IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received December 29, 2020, accepted January 26, 2021, date of publication February 5, 2021, date of current version February 17, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3057486

A Deep Music Recommendation Method Based on Human Motion Analysis

WENJUAN GONG, (Member, IEEE), AND QINGSHUANG YU
Department of Intelligence Science, College of Computer Science and Technology, China University of Petroleum (East China), Qingdao 266580, China
Corresponding author: Wenjuan Gong (wenjuangong@upc.edu.cn)

ABSTRACT Despite its applications in automatic video editing and automatic music composition, the problem of music recommendation from dance motions has seldom been explored. In order to solve this problem, this work proposes a deep music recommendation algorithm based on dance motion analysis and evaluate it through quantitative measures. For quantitative evaluation, this work implements a LSTM-AE based music recommendation method which learns the correspondences between motion and music. In experiments, the two methodologies are compared and the motion analysis based methods outperform their rival by large margins. This work also proposes a quantitative measure of accurately recommended music genre. The proposed motion analysis based method achieves a recommendation accuracy of 91.3% using late fusion of joint and limb features.

INDEX TERMS Graph convolutional neural networks, human motion analysis, music recommendation.

I. INTRODUCTION

Human express feeling by dancing to music and different types of dancing help people express different emotions. The music and motion sequences both express similar amount of tension and emotion [1]. Human dance patterns reflect the musical rhythm [2], e.g., periodic dance movements tend to be synchronized to the speed of musical pulse [2].

Considering the close relationships between dance motions and the accompanying music, researchers explore how to synthesize dance movements for a given audio track [3]–[9]. These algorithms can be used for automatic choreography [10], human-avatar interaction [9], animations [11], [12], and robotics [13].

Music Recommender Systems (MRSs) are often very successful to suggest songs based on the users' preferences [14] but automatic music recommendation methods according to additional information using deep learning algorithms have rather seldom been studied. Studies on sports dance show that music together with dance motions are better representations than dance motions, for example, in sports dance teaching [15]. Music recommendation based on dance motions can be applied to automatic video editing such as automatic background music selection, and automatic music composition, etc. Despite its broad applications, there are only a few studies done on this topic [39], [43]. This work investigates music recommendation method based on dance motions.

Previous dance motion synthesis algorithms learn the correspondence between motions and audio from videos. For example, authors in [16] extract the correlated rhythmic representation from videos. They utilize beats to denote the rhythm of music, and use the dancer's motion trajectory, including the turning and the stop directions, to represent the rhythm of a motion. But RGB image sequences in videos are incoherent and noisy due to variations of human body size, human motion speed, and viewpoints [17]. Thus, it's extremely difficult to extract rhythmic information in dance videos.

This work uses 3D human skeleton data for music recommendation. Spatial information can be precisely extracted among joints in each frame and temporal information can be precisely extracted from consecutive frames.

Inspired by previous work on motion synthesis [3]–[6], [9], [18], we build a music recommendation model which learns motion-music correspondences. Shi *et al.* [38] propose to automatically generate motion from music using Long-short Term Memory (LSTM) and Autoencoder (AE), we instead build an end-to-end motion to music generator. The generator utilizes AE to reduce the input redundancy and LSTM to extract temporal information from the dance motions and the music. This method establishes the correspondences between the music and motion sequences.

The associate editor coordinating the review of this manuscript and approving it for publication was Gangyi Jiang.

Previous methods for motion synthesis usually evaluate the performances with qualitative measures.

We propose a novel deep learning algorithm for music recommendation based on dance motions analysis. The method achieves a good performance by transforming the music recommendation problem into a motion classification problem. After training the model, unseen test dance motions can be fed into the model to predict suitable accompanying music. Experimental results show that the proposed method achieves a high accuracy (91.30%) for music recommendation, which outperforms the music generation method (0% for per music piece accuracy, and 16.67% for per music genre accuracy) with a large margin. The main contributions of this work are as the followings.

1) This work introduces an effective way for music recommendation based on 3D human dance motions using deep learning, which outperforms the LSTM-AE based method.
2) This work includes the quantitative measure in evaluation, while previous methods [6], [9], [14], [38] only perform qualitative evaluations.

II. RELATED WORKS

A. AUDIO PROCESSING

There has been extensive research done in the field of audio processing. One crucial step in audio processing is audio feature extraction. Some of the traditional audio features include Mel-Frequency Cepstral Coefficient (MFCC) [19], constant-Q chromagram [20], tempogram [21], etc. In recent years, researchers have also explored end-to-end audio feature extraction methods.

