MULTI-MODAL ANALYSIS OF DANCE PERFORMANCES FOR MUSIC-DRIVEN CHOREOGRAPHY SYNTHESIS

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

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

ABSTRACT

We propose a framework for modeling, analysis, annotation and synthesis of multi-modal dance performances. We analyze correlations between music features and dance figure labels on training dance videos in order to construct a mapping from music measures (segments) to dance figures towards generating music-driven dance choreographies. We assume that dance figure segment boundaries coincide with music measures (audio boundaries). For each training video, figure segments are manually labeled by an expert to indicate the type of dance motion. Chroma features of each measure are used for music analysis. We model temporal statistics of such chroma features corresponding to each dance figure label to identify different rhythmic patterns for that dance motion. The correlations between dance figures and music measures, as well as, correlations between consecutive dance figures are used to construct a mapping for music-driven dance choreography synthesis. Experimental results demonstrate the success of proposed music-driven choreography synthesis framework.

Index Terms— multimodal dance modeling, music-driven dance choreography synthesis

1. INTRODUCTION

Dance is a collective union of the music and the human body movements. Automatic dance analysis, annotation and synthesis have been studied extensively in the literature with emphasis on human body motion analysis/synthesis and dance music analysis whereas there is relatively little work on the open problem of music-driven automatic dance synthesis as we address in this paper.

One of the early dance notation systems for human body motion, known as Labanotation, defines a data format to record human dance figures with graphical symbols that provides a detailed sequence of changes in human posture during a dance figure [1]. In [2], Li et al. segment body motions into textons, each of which was modeled by a linear dynamic system, 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. In [3], Ruiz and Vachon perform analysis of dance figures in a chain of simple steps using HMMs to perform automatic recognition of basic movements in the contemporary dance.

Dance music analysis in general includes beat and tempo tracking, measure analysis, and rhythm and melody detection. In [4], Gao and Lee propose an adaptive learning approach to analyze music tempo and beat based on maximum a posteriori (MAP) estimation. Ellis 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 [5]. An extensive evaluation of audio beat tracking and music tempo extraction algorithms, which were included in MIREX’06, can be found in [6].

In the context of multi-modal music and dance analysis towards dance motion synthesis, Shiratori et al. propose a method, which automatically detects the musical rhythm with beat and tempo, performs a music rhythm assisted motion segmentation, and classifies motion segments into the primitive motion units [7]. In an earlier work [8], we have addressed the problem of multi-camera audio-visual analysis of dance figures, where a correlation model between body motion and music is extracted by unsupervised temporal segmentation of the recurrent elementary audio and body motion patterns. Later in [9], we have described an automatic music-driven dance animation scheme based on supervised modeling of music and dance figures in a simplified scenario, where a dance performance is assumed to have only a single dance figure which is to be synchronized with the musical beat.

In this present paper, we propose a complete framework for modeling, analysis, annotation and synthesis of multi-modal dance performances, which can handle more complex and realistic scenarios. Specifically, we focus on finding mappings, which are in general many-to-many, between audio patterns and dance figures for music-driven dance choreography synthesis.

2. MUSIC-DRIVEN DANCE SYNTHESIS FRAMEWORK

The overall framework, as depicted in Figure 1, comprises of several blocks that can be grouped into five main tasks: beat extraction and measure localization; audio feature extraction; measure modeling and identification; dance figure labeling and N-gram modeling; multi-modal dance figure estimation.

![Block diagram of the overall multi-modal dance performance analysis-synthesis framework.](image-url)
Audio is processed in two parallel tasks. One of the tasks extracts beat and meter information from audio to determine the measure boundaries whereas the other task extracts features from audio as chroma features. The outputs of this task are used for measure modeling and identification.

On the other hand, video is manually segmented into recurrent dance figures and each dance figure is assigned with a label. The sequence of dance figure labels is then used to calculate the N-gram probabilities so as to capture the underlying figure-to-figure transition structures in the original dance choreography.

