Moreover, musical melody is a rhythmical succession of single tones in different patterns. In this study, we model the melodic pattern in each measure segment with tone-related features using temporal statistical models, i.e., HMMs. We extract tone-related chroma features to characterize the melodic/harmonic content of music. In order to represent the chroma scale, we project the entire spectrum onto 12 bins corresponding to the 12 distinct semi-tones of the musical octave. Theoretically, the frequency of the kth note in the nth octave is defined as $f_k^n = f_0^0 2^{n+k/12}$, where the pitch of the C0 note is $f_0^0 = 16.35$ Hz based on Shepard's helix model in [35] and $n, k \in \mathrm{Z}$, $k = 0, \ldots, 11$. In this study, we extract the chroma features of 60 semi-tones for $n = 4, \ldots, 8$ (over 5 octaves from the C4 note to the B8 note).

We extract chroma features similar to the well-known mel-frequency cepstral coefficient (MFCC) computation [36] by applying cepstral analysis to the semitone spectral energies. Hence our chroma features capture information of fluctuations of semitone spectral energies. We center the triangular energy windows at the locations of the semi-tone frequencies, f_k^n, at different octaves for $k = 0, \ldots, 11$ and $n = 4, \ldots, 8$. Then, we compute the first 12 DCT coefficients of the logarithmic semitone spectral energy vector, that constitute the chromatic scale cepstral coefficient (CSCC) feature set, $\mathbf{f}^m[n]$, for the music frame n. We also compute the first and second time derivatives of these 12 CSCC features, using the following regression formula:

$$\Delta \mathbf{f}^m[n] = \frac{\sum_{r=-2}^{2} r \mathbf{f}^m[n+r]}{\sum_{r=-2}^{2} r^2}. \tag{1}$$

The music feature vector, $\mathbf{f}_\Delta^m[n]$, is then formed by including the first and second time derivatives:

$$\mathbf{f}_\Delta^m[n] = [\mathbf{f}^m[n]^T \ \Delta \mathbf{f}^m[n]^T \ \Delta^2 \mathbf{f}^m[n]^T]^T. \tag{2}$$

Each $\mathbf{F}^{m_t}$, therefore, corresponds to the sequence of music feature vectors $\mathbf{f}_\Delta^m[n]$ that fall into the measure segment m_t. Specifically, $\mathbf{F}^{m_t}$ is a matrix of CSCC features in the form

$$\mathbf{F}^{m_t} = [\mathbf{f}_\Delta^m[1] \ \mathbf{f}_\Delta^m[2] \ \ldots \ \mathbf{f}_\Delta^m[N_{m_t}]] \tag{3}$$

where N_{m_t} is the number of audio frames in measure segment m_t.

B. Motion Feature Extraction

We acquire multiview recordings of a dancing actor for each dance figure in the audiovisual dance database using 8 synchronized cameras. We then employ a motion capture technique for tracking the 3-D positions of the joints of the body based on the markers' 2-D projections on each camera's image plane, using the color information of the markers [33]. The resulting set of 3-D points are used to fit a skeleton structure to the 3-D motion capture data. Through this skeleton structure, we solve the necessary inverse kinematics equations and calculate accurately the set of Euler angles for each joint in its local frame as well as the global translation and rotation of the skeleton structure for the motion trajectory defined by the input 3-D motion capture data. We prefer joint angles as our dance motion features due to their widespread usage in human body motion analysis-synthesis and 3-D character animation literature. We compute 66 angular values associated with 27 key joints of the body as well as 6 values for the global rotation and translation of the body, which leads to a dance motion feature vector $\mathbf{f}^d[n]$ of dimension 72 for each dance motion frame n. However, angular features are generally discontinuous at boundary values due to their 2π-periodic nature and this situation causes a problem in training statistical models to capture the temporal dynamics of a sequence of angular features. Therefore, instead of using the static set of Euler angles $\mathbf{f}^d[n]$, we use their first and second differences computed with following difference equation:

$$\Delta \mathbf{f}^d[n] = \frac{1}{2}(\mathbf{f}^d[n+1] - \mathbf{f}^d[n-1]) \tag{4}$$

where the resulting discontinuities are eliminated by the following conditional update:

$$\Delta \mathbf{f}^d[n] = \begin{cases} \Delta \mathbf{f}^d[n] - 180, & \text{if } \Delta \mathbf{f}^d[n] > 90 \\ \Delta \mathbf{f}^d[n] + 180, & \text{if } \Delta \mathbf{f}^d[n] < -90 \\ \Delta \mathbf{f}^d[n], & \text{otherwise.} \end{cases} \tag{5}$$

Then the 44-dimensional dynamic motion feature vector is formed as

$$\mathbf{f}_\Delta^d[n] = [\Delta \mathbf{f}^d[n]^T \ \Delta^2 \mathbf{f}^d[n]^T]^T. \tag{6}$$

Hence, each $\mathbf{F}^{d_t}$ is a sequence of motion feature vectors $\mathbf{f}_\Delta^d[n]$ that fall into dance figure segment d_t while training temporal models of motion trajectories associated with each dance figure label l_j. That is, $\mathbf{F}^{d_t}$ is a matrix of body motion feature values in the form

$$\mathbf{F}^{d_t} = [\mathbf{f}_\Delta^d[1] \ \mathbf{f}_\Delta^d[2] \ \ldots \ \mathbf{f}_\Delta^d[N_{d_t}]] \tag{7}$$

where N_{d_t} is the number of dance motion frames within dance motion segment d_t. We also calculate the *mean trajectory* for each dance figure label l_j, namely $\boldsymbol{\mu}_j$, by calculating for each

motion feature an average value over all instances (realizations) of the dance figures labeled as l_j. These *mean trajectories* ($\boldsymbol{\mu}_j$) are required later in choreography animation since each dance figure model h_j^d capture only the temporal dynamics of the first and second differences of the Euler angles of the key joints associated with the dance figure label l_j.

III. Choreography Modeling

In this section we present the proposed choreography analysis-synthesis framework. We first describe our choreography analysis procedure which involves statistical modeling of music-to-choreography mapping. We then explain how this statistical modeling is used for choreography synthesis.

A. Multimodal Choreography Analysis

We perform choreography analysis through three statistical models, which together define a many-to-many mapping from musical measures to dance figures. These choreography models are: 1) musical measure models $\mathcal{H}^m$, which capture many-to-one mappings from musical measures to dance figures using HMMs; 2) exchangeable figures model $\mathcal{X}$, which captures one-to-many mappings from musical measures to dance figures, and hence, represents the subjective nature of the dance choreography with possibilities in the choice of dance figures and in their organization; and 3) figure transition model $\mathcal{F}$, which captures the intrinsic dependencies of dance figures. We note that these three choreography models constitute the "*choreography analysis*" block in Fig. 1.

1) Musical Measure Models ($\mathcal{H}^m$): In a dance performance, musical measures that correspond to the same dance figure may exhibit variations and are usually a collection of different melodic patterns. That is, different melodic patterns can accompany the same dance figure, displaying a many-to-one mapping relation from musical measures to dance figures. We capture this many-to-one mapping by employing HMMs to identify and model the melodic patterns corresponding to each dance figure. Specifically, we train an HMM h_j^m over the collection of measures co-occurring with the dance figure l_j, using the musical measure CSCC features, $\mathbf{F}^{m_t}$. Hence, we train an HMM for each dance figure in the dance performance. We define left-to-right HMM structures h_j^m with $a_{ik}^j \neq 0$ for $k = i, i+1, i+2$, where a_{ik}^j is the transition probability from state q_i^j to state q_k^j. The transitions from state q_i^j to q_{i+2}^j account for the differences in measure durations. Emission distributions of the chroma-based music features are modeled by Gaussian mixture density functions with diagonal covariance matrices in each state of h_j^m. The use of Gaussian mixture density in h_j^m enables us to capture different melodic patterns that correspond to a particular dance figure. We denote the collection of musical measure models as $\mathcal{H}^m$, i.e., $\mathcal{H}^m = \{h_j^m | j = 1, \ldots, N\}$. Musical measure models $\mathcal{H}^m$ provide a tool to capture the many-to-one part of the many-to-many musical measure to dance figure mapping problem.