Korzeniowski and Widmer [44] consider Chroma vectors to hold enough information to model harmonic content, but too much noise is generated in the process of calculating them compared with music score representations, such as MIDI. Hence, Korzeniowski and Widmer [22] propose a learned chroma vectors as feature extractor based on artificial neural networks. The proposed feature extractor is specially designed for chord recognition problems.

B. HUMAN MOTION ANALYSIS

Most of the works on human motion analysis rely on image or video sequence data. Vision-based human motion analysis has applications in visual tracking [45], [46], motion recognition [47]–[49], motion synthesis [50], [51], etc. For example, Fathi and Mori [27] propose a method for human action recognition using mid-level motion features. For each frame in the video, low-level optical flow features are extracted at each pixel. The low-level features are further weighted and combined as mid-level features by the AdaBoost algorithm [41]. These features are then utilized to discriminate among different action classes.

There are also works exploring motion analysis from depth data. Shaharyar *et al.* [28] use spatio-temporal features and modify hidden Markov model (HMM) for human detection, tracking and activity recognition.

Only a few works in literature demonstrate human motion analysis using 3D data [38], [58] despite its preciseness in capturing human motions. The main reason is the high cost of the specialized equipment for 3D motion capture. In this work, 3D motion capture data are utilized as additional information, thus circumventing the challenge of predicting 3D dance motions from video sequences [60].

C. MUSIC TO MOTION GENERATION

Research shows that there is a close relationship between dance motions and its accompanying music [7], [42]. Studies in [7] show the neural correlates of dance and music: expert dancers and expert musicians are found to have increased cortical thickness compared to untrained people in superior temporal regions, and gray matter structure in the superior temporal gyrus is correlated with both musical and dancing related tasks. This suggests that dancing and its accompanying music is biologically correlated.

Much work has been done in the area of automatic choreography generation from music. The work can be categorized into searching based methods [3], [4], [12], [29], [30], shallow neural networks like Factored Conditional Restricted Boltzmann Machine (FCRBM) [9], deep neural networks [6], [8], [10], [31], [32], and other methods including dynamic time warping [33] and hidden Markov models [5], [13].

On the contrary, there are very few work on automatic music generation or music recommendation according to human motions [39], [43]. Tsuchida *et al.* [39] propose a music recommendation method according to the dance motions. The method calculates the similarity between the query motion and the motion trajectory in the database, but the motion data are in the form of RGB images and only the 2-dimensional coordinate points are considered. In this work, we use 3D coordinates, and this is the first time to use deep learning algorithms to extract dance information for music recommendation methods.

D. MUSIC RETRIEVAL AND RECOMMENDATION

The most widespread application of music processing is music retrieval [55], [56], [59] and recommendation [14], [53]. Music retrieval and recommendation algorithms can be categorized into the following classes: collaborative methods, content-based methods [52], contextual based methods [54], hybrid methods, and sequential methods [53].

Collaborative methods utilize users' histories and assume that users with similar histories share similar interests. For example, Zhang *et al.* [25] propose a model called "Auralist" that is capable to consider four factors ("the desired goals of accuracy", "diversity", "novelty" and "serendipity") simultaneously. Other factors, such as tags and text descriptions are also exploited. For example, Sergio *et al.* [26] generate knowledge graphs for music recommendations.

Content-based methods extract information from the audio tracks directly. For example, Oord *et al.* [52] propose a deep convolutional neural networks based music recommendation

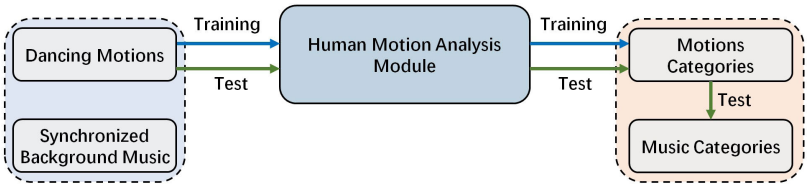


FIGURE 1. The diagram of the proposed music recommendation method.

FIGURE 2. The GCN-TCN based module.

method and the method outperforms a bag-of-words representation on the Million Song Dataset.