For measure modeling and identification, we train HMMs to capture different rhythmic audio patterns that correspond to each dance figure. The trained HMMs are then used in associating each musical measure with a list of likely dance figure ids. The collection of these lists is finally used for dance figure estimation: Given the collection of likely dance figures lists, we estimate a sequence of figure ids based on figure-to-figure transition probability model found by dance figure labeling and N-gram modeling. The following sections explain the aforementioned five main tasks in more detail.

2.1. Beat Extraction and Measure Localization

A musical piece is a collection of measures and a measure is a time segment that is defined as the number of beats in a given duration. Since dance figures are performed in synchrony with musical rhythm, the boundaries of dance figures are expected to match those of the musical measures. For this purpose, we check the accuracy of manually marked dance figure labels by performing measure analysis on music audio. We use one of the recent algorithms proposed by Davies and Plumley in [10] to extract beats and fine-tune measure boundary locations.

2.2. Acoustic Feature Extraction

Chroma features can be used to characterize the melodic or harmonic content of music since they represent musical audio by projecting the entire spectrum onto 12 bins corresponding to the 12 distinct semitones of the musical octave.

We extract chroma features by following an approach which is very similar to the well-known mel-frequency cepstral coefficients (MFCC) calculation. The difference is in how we choose the triangular overlapping windows while calculating the chroma coefficients from the magnitude spectrum of DFT of the audio signal. We basically center the triangular weight windows at the locations of semitone frequencies at different octaves as shown in Figure 2. Then, we take log-average of the harmonics of the calculated semitone coefficients, that gives us 12-bin chroma features.

2.3. Measure Modeling and Identification

In a dance performance, musical measures that correspond to the same dance figure may exhibit variations and are usually a collection of different rhythmic audio patterns. We employ HMMs to identify and model the audio measure patterns corresponding to the dance figures. Each HMM is trained over the collection of measures co-occuring with the same dance figure. In other words, each HMM computes $P(F|a)$, i.e., probability of dance figure $F$ given $a$, acoustic chroma features. Hence, we train as many HMMs as the number of different dance figures that exist in the dance performance. Figure 3 shows the HMM structure we use for training models for the collection of measures.

For measure identification part of this task, we use the trained HMMs to assign figure ids to the sequence of measures extracted from the input music audio. Instead of identifying each measure with the label of the model that gives the best acoustic score, i.e., the highest likelihood probability of the model match, we create a list of model labels with the highest-N acoustic scores. That is, we generate $N$ alternative transcriptions for each audio frame, i.e., musical measure. We then form a lattice, call it $M$, where the vertical dimension represents the dance figures and the horizontal dimension represents the frames of music (i.e., measures). The entries of $M$ are the acoustic scores (i.e., likelihood probabilities) of the corresponding models at the corresponding music frames. This lattice will also be used for dance figure estimation.

2.4. N-Gram Dance Figure Modeling

The N-gram dance figure model provides us with some rules that specify the structure of a dance choreography. For instance, two particular dance figures that never appear in consequence in the training video do not either appear one after the other in the estimated synthesis. We can also enforce a dance figure to always follow a particular figure if it is also the case in the training video with the help of the N-gram model. In this work, we pick $N = 2$ and create a bigram probability matrix, call it $A$, for the input dance figure sequence to capture the dependency relation of the current dance figure with the previous one. More specifically, each entry in $A$, namely $a_{i,j}$, is the probability of performing the figure $F_j$ after the figure $F_i$.

2.5. Multi-modal Dance Figure Estimation

Using the bigram matrix $A$ (computed as explained in Section 2.4) together with the lattice $M$ (constructed as explained in Section 2.3), we estimate an output dance figure sequence by finding a path along $M$ in two different ways. In the first one, we follow the single best path along $M$, i.e., the label sequence that has the maximum total likelihood. In the other one, we follow a path in which we pick the likely figure, i.e., the figure that is randomly selected according to a predefined distribution, at each music frame.