2) Exchangeable Figures Model ($\mathcal{X}$): In a dance performance, it is possible that several distinct dance figures can be performed equally well along with a particular musical measure pattern, exhibiting a one-to-many mapping relation from musical measures to dance figures [1]. To represent this one-to-many mapping relation, we introduce the notion of *exchangeable figure groups*, each containing a collection of dance figures that can be replaced with one another without causing an artifact in a dance performance. To learn exchangeable figure groups, we cluster the measure segments in each musical piece with respect to their melodic similarities. The melodic similarity s_{ij} between two different measure segments m_i and m_j is computed as the local match score obtained from dynamic time warping (DTW) [37] of the chroma-based feature matrices $\mathbf{F}^{m_i}$ and $\mathbf{F}^{m_j}$, corresponding to m_i and m_j, respectively, in the musical piece S_k. Then, based on the melodic similarity scores between pairs of musical measure segments in S_k, we form an affinity matrix $\mathbf{Y}_k = \left(y_{ij}^k\right)_{i,j=1,\ldots,N}$, where $y_{ij}^k = \exp(-s_{ij})$ if $i \neq j$, and $y_{ii}^k = 0$. Finally, we apply the spectral clustering algorithm described in [38] over $\mathbf{Y}_k$ to cluster the measure segments in S_k. The spectral clustering algorithm in [38] assumes that the number of clusters is known a priori and employs k-means clustering algorithm [39]. Since we do not know the number of clusters a priori, we measure the "quality" of the partition in the resulting clusters using the internal indexes, *silhouettes* [40], to determine the appropriate number of clusters. The silhouette value for each point is a measure of how similar that point is to the points in its own cluster compared to the points in the other clusters, and ranges from -1 to $+1$. Averaging over all the silhouette values, we compute the overall quality of the clustering for a range of cluster numbers and pick the one that results in the highest silhouette value.

We perform separate clustering for each musical piece S_k in order to increase the accuracy of musical measure clustering since similar measure patterns are likely to occur in the same musical piece rather than spread among different musical pieces. Once we obtain clusters of measures in all musical pieces, we can then use all of the measure clusters in all musical pieces to determine the exchangeable figures group $\mathcal{G}_j$ for each dance figure l_j by collecting the dance figure labels that co-appear with l_j in any of the resulting clusters. Note that a particular dance figure can appear in more than one musical piece (see also Fig. 5). Based on the exchangeable figure groups $\mathcal{G}_j$, we define the exchangeable figures model x_j as an indicator random variable:

$$x_j(i) = \mathrm{I}(l_i) = \begin{cases} 1, & \text{if } l_i \in \mathcal{G}_j \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

where $\mathcal{G}_j$ is the exchangeable figure group associated with the dance figure l_j. The collection of x_j for all dance figure labels in $\mathcal{L}$ gives us the exchangeable figures model $\mathcal{X}$.

The notion of exchangeable figures is the key to reflect the subjective nature of the dance choreography with possibilities in the choice of dance figures and their organization throughout the choreography estimation process. The use of exchangeable figures model allows us to create a different artistic dance performance content each time we estimate a dance choreography.

3) Figure Transition Model ($\mathcal{F}$): The figure transition model is built to capture the intrinsic dependencies of the dance figure sequences within the context of dance choreographies. The intrinsic dependencies of the choreography are defined with figure-to-figure transition probabilities. The figure-to-figure transition probability density functions

are modeled in n-gram language models, where the probability of the dance figure l_j at d_t given the dance figure sequence $i_1, i_2, \ldots, i_{n-1}$ at d_{t-1}, d_{t-2}, $\ldots$, d_{t-n+1}, i.e., $P(d_t = l_j | d_{t-1} = i_1, \ldots, d_{t-n+1} = i_{n-1})$, defines the n-gram dance language model. This model provides a number of rules that specify the structure of a dance choreography. For instance, a dance figure that never appears after a particular sequence of $n-1$ dance figures in the training video does not appear in the synthesized choreography either. We can also enforce a dance figure to always follow a particular sequence of $n-1$ dance figures if it is also the case in the training video with the help of the n-gram dance language model.

B. Multimodal Choreography Synthesis

We formulate the choreography synthesis problem as estimating a dance figure sequence from a sequence of musical measures. The core of the choreography synthesis is defined as a Viterbi decoding process on a discrete HMM, which is constructed by the musical measure models, $\mathcal{H}^m$, and the figure transition model, $\mathcal{F}$. Furthermore, the exchangeable figures model, $\mathcal{X}$, is utilized to introduce acceptable variations into the Viterbi decoding process to enrich the synthesized choreography. Based on this Viterbi decoding process, we define three separate choreography synthesis scenarios: 1) *single best path*, 2) *likely path*, and 3) *exchangeable path*, which are explained in detail in the following subsections. We note that all these choreography synthesis scenarios are multimodal, that is, they depend on the joint statistical models of music and choreography, and they have different refinements and contributions to enrich the choreography synthesis.

Besides these three synthesis scenarios, we investigate two other reference (baseline) choreography synthesis scenarios: one using only the musical measure models $\mathcal{H}^m$ to map each measure segment in the test musical piece to a dance figure label (which we refer to as *acoustic-only* choreography), and another one using only the figure transition model $\mathcal{F}$ (which we refer to as *figure-only* choreography). The *acoustic-only* choreography corresponds to a synthesis scenario in which only the correlations between musical measures and dance figure labels are taken into account, but the correlations between consecutive figures are ignored. In contrast to the *acoustic-only* choreography, the *figure-only* choreography scenario predicts the dance figure for the next measure segment only according to figure-to-figure transition probabilities, which are modeled as bigram probabilities of $\mathcal{F}$, by discarding the correlations between musical measures and dance figures. Note that the *figure-only* synthesis can be regarded as a synchronization-based technique discussed in Section I-A, such as the ones proposed in [26] and [27]. In *figure-only* synthesis, the dance figure sequence is generated randomly by only respecting figure-to-figure transition probabilities to ensure visual continuity of the resulting character animation. The dance figure sequences resulting from these two baseline scenarios constitute reference choreographies that help us comparatively assess the contributions of our choreography analysis-synthesis framework.

1) Single Best Path Synthesis: We construct a discrete HMM, $\mathcal{C}$, using the musical measure models, $\mathcal{H}^m$, and the figure transition model, $\mathcal{F}$. In the figure transition model $\mathcal{F}$, the figure-to-figure transition probability distributions are computed with bigram models. This choice of model is due to the scale of choreography database that we use in training, and a much larger database can be used to model higher order n-gram dance language models. The discrete HMM, $\mathcal{C} = (\mathcal{A}, \mathcal{B}, \pi)$, is defined with the following parameters:

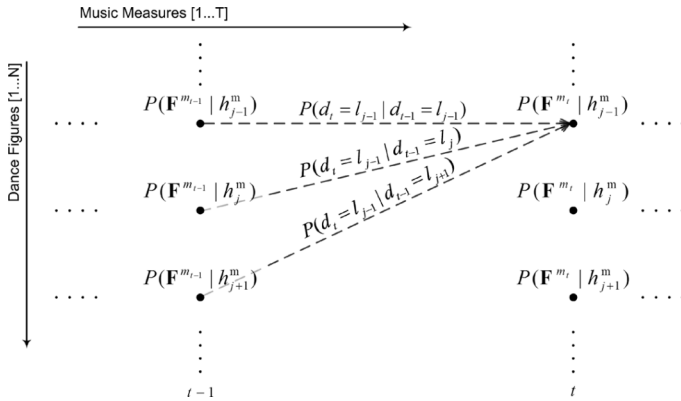

Fig. 3. Lattice structure $\mathbf{M}$ of the discrete HMM $\mathcal{C}$.