Previous music retrieval and recommendation methods only depend on audio features and metadata [57]. There are very few related works on cross-media solutions, such as music recommendation based on dance motions [39], [43]. Ohkushi *et al.* [43] propose to model the correlation between the dance video and the music based on kernel canonical correlation analysis (CCA), and recommend music pieces based on a query video sequence. Tsuchida *et al.* [39] utilize human poses predicted from video sequences as features to calculate similarities among dance video, and predict dance music given input dance videos.

The research on dance motion analysis for music recommendations using deep learning methods is still lacking. In this work, we will explore how to tackle this problem.

III. MOTION TO MUSIC

This work proposes a motion analysis based music recommendation method. The proposed method is designed based on the work of Shi *et al.* [38] on action recognition. Fig. 1 illustrates the diagram of the proposed music recommendation method based on human motion analysis. To compare with the proposed method, we also design a LSTM-AE based music recommendation method. The LSTM-AE based method uses motion as input, and predict music features, which are later compared with music features from the training data for music recommendation. This section describes the feature extractions and network models utilized for these approaches.

A. FEATURE EXTRACTIONS

This section explains the motion feature extraction and the music feature extraction procedures.

1) MOTION FEATURE EXTRACTION

The input motion data is denoted by 3D positions of skeleton joints. The original motion set in the synchronized dataset contains 25 joints [31]. We select 23 key action joints as

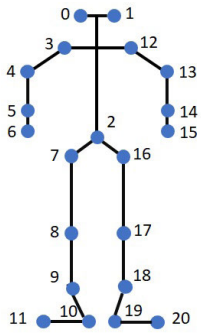


FIGURE 3. The human skeletal model.

our input. The first 21 key joints (#0 $\sim$ #20) are illustrated in Fig. 3 and the extract joints (#21 and #22) are used to describe the position of the hands. Then the skeleton is translated for normalization. All joints in each frame are subtracted by the geometric center of the frame.

To extract spatial and temporal features of a dance motion sequence, we explore a GCN (Graph Convolutional Network) [34] based framework. The input dimensions of the motion sequence are $\#frame \times 23 \times 3$, where $\#frame$ is the number of frames, 23 is the number of human body joints, and 3 denotes the human body joints in 3 dimensions. Input features are processed through a fully connected layer, a graph convolutional layer, and a TCN (Temporal Convolutional Network) layer consecutively, as shown in Fig. 2. A residual connection is drawn from the output of the fully connected layer to the output of the TCN. The TCN module applies a normal 2D convolutions along the temporal axis.

The GCN module incorporates kinetic information from the human body and the joint correlations of various motions from input data. The GCN module considers the graphical structure of the human skeleton and operates to extract spatial features. The structure of the GCN module is illustrated in Fig. 4. The input dimensions of the GCN module are (N, C, T, V), where N is the batch size, C is the number of input channels, T is the number of frames, and V is the number of joints. The input stream is processed by two