We employ a Viterbi algorithm to traverse through the columns of $M$ using $A$. Recall that an entry in $M$, namely $m_{i,j}$, represents the likelihood of figure $F_j$ being performed at music frame $j$. Let $N$ denote the number of rows in $M$ (which is also the number of different dance figures); $T$ denote the number of columns in $M$ (which is also the total number of measure segments); and $\phi_j(t)$ represent the partial likelihood score of performing dance figure $F_j$ at frame $t$ along a single path that accounts for the highest partial likelihood.
from frame 1 to frame \( t \). This partial likelihood can be computed efficiently using the following recursion:

\[
\psi_j(t) = \max_i \{ \phi_j(t-1)a_{ij} \} m_{jt},
\]

(1)

At time \( t \), each partial likelihood score \( \phi_j(t-1) \) is known for all dance figures \( F_j \), hence Equation 1 can be used to compute \( \phi_j(t) \) thereby extending the partial paths by one music frame. We also define a structure \( \psi_j(t) \) to keep track of the argument which maximizes Equation 1, for each \( j \) and \( t \), in order to retrieve the dance figure sequence. The overall algorithm for finding the single best dance figure sequence can be summarized as follows:

1. Initialization:

\[
\psi_1(1) = m_{j1}, \quad 1 \leq j \leq N
\]

\[
\psi_j(1) = 0, \quad 1 \leq j \leq N
\]

(2)

2. Recursion:

\[
\phi_j(t) = \max_i \{ \phi_i(t-1)a_{ij} \} m_{jt}, \quad 2 \leq t \leq T
\]

\[
\psi_j(t) = \arg\max_i \{ \phi_i(t-1)a_{ij} \}, \quad 2 \leq t \leq T
\]

(3)

3. Termination:

\[
\Phi = \max_i \{ \phi_i(T) \}
\]

\[
\Psi(T) = \arg\max_i \{ \phi_i(T) \}
\]

(4)

4. Path (dance figure sequence) backtracking:

\[
\Psi(t) = \psi_{\Psi(t+1)}(t+1), \quad t = T-1, T-2, \ldots, 1
\]

(5)

Even though this procedure is designed for the first synthesis scenario, i.e., picking the single best path along \( M \), we can easily modify it for the second synthesis scenario, i.e., picking a likely path along \( M \). Instead of picking the maximum in Equation 1, we can randomly pick one of the 'likely' dance figures according to a pre-specified distribution \( P \). It is also necessary to update the recurrence relation for \( \psi_j(t) \) accordingly.

3. EXPERIMENTS AND RESULTS

In this study, we investigate the Turkish folk dance, \emph{kasik}.\(^1\) Our audiovisual database is 36 minutes long and consists of 20 dance performances with 20 different musical pieces. There are 31 different dance figures (i.e., \( N = 31 \)) and a total of 1265 musical measure segments (i.e., \( T = 1265 \)).

Table 1 provides the distribution of dance figures for different musical pieces (denoted by MP\(_1\) through MP\(_{20}\) in the first column), showing the many-to-many nature of the mapping between dance figures and music measures. That is, there are different dance figures performed with the same musical piece and some dance figures are performed with more than one musical piece. More importantly, each row in the third column of Table 1 shows the expert-specified group(s) of figures that are alternatives to one another for the corresponding musical piece. For instance, dance figure a2 can be performed in places where b2 is performed, or vice versa, with the first musical piece and the change of places between these two figures creates a different but still acceptable choreography according to the expert. This information will be useful in evaluating the output of the dance figure estimation task.

We follow 5-fold cross-validation procedure for measure modeling and identification. We create an HMM for each "measure collection" using four fifth of the audio data and use these models to identify the remaining one fifth. We repeat this procedure five times, each time using different parts of the audio data for training and testing. This way, the entire audio data is identified with recognition ids upon which we base our dance figure estimation task as mentioned earlier in Section 2.5.