- T is the number of time frames (measure segments). For each time frame (measure), the choreography synthesis process outputs exactly one dance figure label. Recall that we denote the individual dance figures as d_t and individual measures as m_t for $t = 1, \ldots, T$.
- N is the number of distinct dance figure labels, i.e., l_j, where $j = 1, \ldots, N$. Dance figure labels are the outputs of the process being modeled.
- $\mathcal{A} = \{a_{ij}\}$ is the dance figure transition probability distribution with elements defined as

$$a_{ij} = P(d_t = l_j | d_{t-1} = l_i), \quad i, j = 1, \ldots, N \tag{9}$$

  where the elements, a_{ij}, are the bigram probabilities from the figure transition model $\mathcal{F}$ and they satisfy $\sum_{j=1}^{N} a_{ij} = 1$.
- $\mathcal{B} = \{b_t(j)\}$ is the dance figure emission distribution for measure m_t. Elements of $\mathcal{B}$ are defined using the musical measure models $\mathcal{H}^m$ as

$$b_t(j) = P(\mathbf{F}^{m_t} | h_j^m), \;\; j = 1, \ldots, N; \; t = 1, \ldots, T. \tag{10}$$

- $\pi = \{\pi_i\}$ is the initial dance figure distribution, where

$$\pi_i = P(d_1 = l_i), \;\; i = 1, \ldots, N. \tag{11}$$

The discrete HMM $\mathcal{C}$ constructs a lattice structure, say $\mathbf{M}$, as given in Fig. 3. The proposed choreography synthesis can be formulated as finding a path through the lattice $\mathbf{M}$. Assuming a uniform initial figure distribution π, the *single best path* synthesis scenario decodes the Viterbi path along the lattice $\mathbf{M}$ to estimate the synthesized figure sequence $\tilde{\mathbf{r}}$. The Viterbi algorithm for finding the *single best path* synthesis can be summarized as follows:

1) Initialization:

$$\begin{aligned} \phi_j(1) &= b_1(j) \qquad j = 1, \ldots, N \\ \psi_j(1) &= 0 \qquad\quad j = 1, \ldots, N. \end{aligned} \tag{12}$$

2) Recursion: For $t = 2, \ldots, T$

$$\phi_j(t) = \max_i\{a_{ij}\phi_i(t-1)\}b_t(j) \;\; j = 1, \ldots, N \quad (13)$$

$$\psi_j(t) = \operatorname*{argmax}_i\{a_{ij}\phi_i(t-1)\} \;\; j = 1, \ldots, N. \quad (14)$$

3) Termination:

$$\tilde{r}_T = \operatorname*{argmax}_i\{\phi_i(T)\}. \quad (15)$$

4) Path (dance figure sequence) backtracking:

$$\tilde{r}_t = \psi_{\tilde{r}_{t+1}}(t+1) \;\; t = T-1, T-2, \ldots, 1. \quad (16)$$

Here $\phi_j(t)$ represents the partial likelihood score of performing the dance figure l_j at frame t, and $\psi_j(t)$ is used to keep track of the best path retrieving the dance figure sequence. The path $\tilde{\mathbf{r}} = \{\tilde{r}_t\}_{t=1}^{t=T}$ decodes the resulting dance figure label sequence as the desired output choreography. Note that the resulting dance choreography $\tilde{\mathbf{r}}$ is unique for the *single best path* synthesis scenario since it is the Viterbi path along the lattice $\mathbf{M}$.

2) Likely Path Synthesis: In the second synthesis scenario, we find a *likely path* along $\mathbf{M}$ in which we follow one of the *likely* partial paths in lieu of following the partial path that has the highest partial likelihood score at each time frame. The *likely path* synthesis is expected to create variations along the *single best path* synthesis, which results in an enriched set of synthesized choreography sequences with high likelihood scores.

We modify the recursion step of the Viterbi algorithm to define the *likely path* synthesis:

Recursion: For $t = 2, \ldots, T$

$$i^1 = \operatorname*{argmax}_i\{a_{ij}\phi_i(t-1)\}$$

$$i^2 = \operatorname*{argmax}_{i-\{i^1\}}\{a_{ij}\phi_i(t-1)\} \quad (17)$$

$$\psi_j(t) = U(i^1, i^2) \;\; j = 1, \ldots, N \quad (18)$$

$$\phi_j(t) = \phi_{\psi_j(t)}(t-1)a_{\psi_j(t)j}b_t(j) \;\; j = 1, \ldots, N \quad (19)$$

where i^1 and i^2 are the figure indices with the top two partial path scores and $U(i^1, i^2)$ returns randomly one of the arguments with uniform distribution. Note that i^1 corresponds to $\psi_j(t)$ in (14). The likely path scenario is expected to synthesize different dance choreographies since it propagates randomly among top two ranking transitions at each time frame. This intricately introduces variation into the choreography synthesis process.

3) Exchangeable Path Synthesis: In this scenario, we find an *exchangeable path* by letting the exchangeable figures model $\mathcal{X}$ replace and update the *single best path*. Unlike the *likely path* synthesis, the *exchangeable path* scenario introduces random variations to the *single best path* that respect one-to-many mappings from musical measures to dance figures as defined in $\mathcal{X}$.

The *exchangeable path* synthesis is implemented with the following procedure:

1) Compute the *single best path* synthesis for a given musical measure sequence and set the measure segment index $t = 1$.
2) The figure l_{i^1} at measure segment t is replaced with another figure l_{i^*} from its exchangeable figure group $\mathcal{G}_{i^1}$

$$i^* = U'(\{i\}|l_i \in \mathcal{G}_{i^1}) \quad (20)$$

where $U'()$ returns randomly one of the arguments according to the distribution of acoustic scores $P(\mathbf{F}^{m_t}|h_i^m)$ of the dance figures $l_i \in \mathcal{G}_{i^1}$.

3) The rest of the figure sequence, $\tilde{r}_{t+1}, \ldots, \tilde{r}_T$, is updated by determining a new *single best path* using the Viterbi algorithm.
4) The steps 2) and 3) are repeated for measure segments $t = 2, \ldots, T$.

The *exchangeable path* synthesis yields an alternative path by modifying the *single best path* in the context of the exchangeable figures model. Its key difference from the *likely path* is that the collection of the *candidate* dance figures that can replace a particular dance figure in the choreography, say l_j, is constrained with the dance figures for which the exchangeable figures model x_j yields 1. Hence, it is expected to introduce more acceptable variations into the synthesized choreography than the *likely path*.

IV. Animation Modeling

In this section, we address character animation of dance figures to visualize and evaluate the proposed choreography analysis and synthesis framework. First we define a dance figure model, which captures the variations in the way particular dance figures are performed, by modeling the motion trajectory of a dance figure via an HMM. Then we define a character animation system, which generates a sequence of dance motion features from a given sequence of dance figures, so as to animate a 3-D character model.