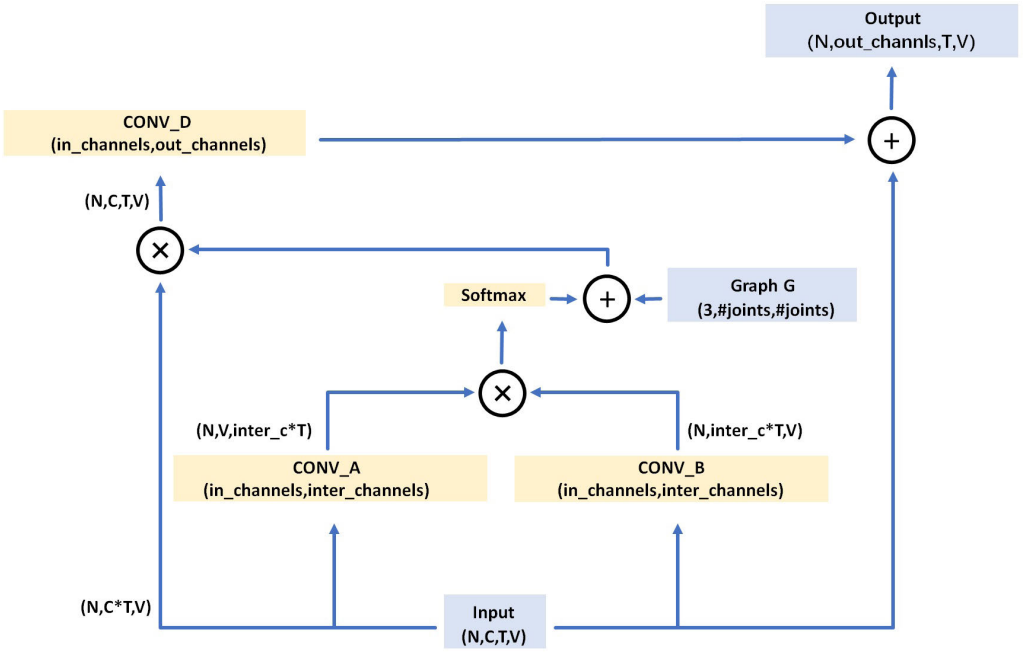


FIGURE 4. The GCN module.

FIGURE 5. The pipeline of the motion classification method.

independent pipelines. Each applies a convolution filter of size (*in_channels*, *inter_channels*), where *in_channels* is the number of input channels, and *inter_channels* is the number of the intermediate channels. Two streams are combined and a softmax layer is applied to calculate a vector of probabilities. The resulted probabilities are added to the graph matrix G. The value of V is equal to the value of *#joints* (notations are chosen based on different naming conventions). Then the tensor is multiplied by a rearranged input tensor. Afterwards, the tensor is rearranged and processed with a convolutional filter of size (*in_channels*, *out_channels*). And the result tensor is added to the initial input tensor similar to a residual block. The procedure described in Fig. 4 is repeated for each slice of the graph matrix G, and the three outputs are added together to get the final output.

This module incorporates human motion kinetics by introducing a pre-defined graph structure G. The graph consists of adjacency matrices and encodes the connectivities among joints defined by the human skeleton structure. The dimensions of the graph tensor are $3 \times$ *#joints* $\times$ *#joints*, where *#joints* is the number of joints. The three slices of the graph tensor correspond to the joint connection matrix ($G(1, :, :)$) which describes the connection between joints, the joint matrix ($G(2, :, :)$) which is an identity matrix, and the contra-connection matrix ($G(3, :, :)$) which is the transpose matrix of $G(1, :, :)$.

The TCN module is implemented using a standard convolution module. The input dimensions of the first TCN module are rearranged to (N, *channels*, *#frames*, V), where N is the batch size, *channels* is the number of feature channels, *#frames* is the number of temporal frames, and V is the number of joints. By reshaping the tensor, the TCN module carries out convolution along the temporal axis.

The motion analysis module defines the loss as the cross entropy between the prediction probability distribution and the ground truth probability distribution. The batch size is set as 10, the number of training epochs is set as 1000, the learning rate is set as 0.1, and after 40 epochs, the learning rate is set as 0.01. Fig. 5 shows the overall structure of the motion analysis module.

2) MUSIC FEATURE EXTRACTION

It is a one-dimensional array whose length is determined by the audio length and sampling rate. Audio features are usually extracted from the audio waveform. Fig. 6 visualizes one example for each dance category in the synchronized dataset [31]. Four audio features are extracted: Mel Frequency Cepstral Coefficients(MFCC) [19], constant-Q chromagram [20], onset strength envelope [35] and tempogram [21].

Spectrogram is the collection of short-term power spectrum resulted from the audio signal processed by the