We define the following five assessment levels to evaluate each figure label \( F_i \) in the synthesized figure sequence compared to the respective figure label \( F_i \) assigned by the expert:

- **L0:** \( F_i \) is marked as L0 if \( F_i \) matches \( F_i \).
- **L1:** \( F_i \) is marked as L1 if \( F_i \) does not match \( F_i \), but it is in one of the expert-specified exchangeable figure groups together with \( F_j \) i.e., \((F_i, F_j) \in H\), where \( H = \{(F_i, F_j) | F_i \neq F_j; F_i \text{ and } F_j \text{ are in one of the expert-specified exchangeable figure groups}\}.\)
- **L2:** \( F_i \) is marked as L2 if \( F_i \) does not match \( F_i \), and it is not in one of the expert-specified exchangeable figure groups together with \( F_i \), either. However, \( F_i \) and \( F_i \) are performed with the same musical piece; i.e., \((F_i, F_i) \in O \setminus H\), where \( O = \{(F_i, F_j) | F_i \neq F_j; F_i \text{ and } F_j \text{ are performed with the same musical piece}\}.\)
- **L3:** \( F_i \) is marked as L3 if \( F_i \) and \( F_i \) should not be performed with the same musical piece, and yet, they are exchanged due to a recognition error because the musical pieces with which they are actually performed have similar rhythmic audio patterns; i.e., \((F_i, F_a) \in R \setminus O\), where \( R = \{(F_i, F_j) | F_i \neq F_j; \text{the entry } (F_i, F_j) \text{ is non-zero in the confusion matrix resulting from measure modeling and identification}\}.\)
- **L4:** \( F_i \) is marked as L4 if it is not marked as one of L0 through L3.

\begin{table}[h]
\centering
\caption{List of figures for each musical piece and the corresponding groups of exchangeable figures.}
\begin{tabular}{|c|c|}
\hline
List of Figures & Exchangeable Figure Groups \\
\hline
MP\(_1\) & \{b2, a2, e2, f1, i1, i2, z2, z1\} \{b2, f2, e2\} \{b2, f1, z1\} \\
MP\(_2\) & \{b4, e4, h2, z2\} \{b4, e4, h2\} \\
MP\(_3\) & \{b2, e2, f3, n1, n2, z1\} \{b2, e2, f3, n1\} \{b2, e2\} \\
MP\(_4\) & \{b4, e4, f2, h5, o, z2\} \{b4, e4, f2, h5, o\} \\
MP\(_5\) & \{b2, e2\} \{b2, e2\} \\
MP\(_6\) & \{b2, e2, n1, z2\} \{b2, e2, n1\} \\
MP\(_7\) & \{b2, e2, f3, n1\} \{b2, e2, f3\} \\
MP\(_8\) & \{b2, e2, f3, n1\} \{b2, e2, f3\} \\
MP\(_9\) & \{b4, e4, h2, n2, o\} \{b4, e4, h2\} \{b4, e4\} \{b4, h2\} \\
\hline
\end{tabular}
\end{table}

\(^1\)Kasik means spoon in English. The dance is named so because the dancers clap spoons while dancing.
In order to visualize the quality of the synthesized choreographies, we created demo videos according to the output dance figure sequences. These demo videos are available online at http://mvgl.ku.edu.tr/bodymotionanalysis/icassp10/.

4. CONCLUSIONS

In this paper, we propose a mapping from music measures to dance figures based on correlations between dance figures and music measures as well as 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. The output sequence of dance figure labels can be considered as a dance choreography that is in synchrony with the driving audio signal. The experimental results show that the proposed framework is successful at creating acceptable alternative dance choreographies.

Possible applications of the proposed framework include: i) an automatic dance tutor that evaluates recorded dance performances of dance students, ii) synthesis of 3D dancing avatars for visual evaluation of synthesized choreographies, and iii) automatic synthesis of dance performances from audio only for on-line games and other entertainment applications, such as ‘Second Life’.

5. REFERENCES