A. Dance Figure Models ($\mathcal{H}^d$)

The way a dancer performs a particular dance figure may exhibit variations in time in a dance performance. Therefore, it is important to model the temporal statistics of each dance figure to capture the variations in the dance performance. Note that these models will also capture the personalized dance figure patterns of a dancer. We use the set of motion features $\mathbf{F}^{d_t}$ to train an HMM, h_j^d, for each dance figure label l_j to capture the dynamic behavior of the dancing body. Since a dance figure contains typically a well-defined sequence of body movements, we employ a left-to-right HMM structure (i.e., $a_{ik}^j \neq 0$ for $k = i, i+1$, where a_{ik}^j is the transition probability from state q_i^j to state q_k^j in h_j^d) to model each dance figure. Emission distributions of motion parameters are modeled by a Gaussian density function with full covariance matrix in each state of h_j^d. We denote the collection of dance figure models as $\mathcal{H}^d$, i.e., $\mathcal{H}^d = \{h_j^d | j = 1, \ldots, N\}$.

B. Character Animation

The synthesized choreography $\tilde{\mathbf{r}}$ (i.e., $\{\tilde{r}_t\}_{t=1}^{t=T}$) specifies the label sequence of dance figures to be performed with each measure segment whose duration is known beforehand in the proposed framework. The body posture parameters corresponding to each dance figure in the synthesized choreography $\{\tilde{r}_t\}_{t=1}^{t=T}$ are then generated such that they fit to the statistical dance figure models $\mathcal{H}^d$.

To generate body posture parameters using the dance figure model h_j^d for the dance figure l_j, we first determine the number of dance motion frames L required for the given segment duration. Next, we distribute the required number of motion frames L among the states of the dance figure model h_j^d according to the expected state occupancy duration:

$$o_i^j = \frac{1}{1 - a_{ii}^j}, \qquad 1 \leq i \leq P \quad (21)$$

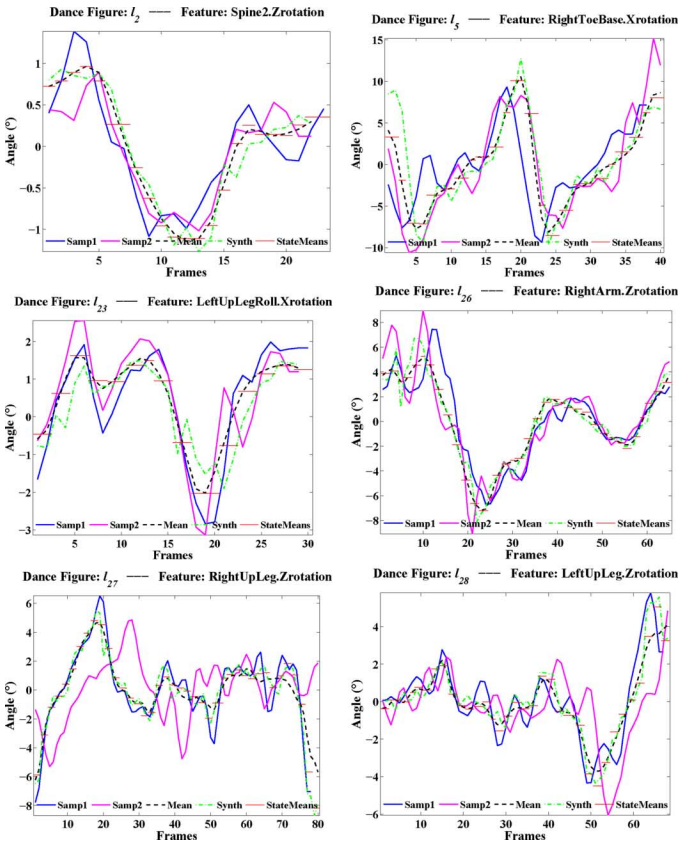

Fig. 4. Plots compare a synthesized trajectory with two sample trajectories, as well as with the *mean trajectory* for different motion features from different dance figures in the database. The expected state durations, all associated with the HMM structure trained for the same dance figures, are also displayed in the plot with horizontal solid lines. The corresponding angular values of each horizontal solid line are the means of the Gaussian distributions, associated with the particular motion feature in each state of the HMM structure trained for the given dance figure.

where o_i^j is the expected duration in state q_i^j, a_{ii}^j is the self-state-transition probability for state q_i^j (assuming $a_{ii}^j \neq 0$), and P is the number of states in h_j^d.

In order to avoid generation of noisy parameters, we first increase the time resolution of the dance motion by oversampling the dance motion model. That is, we generate parameters for a multiple of L, say KL, where K is an integer scale factor. Then, we generate the body motion parameters along the states of h_j^d according to the distribution of KL motion frames to these states, using the corresponding Gaussian distribution at each state. To reverse the effect of oversampling, we perform a downsampling by K that eventually yields smoother state transitions, and hence, more realistic parameter generation that avoid motion jerkiness.

The dance figure models h_j^d are trained over the first and second differences of the Euler angles of the joints, which are defined in Section II-B. Therefore, to obtain the final set of body posture parameters for a dance figure l_j, we simply need to sum the generated first differences with the *mean trajectory* associated with l_j, i.e., $\boldsymbol{\mu}_j$. Each plot in Fig. 4 depicts a synthesized trajectory against two sample trajectories for one of the motion features from a dance figure in the database along with the *mean trajectory* associated with the same motion feature from the same dance figure. The length of the horizontal solid lines represent the expected state durations (in terms of the number of frames), and the corresponding angular values represent the means of the Gaussian distributions, associated with the particular motion feature in each state of the HMM structure trained for the particular dance figure. In these plots, the two sample trajectories exemplify the temporal variations between different realizations of the same dance figure. Deriving from the trained dance figure HMMs, the synthesized dance figure trajectories mimic the underlying temporal dynamics of a given dance figure. Hence, modeling the temporal variations using HMMs for each dance figure allows us to synthesize more realistic and personalized dance motion trajectories.

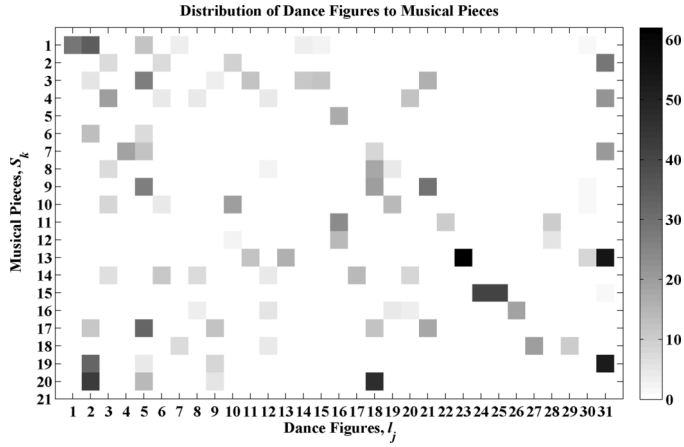

Fig. 5. Distribution of dance figures to musical pieces is visualized using an image plot. Columns of the plot represent the dance figures (l_j) whereas the rows represent the musical pieces (S_k) in the database. Note that l and S are dropped in the figure for clarity of the representation. Consequently, each cell in the plot indicates how many times a particular dance figure is performed with the corresponding musical piece. Cells with different gray values in a column imply that the same dance figure can be accompanied with different musical pieces. Similarly, cells with different gray values in a row imply that different dance figures can be performed with the same musical piece.

After repeating the described procedure for each dance figure in the synthesized choreography, the body posture parameters at the dance figure boundaries are smoothed via cubic interpolation within a Δ-neighborhood of each dance figure boundary in order to generate smoother figure-to-figure transitions.

We note that the use of HMMs for dance figure synthesis provides us with the ability of introducing random variations in the synthesized body motion patterns for each dance figure. These variations make the synthesis results look more natural due to the fact that humans perform slightly varying dance figures at different times for the same dance performance.