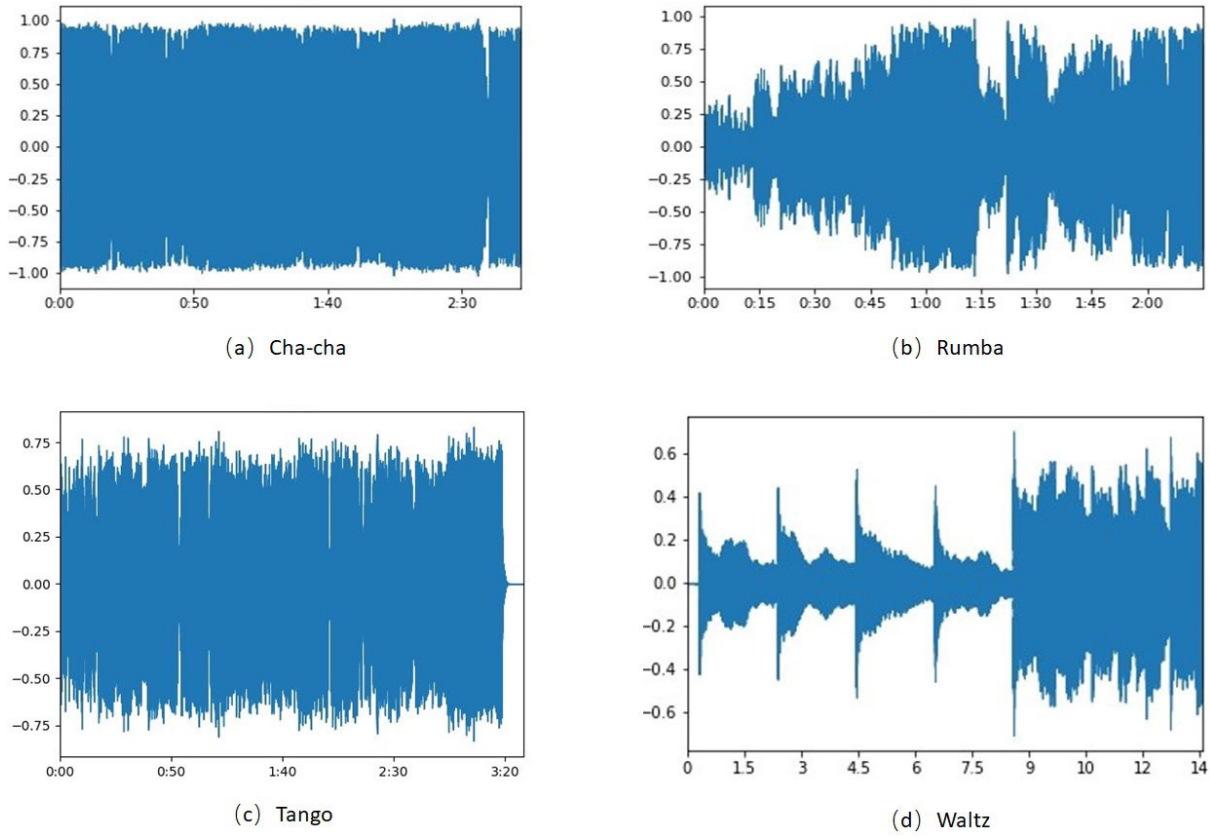


FIGURE 6. Exemplar audio waveform for each category in the synchronized dataset [31]. These four examples are from the four categories, respectively: (a) Cha-cha (b) Rumba (c) Tango (d) Waltz.

Short-time Fourier Transform (STFT). It transforms the original audio signal from time series into frequency domain. The original feature is 1-dimensional while the transformed feature is 2-dimensional. Mel-filter simulates the human perception and transforms the linear spectrum map into non-linear Mel-spectrum. MFCC (Mel Frequency Cepstral Coefficients) is the low frequency component after applying cepstrum analysis on Mel-spectrum. Cepstrum analysis is the reverse Fourier transformation of the log Mel-Spectrum. Cepstrum analysis transforms the non-linear problem back into linear problem.

Constant-Q chromagram is a chromagram after Constant-Q transform. The chroma is computed by summing the log-frequency magnitude spectrum across octaves and the chromagram is the sequence of chroma vectors.

Onset detection locates note onsets or percussive events and is the first stage of most beat tracking algorithms. The calculated Mel-spectrogram is converted to dB, and the first order differentiation is calculated along time in each frequency band. Each different equation is half-wave rectified. The onset strength envelope is then obtained by summing the positive differences across all frequency bands.

Tempogram is the basic method of analyzing the music beat interval. It indicates the number of the music beat interval within a period of time.

B. MUSIC RECOMMENDATION FROM DANCE MOTIONS

In this work, two types of methods are exploited: the motion-music mapping learning method based on LSTM-AE and the music recommendation method based on motion analysis. This section describes the approach procedures.

1) MOTION-MUSIC MAPPING MODELS BASED ON LSTM-AE

Inspired by [38], which automatically generates motion from music, we build an end-to-end motion to music generator using LSTM and AE. The kinematic constraints in the motion of rigid bodies result in redundancy [36]. This approach utilizes an autoencoder to reduce feature dimensions. An autoencoder reduces our motion feature to a lower-dimensional vector [37]. An autoencoder consists of two networks: the encoder and the decoder. Both the encoder and decoder are fully-connected feedforward neural networks and the encoder is usually symmetric to the decoder in terms of layer structure. LSTM Autoencoder is an autoencoder for sequence data using an encoder-decoder architecture where LSTM layers are utilized to build the encoder and the decoder networks.