V. Experiments and Results

We investigate the effectiveness of our choreography analysis and synthesis framework using the Turkish folk dance, *Kasik.*[1] The *Kasik* database consists of 20 dance performances with 20 different musical pieces with a total duration of 36 min. There are 31 different dance figures (i.e., $N = 31$) and a total of 1258 musical measure segments (i.e., $T = 1258$). Fig. 5 shows the distribution of dance figures to different musical pieces where each column represents a dance figure label l_j and each row represents a musical piece S_k. Hence, entries with different colors in a column indicate that the same figure can be performed with

[1]*Kasik* means *spoon* in English. The dance is named so, since the dancers clap spoons while dancing.

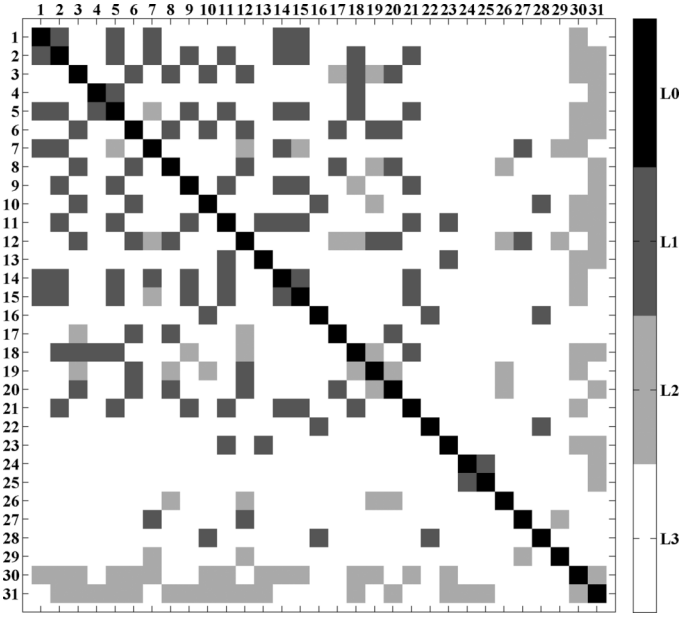

Fig. 6. Matrix plot demonstrates the assessment levels associated with any pair of dance figures in the database. Assessment levels are indicated with different colors. The pairs of dance figures that fall into assessment level $L1$ in a row correspond to the group of exchangeable dance figures for that particular dance figure. For instance, by looking at the first row, one can say that the dance figure l_1 is exchangeable with the dance figures l_2, l_5, l_7, l_{14}, and l_{15}. In other words, $\{l_2, l_5, l_7, l_{14}, l_{15}\}$ is the group of exchangeable figures for l_1, i.e., $\mathcal{G}_1$.

TABLE I
AVERAGE PENALTY SCORES (APS) OF VARIOUS CHOREOGRAPHY SYNTHESIS SCENARIOS

Synthesis Scenario	APS
Single Best Path	0.56
Likely Path	0.91
Exchangeable Path	0.63
Acoustic-Only	0.82
Figure-Only	2.07

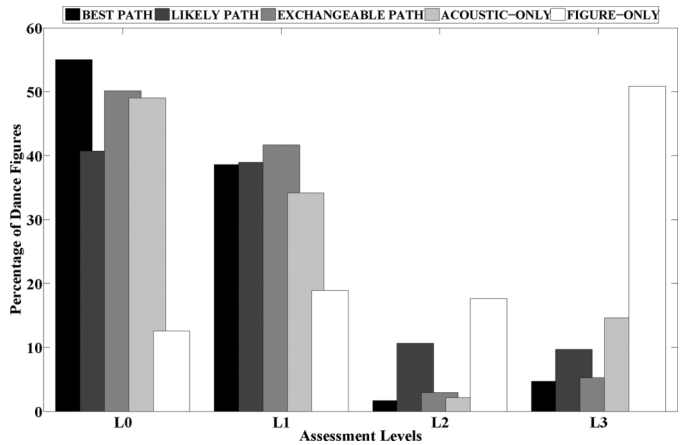

Fig. 7. Percentage of figures that fall into each assessment level for the proposed five different synthesis scenarios.

different melodic patterns whereas entries with different colors in a row indicate that different dance figures can be performed with the same melodic pattern. Therefore, Fig. 5 can be seen as a means to provide evidence for our basic assumption that there is a many-to-many relationship between dance figures and musical measures.

We follow a 5-fold cross-validation procedure in the experimental evaluations. We train musical measure models with four-fifths of the musical audio data in the analysis part and use these musical measure models in the process of choreography estimation for the remaining one-fifth of the musical audio data in the synthesis part. We repeat this procedure five times, each time using different parts of the musical audio data for training and testing. This way, we synthesize a new dance choreography for the entire musical audio data.

A. Objective Evaluation Results

We define the following four assessment levels to evaluate each dance figure label $\tilde{r}_t$ in the synthesized figure sequence $\tilde{\mathbf{r}}$, compared to the respective figure label r_t in the original dance choreography $\mathbf{r}$, assigned by the expert:

- $L0$ (*Exact-match*): $\tilde{r}_t$ is marked as $L0$ if $\tilde{r}_t$ matches r_t.
- $L1$ (*X-match*): $\tilde{r}_t$ is marked as $L1$ if $\tilde{r}_t$ does not match r_t, but it is in r_t's exchangeable figure group $\mathcal{G}_{r_t}$; i.e., $\tilde{r}_t \in \mathcal{G}_{r_t}$.
- $L2$ (*Song-match*): $\tilde{r}_t$ is marked as L_2 if $\tilde{r}_t$ neither matches r_t nor is in $\mathcal{G}_{r_t}$; but, $\tilde{r}_t$ and r_t are performed within the same musical piece; i.e., $\tilde{r}_t, r_t \in S_k$.
- $L3$ (*No-match*): $\tilde{r}_t$ is marked as L_3 if it is not marked as one of $L0$ through $L2$.

Fig. 6 displays all assessment levels associated with any possible pairing of dance figures in a single matrix. Note that the matrix in Fig. 6 defines a distance metric on dance figure pairs by mapping the four assessment levels $L0$ through $L3$ into penalty scores from 0 to 3, respectively. Hence in this distance metric, low penalty scores indicate desirable choreography synthesis results. Average penalty scores are reported to measure the "goodness" (coherence) of the resulting dance choreography.

Recall that we propose three alternative choreography synthesis scenarios together with the two other reference choreography synthesis techniques, as discussed in Section III-B. The average penalty scores of these five choreography synthesis scenarios are given in Table I. Furthermore, the distribution of the number of figures that fall into each assessment level for all synthesis scenarios are given in Fig. 7. The average penalty scores for the reference *acoustic-only* and *figure-only* choreography synthesis scenarios are 0.82 and 2.07, respectively. The high average penalty score of the *figure-only* choreography is mainly due to the randomness in this synthesis technique. Hence, the unimodal learning phase of the *figure-only* choreography synthesis, which takes into account only the figure-to-figure transition probabilities, does not carry sufficient information to automatically generate dance choreographies which are similar to the training *Kasik* database. On the other hand, the *acoustic-only* choreography, which employs the musical measure models $\mathcal{H}^m$ for synthesis, attains a lower average penalty score. We also note that the average dance figure recognition rate using the musical measure models $\mathcal{H}^m$ through five-fold cross validation is obtained as 49.05%. This indicates that the musical measure models learn the correlation between measures and figures. However, the *acoustic-only* choreography synthesis that depends only on the correlation of measures and figures fails to sustain motion continuity at figure transition boundaries. This creates visually unacceptable character animation for the *acoustic-only* choreography.

The proposed *single best path* synthesis scenario refines the *acoustic-only* choreography with the inclusion of the figure transition model $\mathcal{F}$, which defines a bigram model for figure-to-figure transitions. The average penalty score of the *single best path* synthesis scenario is 0.56, which is the smallest average penalty score among all scenarios. This is expected, since the *single best path* synthesis generates the optimal Viterbi path along the multimodal lattice structure $\mathbf{M}$. The *likely path* synthesis introduces variation into the *single best path* synthesis. The average penalty score of the *likely path* synthesis increases to 0.91. This increase is an indication of the variation introduced to the optimal Viterbi path. The *exchangeable path* synthesis refines the *likely path* synthesis by introducing more acceptable random variations to the *single best path* synthesis. Recall that random variations of the *exchangeable path* synthesis depend on the exchangeable figures model $\mathcal{X}$. The average penalty score of the *exchangeable path* synthesis is 0.63, which is an improvement compared to the *likely path* synthesis.

In Fig. 7, we observe that among all the assessment levels, the levels $L0$ and $L1$ are indicators of the diversity of alternative dance figure choreographies, rather than being an error indicator, whereas the assessment levels $L2$ and $L3$ indicate an error in the dance choreography synthesis process. In this context, we observe that only 31% of the *figure-only* choreography and only 83% of the *acoustic-only* choreography fall into the first three assessment levels. On the other hand, using the mapping obtained by our framework increases this ratio to 94% and 92% for the *single best path* and the *exchangeable path* synthesis scenarios, respectively. The percentage drops to 80% for the *likely path* synthesis scenario, yet it is still a high percentage of the entire dance sequence.

B. Subjective Evaluation Results

We performed a subjective A/B comparison test using the music-driven dance animations to measure the opinions of the audience on the coherence of the synthesized dance choreographies with the accompanying music. During the test, the subjects were asked to indicate their preference for each given A/B test pair of synthesized dance animation segments on a scale of $(-2;\ -1;\ 0;\ 1;\ 2)$, where the scale corresponds to *strongly prefer A*, *prefer A*, *no preference*, *prefer B*, and *strongly prefer B*, respectively. We compared dance animation segments from five different choreographies, namely, *original*, *single best path*, *likely path*, *exchangeable path*, and *figure-only* choreographies. We, therefore, had ten possible pairings of different dance choreographies, e.g., *original* versus *single best path*, or *likely path* versus *figure-only*, etc. For each possible pair of choreographies, we used short audio segments from three different musical pieces from the *Kasik* database to synthesize three A/B pairs of dance animation video clips for the respective dance choreographies. This yielded us a total of 30 A/B pairs of dance animation segments. We also included one A/B pair of dance animation segments for pairing each choreography with itself, i.e., *original* versus *original*, etc., in order to test if the subjects were careful enough and show almost *no preference* over five such possible self-pairings of the aforementioned choreographies. As a result, we extracted 35 short segments from the audiovisual database, where each segment was approximately 15 s. We picked at most two non-overlapping segments from each musical piece in order to make a full coverage of the audiovisual database in the subjective A/B comparison test.

TABLE II
SUBJECTIVE A/B PAIR COMPARISON TEST RESULTS

		B				
		O	**SBP**	**LP**	**XP**	**FO**
A	Original (O)	**0.1**	-0.6	**0.1**	**0.7**	-0.2
	Single Best Path (SBP)		**0.2**	**0.2**	**0.7**	-0.2
	Likely Path (LP)			**0.2**	-0.3	-0.7
	Exchangeable Path (XP)				**0.1**	-0.7
	Figure-Only (FO)					0.0

The subjective tests are performed over 18 subjects. The average preference scores for all comparison sets are presented in Table II. Note that the rows and the columns of Table II, respectively, correspond to A and B of the A/B pairs. Also, the average preference scores that tend to favor B are given in bold to ease the visual inspection. The first observation is that the animations for the original choreography and for the choreographies resulting from the proposed three synthesis scenarios (i.e., *single best path*, *likely path*, and *exchangeable path* choreographies) are preferred over the animations for the *figure-only* choreography. We also note that the *likely path* and *exchangeable path* choreography animations are strongly preferred against the *figure-only* choreography animation. This observation is an evidence of the fact that audience is generally appealed by variations in the dance choreography as long as the overall choreography is coherent with the accompanying music. Hence we observe the *likely path* and the *exchangeable path* synthesis as the most preferable scenarios in subjective tests, and they manage to create alternative choreographies that are coherent and appealing to the audience.

C. Discussion

The objective evaluations carried out using the assessment levels in Section V-A indicate that the most successful choreography synthesis scenarios are the *single best path* and the *exchangeable path* scenarios. We note that the *exchangeable path* synthesis as well as the *likely path* can be seen as variations or refinements of the *single best path* synthesis approach. On the other hand, according to the subjective evaluations presented in Section V-B, the most preferable scenarios are the *likely path* and the *exchangeable path*. Hence the *exchangeable path* synthesis, being among the top two according to both objective and subjective evaluations, can be regarded as our best synthesis approach. Note also that the *exchangeable path* synthesis includes all the statistical models defined in Section III.

The other two scenarios (*acoustic-only* and *figure-only*) are mainly used to demonstrate the effectiveness of the musical measure models and the figure transition model, hence they both serve as reference synthesis methods. The *acoustic-only* choreography however depends only on the correlations between measures and figures, and therefore fails to sustain motion continuity at figure-to-figure boundaries, which is indispensable to create realistic animations. Hence we have chosen the *figure-only* synthesis as the baseline method to compare with our best result, i.e., with the *exchangeable path* synthesis,

and prepared a demo video (submitted as supplemental material) that compares side by side the animations resulting from these two synthesis scenarios for two musical pieces in the *Kasik* database.

We have also prepared another video which demonstrates the *likely path*, the *exchangeable path*, and the *single best path* synthesis results with respect to sample original dance performances. The demo video starts with a long excerpt from the original and the synthesized choreographies driven by a musical piece that is available in the *Kasik* database. The long excerpt is followed by several short excerpts from the original and the synthesized choreographies driven by the musical pieces that are available in the *Kasik* database. The demo is concluded with two long excerpts from the synthesized choreographies driven by two musical pieces that are **not** available in the *Kasik* database. Both demo videos are also available online [41].

VI. Conclusions

We have described a novel framework for music-driven dance choreography synthesis and animation. For this purpose, we construct a many-to-many statistical mapping from musical measures to dance figures based on the correlations between dance figures and musical measures as well as the correlations between successive dance figures in terms of figure-to-figure transition probabilities. We then use this mapping to synthesize a music-driven sequence of dance figure labels via a constraint based dynamic programming procedure. With the help of exchangeable figures notion, the proposed framework is able to yield a variety of different dance figure sequences. These output sequences of dance figures can be considered as alternative dance choreographies that are in synchrony with the driving music signal. The subjective evaluation tests indicate that the resulting music-driven dance choreographies are plausible and compelling to the audience. To further evaluate the synthesis results, we have also devised an objective assessment scheme that measures the "goodness" of a synthesized dance choreography with respect to the original choreography.

Although we have demonstrated our framework on a folk dance database, the proposed music-driven dance animation method can also be applied to other dance genres such as ballroom dances, Latin dances and hip hop, as long as the dance performance is musical measure-based (i.e., the metric orders in the course of music and dance structure coincide [1]), and the dance database contains sufficient amount of data to train the statistical models employed in our framework. One possible source of problem in our current framework might be due to the dance genres which do not have a well-defined set of dance figures. Hip hop, which is in fact a highly measure-based dance genre, is a good example of this. In such cases, figure annotation may become a very tedious task due to possibly very large variations in the way a particular movement (supposedly a dance figure) is performed. More importantly, if the dance genre does not have a well-defined set of dance figure, the 3-D motion capture data needed to train the dance figure models (described in Section IV-A) must be carefully prepared, since otherwise joining up individual figures smoothly during character animation can be very difficult and the continuum of the synthesized dance motion may not be guaranteed.