The reduced-dimensional vector is input to the decoder to reconstruct the original vector. The loss function of the model is defined as the mean Euclidean distance between the predicted vector $\vec{x'}$ and the original vector $\vec{x}$. To reduce the overall loss, the encoder tends to pre-

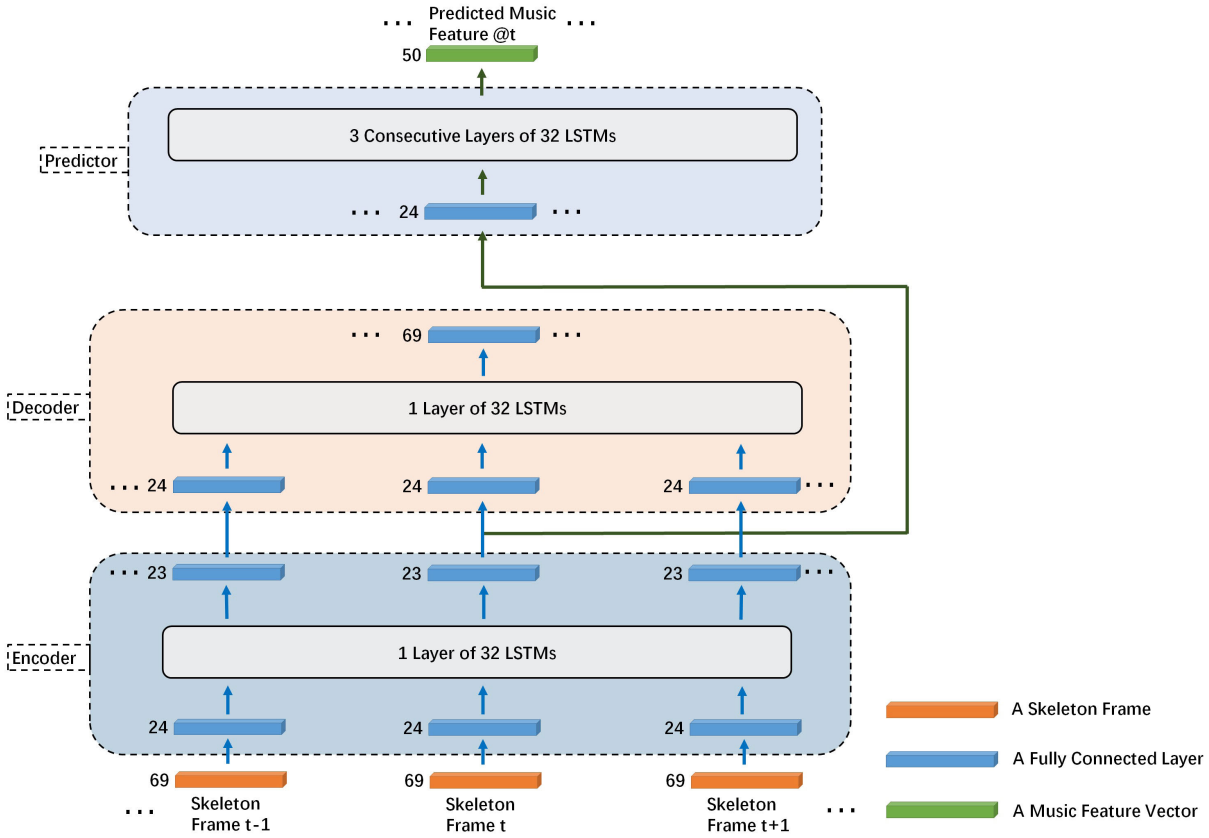


FIGURE 7. The network structure of the LSTM-AE based music recommendation model. An autoencoder reduces the input motion feature to a lower-dimensional vector, then the obtained features are used to predict the music features.

serve as much information as possible, while the decoder ensures that the feature vector with reduced dimensions contains sufficient information to represent the original vector.

Fig. 7 displays the architecture of the LSTM-Autoencoder based music recommendation method. The encoder model is comprised of one fully-connected layer, one LSTM layer and another fully connected layer. The decoder model utilizes the same structure to restore the skeletal data. The encoder compresses the input and produces the code. The code from the encoder is further processed by the predictor module. The predictor module is composed of one fully-connected layer, three consecutive layers of LSTMs and another fully-connected layer. The autoencoder module and the predictor module are trained alternately. According to the study in [31], when the loss of the autoencoder module is reduced below a certain value, the predictor performance will get worse. We choose a stopping threshold of 0.4 as in [31]. The optimization of the autoencoder network will stop once the loss of the autoencoder module reaches the threshold.