The performance of the proposed framework strongly depends on the quality, the complexity and the size of the audiovisual dance database. For instance, higher-order n-gram models can be integrated in the presence of sufficient training data, to better exploit the intrinsic dependencies of the dance figures. Currently we employ only bigrams (with $n = 2$) to model intrinsic dependencies of the dance figures. This choice of model is mainly due to the scale of the choreography database that we use in training. We also note that, in our experiments, the bigram statistics proved to be sufficient for the current state of the framework and for the particular Turkish folk dance genre, i.e., *Kasik*.

Our music-driven dance animation scheme currently supports only the single dancer scenario; but the framework can be extended to handle multiple dancers as well by learning the correlations between the movements of the dancers through the use of additional statistical models, that would however increase the complexity of the overall learning process. Having multiple dancers will also increase the complexity of the animation since then the spatial positioning of the dancers relative to each other will also have to be taken into account.

The proposed framework currently requires expert input and musical transcription prior to the audiovisual feature extraction and modeling tasks. This tedious pre-processing can be eliminated by introducing automatic measure/dance figure segmentation capability into the framework. However, such automatic segmentation techniques are not yet currently available in the literature, and they seem to remain as open research areas in the near future. In this study, HMM structures are used to model the dance motion trajectories, since they can represent variations among different realizations of dance figures in personalized dance performances. However, one can consider other methods such as *style machines* that will also represent stylistic variations associated with dance figures.

We define a dance figure as the dance motion trajectory corresponding to a single measure segment. The choice of measures as elementary music primitives simplifies the task of statistical modeling and allows us to use the powerful HMM framework for music-to-dance mapping. However this choice can also be seen as a limiting assumption that ignores higher levels of semantics and correlations which might exist in a musical piece such as chorus and verses. This current limitation of our framework could be addressed by using, for example, hierarchical statistical modeling tools and/or higher order n-grams (with $n > 2$). Yet, modeling higher levels of semantics remains as an open challenge for further research.

The proposed framework can trigger interdisciplinary studies with collaboration of dance artists, choreographers, and computer scientists. Certainly, it has the potential of creating appealing applications, such as fast evaluation of dance choreographies, dance tutoring, entertainment, and more importantly digital preservation of folk dance heritage by safeguarding irreplaceable information that tend to perish. As a final remark, we think that the proposed framework can be modified to be used for other multimodal applications such as speech-driven facial expression or body gesture synthesis and animation.

REFERENCES

[1] W. C. Reynolds, "Foundations for the analysis of the structure and form of folk dance: A syllabus," *Yearbook Int. Folk Music Council*, vol. 6, pp. 115–135, 1974.

[2] S. Gao and C.-H. Lee, "An adaptive learning approach to music tempo and beat analysis," in *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing.*, 2004, vol. 4, pp. 237–240.

[3] D. P. W. Ellis, "Beat tracking by dynamic programming," *J. New Music Res.*, vol. 36, no. 1, pp. 51–60, 2007.

[4] M. F. McKinney, D. Moelants, M. E. P. Davies, and A. Klapuri, "Evaluation of audio beat tracking and music tempo extraction algorithms," *J. New Music Res.*, vol. 36, no. 1, pp. 1–16, 2007.

[5] A. Klapuri, A. Eronen, and J. Astola, "Analysis of the meter of acoustic musical signals," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 14, no. 1, pp. 342–355, Jan. 2006.

[6] M. Gainza, "Automatic musical meter detection," in *IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009)*, 19–24, 2009, pp. 329–332.

[7] T. Fujishima, "Realtime chord recognition of musical sound: A system using common lisp music," in *Proc. Int. Computer Music Conf.*, 1999, pp. 464–467.

[8] K. Lee and M. Slaney, "Automatic chord recognition from audio using a supervised HMM trained with audio-from-symbolic data," in *Proc. 1st ACM Workshop Audio and Music Computing Multimedia (AMCMM '06)*, New York, 2006, pp. 11–20.

[9] D. Ellis and G. Poliner, "Identifying 'cover songs' with chroma features and dynamic programming beat tracking," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2007 (ICASSP 2007)*, 15–20, 2007, vol. 4, pp. IV–1429–IV–1432.

[10] S. Kim, P. Georgiou, and S. Narayanan, "A robust harmony structure modeling scheme for classical music opus identification," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing, 2009 (ICASSP 2009)*, 19–24, 2009, pp. 1961–1964.

[11] C. Bregler, S. M. Omohundro, M. Covell, M. Slaney, S. Ahmad, D. A. Forsyth, and J. A. Feldman, "Probabilistic models of verbal and body gestures," in *Computer Vision in Man-Machine Interfaces*. Cambridge, U.K.: Cambridge Univ, Press, 1998, pp. 267–290.

[12] O. Arikan and D. A. Forsyth, "Interactive motion generation from examples," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 483–490, 2002.

[13] L. Kovar, M. Gleicher, and F. Pighin, "Motion graphs," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 473–482, 2002.

[14] Y. Li, T. Wang, and H.-Y. Shum, "Motion texture: A two-level statistical model for character motion synthesis," *ACM Trans. Graph.*, vol. 21, no. 3, pp. 465–472, 2002.

[15] M. Brand and A. Hertzmann, "Style machines," in *Proc. 27th Annu. Conf. Computer Graphics and Interactive Techniques (SIGGRAPH '00)*, New York, 2000, pp. 183–192.

[16] J. Min, H. Liu, and J. Chai, "Synthesis and editing of personalized stylistic human motion," in *Proc. 2010 ACM SIGGRAPH Symp. Interactive 3D Graphics and Games (I3D '10)*, New York, 2010, pp. 39–46.

[17] A. Ruiz and B. Vachon, "Three learning systems in the reconnaissance of basic movements in contemporary dance," in *Proc. 5th Biannu. World Automation Congr., 2002* , 2002, vol. 13, pp. 189–194.

[18] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proc. 24th Annual Conf. Computer Graphics and Interactive Techniques (SIGGRAPH '97)*, New York, 1997, pp. 353–360.

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

[20] M. Brand, "Voice puppetry," in *Proc. 26th Annual Conf. Computer Graphics and Interactive Techniques (SIGGRAPH '99)*, New York, 1999, pp. 21–28.

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

[22] J. Xue, J. Borgstrom, J. Jiang, L. Bernstein, and A. Alwan, "Acoustically-driven talking face synthesis using dynamic Bayesian networks," in *Proc. IEEE Int. Conf. Multimedia and Expo, 2006*, Jul. 2006, pp. 1165–1168.

[23] M. E. Sargin, Y. Yemez, E. Erzin, and A. M. Tekalp, "Analysis of head gesture and prosody patterns for prosody-driven head-gesture animation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1330–1345, Aug. 2008.

[24] M. Neff, M. Kipp, I. Albrecht, and H.-P. Seidel, "Gesture modeling and animation based on a probabilistic re-creation of speaker style," *ACM Trans. Graph.*, vol. 27, pp. 5:1–5:24, Mar. 2008.

[25] S. Levine, P. Krähenbühl, S. Thrun, and V. Koltun, "Gesture controllers," *ACM Trans. Graph.*, vol. 29, pp. 124:1–124:11, Jul. 2010.