Motion features are fed into the encoder for dimensionality reduction, then the processed features are used to predict its corresponding music features, and the predicted music features are compared with the training music features in the dataset to recommend a music genre.

2) MUSIC RECOMMENDATION BASED ON MOTION CLASSIFICATION

This work adopts an adaptive graph convolutional network for dance motion analysis [38]. As illustrated in Fig. 5, the pipeline consists of several GCN-TCN modules. The first step is batch normalization. After applying BatchNorm, the resulting data are further processed with 10 consecutive GCN-TCN modules with a specific *inter_channels* value for each module. For simplicity of illustrated, we use a module of (GCN-TCN)$*$#*module* to denote a module formed by #*module* consecutive GCN-TCN modules.

In the inference stage, we feed the motion features into the network for classification, and the music genre is predicted as its corresponding motion genre. For each dance motion in the test set, we randomly select a piece from the predicted music category as their background music.

IV. EXPERIMENTAL RESULTS

A. DATASET

There is a great deal of datasets available on music classification and action recognition, and there are a few dataset consisting of dance videos with accompanying music [39], [43]. There are very few dataset consisting of motion and music data that are synchronized. Tang *et al.* [31] collect a synchronized music-dance dataset, totalling 907,200 frames,

TABLE 1. Quantitative evaluations on the test data. Experiment 0 is a LSTM-AE based method, and experiment 1-4 are motion analysis based methods. Experiment 1 uses skeleton node of joint positions as input, experiment 2 uses limbs, experiment 3 is a feature level fusion method, and experiment 4 is a decision level fusion method.

Experiment	Input Dimensions	Method	Accuracy (%) per music piece	Accuracy (%) per music genre
0	69×120	LSTM-AE	0	16.67
1	$46 \times 3 \times 229 \times 23$	Motion (Joints)	-	45.65
2	$46 \times 3 \times 229 \times 23$	Motion (Limbs)	-	89.13
3	$46 \times 6 \times 229 \times 23$	Motion (Joints&Limbs)	-	78.26
4	$46 \times 3 \times 229 \times 23$	Motion (Exp1_Exp2)	-	**91.30**

containing 60 complete dance choreographies for 4 dance types (waltz, tango, cha-cha, and rumba). The dataset is collected with optical motion capture equipment (Vicon) and the frame rate is 25 fps. Tang *et al.* [31] ask professional dancers to choose the music that will accompany a dance motion sequence.

With a 3D human pose estimation module, the proposed method is capable of recommending background music for dance videos. The additional module will reduce the overall performance due to imprecise pose predictions. In this scenario, it is difficult to explore the extent to which dance motion cues is beneficial for music recommendation. As a result, we use the dataset [31] with synchronized music and 3D dance motions for evaluation. The dataset is randomly split into three subsets: training (1/3), validating (1/3) and testing (1/3).

B. QUANTITATIVE EVALUATIONS

Unlike the previous dance-driven music recommended approaches [39], which only provide qualitative evaluation, we develop a quantitative measure for evaluation. The proposed qualitative measure provides insights on style qualities of the recommended music.

We provide quantitative evaluations to judge whether the recommended music piece belongs to the same style with the dance motions. Recommendation accuracy is defined as the total number of correct predictions divided by the total number of predictions. We evaluate the designed LSTM-AE based approach and the motion analysis based approach. In the motion analysis based approach, all the motion sequences and the audio waveforms are truncated to have the same length (229 frames). The data samples have varying lengths. The minimum frame number is 229, so all samples are truncated to have the same length. For each category, we choose 34 pairs of music-motion segments. The music-motion segments are labelled based on their music types: ''waltz'', ''tango'', ''cha-cha'', and ''rumba''. Various configurations of the motion analysis based approach are listed as the followings. The code and the data for rerunning all experiments are available at: https://github.com/wenjuangong/qingshuangyu-HMMR.

The motion analysis based experiments evaluate two types of data as input: joints and bones. These two features are combined using early fusion (feature fusion) and late fusion (decision fusion).