[26] M. Cardle, L. Barthe, S. Brooks, and P. Robinson, "Music-driven motion editing: Local motion transformations guided by music analysis," in *Proc. Annu. Eurographics UK Conf.*, 2002, vol. 0, pp. 38–44.

[27] H. C. Lee and I. K. Lee, "Automatic synchronization of background music and motion in computer animation," *Comput. Graph. Forum*, vol. 24,, pp. 353—361,, 2005.

[28] T.-H. Kim, S. I. Park, and S. Y. Shin, "Rhythmic-motion synthesis based on motion-beat analysis," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 392–401, 2003.

[29] G. Alankus, A. A. Bayazit, and O. B. Bayazit, "Automated motion synthesis for dancing characters," *Comput. Animat. Virtual Worlds*, vol. 16, no. 3-4, pp. 259–271, 2005.

[30] T. Shiratori, A. Nakazawa, and K. Ikeuchi, "Dancing-to-music character animation," *Comput. Graph. Forum*, vol. 25, no. 3, pp. 449–458, 2006.

[31] J. W. Kim, H. Fouad, J. L. Sibert, and J. K. Hahn, "Perceptually motivated automatic dance motion generation for music," *Comput. Animat. Virtual Worlds*, vol. 20, no. 2–3, pp. 375–384, 2009.

[32] F. Ofli, Y. Demir, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multicamera audio-visual analysis of dance figures," in *Proc. IEEE Int. Conf. Multimedia and Expo, 2007*, 2007, pp. 1703–1706.

[33] F. Ofli, Y. Demir, E. Erzin, Y. Yemez, A. M. Tekalp, K. Balci, I. Kiziloglu, L. Akarun, C. Canton-Ferrer, J. Tilmanne, E. Bozkurt, and A. Erdem, "An audio-driven dancing avatar," *J. Multimodal User Interfaces*, vol. 2, no. 2, pp. 93–103, Sep. 2008.

[34] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp, "Multi-modal analysis of dance performances for music-driven choreography synthesis," in *Proc. IEEE Int. Conf. Acoustics Speech and Signal Processing (ICASSP), 2010*, 2010, pp. 2466–2469.

[35] R. Shepard, "Circularity in judgements of relative pitch," *J. Acoust. Soc. Amer.*, vol. 36, no. 12, 1964.

[36] S. Davis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 28, no. 4, pp. 357–366, Aug. 1980.

[37] H. Sakoe and S. Chiba, "Dynamic programming algorithm optimization for spoken word recognition," *IEEE Trans. Acoust., Speech, Signal Process.*, vol. 26, no. 1, pp. 43–49, Feb. 1978.

[38] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems*. Cambridge, MA: MIT Press, 2001, vol. 14, pp. 849–856.

[39] J. B. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. 5th Berkeley Symp. Mathematical Statistics and Probability*, 1967, pp. 281–297.

[40] P. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, 1987.

[41] F. Ofli, E. Erzin, Y. Yemez, and A. Tekalp, Music-Driven Dance Choreography Synthesis Demo, 2011. [Online]. Available: http://mvgl.ku.edu.tr/Learn2Dance.

Ferda Ofli (S'07–M'11) received the B.Sc. degrees, both in electrical and electronics engineering and computer engineering, and the Ph.D. degree in electrical engineering from Koç University, Istanbul, Turkey, in 2005 and 2010, respectively.

He is currently a postdoctoral researcher in the Tele-Immersion Group of the University of California at Berkeley, Berkeley, CA. His research interests span the areas of multimedia signal processing, computer vision, pattern recognition, and machine learning. He received the Graduate Studies Excellence award in 2010 for outstanding academic achievement at Koç University.

Engin Erzin (S'88–M'96–SM'06) received the B.Sc., M.Sc., and Ph.D. degrees from the Bilkent University, Ankara, Turkey, in 1990, 1992, and 1995, respectively, all in electrical engineering.

During 1995–1996, he was a postdoctoral fellow in the Signal Compression Laboratory, University of California, Santa Barbara. He joined Lucent Technologies in September 1996, and he was with the Consumer Products for one year as a Member of Technical Staff of the Global Wireless Products Group. From 1997 to 2001, he was with the Speech and Audio Technology Group of the Network Wireless Systems. Since January 2001, he has been with the Koç University, Istanbul, Turkey. His research interests include speech signal processing, audio-visual signal processing, human-computer interaction, and pattern recognition.

Dr. Erzin is serving as an Associate Editor of the IEEE TRANSACTIONS ON AUDIO, SPEECH, AND LANGUAGE PROCESSING (2010–2013).

Yücel Yemez (M'03) received the B.Sc. degree from Middle East Technical University, Ankara, Turkey, in 1989, and the M.Sc. and Ph.D. degrees from Boğaziçi University, Istanbul, Turkey, in 1992 and 1997, respectively, all in electrical engineering.

From 1997 to 2000, he was a postdoctoral researcher in the Image and Signal Processing Department of Télécom Paris (ENST), Paris, France. Currently he is an Associate Professor of the Computer Engineering Department at Koç University, Istanbul. His research is focused on various fields of computer vision and graphics.

A. Murat Tekalp (S'80–M'84–SM'91–F'03) received the M.Sc. and Ph.D. degrees in electrical, computer, and systems engineering from Rensselaer Polytechnic Institute (RPI), Troy, NY, in 1982 and 1984, respectively.

He has been with Eastman Kodak Company, Rochester, NY, from December 1984 to June 1987, and with the University of Rochester from July 1987 to June 2005, where he was promoted to Distinguished University Professor. Since June 2001, he has been a Professor at Koç University, Istanbul, Turkey. His research interests are in the area of digital image and video processing, including video compression and streaming, motion-compensated video filtering for high-resolution, video segmentation, content-based video analysis and summarization, 3DTV/video processing and compression, multicamera surveillance video processing, and protection of digital content. He authored the book *Digital Video Processing* (Englewood Cliffs, NJ: Prentice-Hall, 1995) and holds seven U.S. patents. His group contributed technology to the ISO/IEC MPEG-4 and MPEG- 7 standards.

Dr. Tekalp was named Distinguished Lecturer by the IEEE Signal Processing Society in 1998, and awarded a Fulbright Senior Scholarship in 1999. He received the TUBITAK Science Award (highest scientific award in Turkey) in 2004. He chaired the IEEE Signal Processing Society Technical Committee on Image and Multidimensional Signal Processing (January 1996–December 1997). He served as an Associate Editor for the IEEE TRANSACTIONS ON SIGNAL PROCESSING (1990 to 1992) and the IEEE TRANSACTIONS ON IMAGE PROCESSING (1994 to 1996), and the Kluwer journal *Multidimensional Systems and Signal Processing* (1994 to 2002). He was an Area Editor for *Graphical Models and Image Processing* (1995 to 1998). He was also on the Editorial Board of the Academic Press journal *Visual Communication and Image Representation* (1995 to 2002). He was appointed as the Special Sessions Chair for the 1995 IEEE International Conference on Image Processing, the Technical Program Co-Chair for IEEE ICASSP 2000 in Istanbul, the General Chair of IEEE International Conference on Image Processing (ICIP) in Rochester in 2002, and Technical Program Co-Chair of EUSIPCO 2005 in Antalya, Turkey. He is the Founder and First Chairman of the Rochester Chapter of the IEEE Signal Processing Society. He was elected as the Chair of the Rochester Section of IEEE for 1994 to 1995. At present, he is the Editor-in-Chief of the EURASIP journal *Signal Processing: Image Communication* (Elsevier). He is serving as the Chairman of the Electronics and Informatics Group of the Turkish Science and Technology Foundation (TUBITAK) and as an independent expert to review projects for the European Commission.