EXPERIMENT 0

This experiment runs the LSTM-AE based method. It takes human body joint positions as input. The input dimensions are set to 69×120, where 69 denotes the 23 human body joints in 3 dimensions, and 120 is the number of frames for each motion segment. The motion segment of consecutive frames are fed into a layer of LSTMs to extract features. The output is acoustic features. The predicted acoustic features are compared with the training acoustic features and the music genre labels are predicted based on feature distances.

EXPERIMENT 1

This experiment takes human body joint positions as input. The input dimensions are set to $46 \times 3 \times 229 \times 23$, where 3 denotes the human body joints in 3 dimensions, 299 is the number of frames for each motion segment, and 46 is the number of test samples (45 is the number of validation samples).

EXPERIMENT 2

This experiment utilizes human body limbs as input instead of human body joints. And the input dimensions are the same as in experiment 1.

EXPERIMENT 3

This experiment uses both joint and limb positions as input so the input dimensions are $46 \times 6 \times 229 \times 23$.

EXPERIMENT 4

In this experiment, we fuse the predictions from both experiment 1 and experiment 2. Two predictions are added together to form the final prediction. This experiment gives the highest accuracy (91.3%).

The experimental results show that the motion analysis based methods have much higher recommendation accuracies than the LSTM-AE based method, as shown in Table 1.

In experiment 1, skeleton node only keeps the joint positions, but loses both joint connection information and limb rotation information. Consequently, experiment 2 (using limbs as input) outperforms experiment 1 (using joints as input) by a large margin (43.48%). Experiment 3 is feature level fusion and experiment 4 is decision level fusion. The best prediction accuracy is achieved in experiment 4 (91.30%). Overall, the results indicate that the late fusion effectively improves recommendation accuracy.

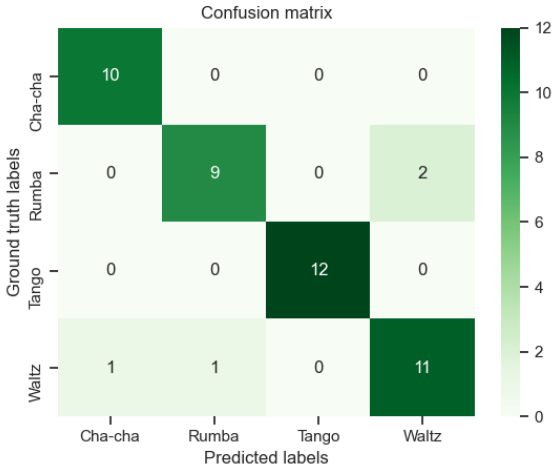


FIGURE 8. The confusion matrix of experiment 4.

To have a clear understanding of the relationships between categories, we visualize the confusion matrix for experiment 4 in Fig. 8. From the results, we can see that the proposed method performs well in distinguishing the Cha-cha and the Tango. On the contrary, it's hard to discriminate between Rumba and Waltz. The confusion between these classes are conceivable, e.g., the Rumba, like the Waltz is one of the slowest dances in the dataset.

V. QUALITATIVE COMPARISONS

We show the qualitative evaluations in this section. For each test data, the method recommends 3 background music pieces. The ground truth or a recommended music piece is integrated with the dancing skeletons to produce a music video. These videos are available at: https://sites.google.com/view/wenjuangong/projects/music-processing.

A. PROCESSING

We show these videos to the participants and ask them to rate the videos in a questionnaire. We recruited 9 participants from the university. Table 2 shows the profiles of the subjects. After watching the videos, the participants are asked to evaluate their style and tempo qualities on a 5-point scale. Also, the participants are asked to evaluate the overall quality compared with the ground truth data. The evaluation questions are as the following.

Q1: Considering the atmosphere of the dance and the recommended music piece, participants determine whether the dance movements match with the recommended music piece in style. Specifically, the participants rate the quality on a 5-point scale: the worst match (1 point), slightly better than the worst (2 points), a reasonable match (3 points), a very good match (4 points), or a perfect match (5 points).

Q2: Participants determine whether the dance movements match with the recommended music piece in tempo. Specifically, participants rate the quality on a 5-point scale: the tempos do not match at all (1 point), the tempos match about 40% of the time (2 points), the tempos match about 60% of the time (3 points), the tempos match about 80% of the time (4 points), or the tempos match perfectly (5 points).

Q3: Participants determine whether the recommended background music suits the dance movements better than the ground truth data. Specifically, the participants answer yes or no.

TABLE 2. The profiles of the subjects. The profiles of the recruited participants, including the total number of participants, their gender, age, nationality, and major.

Number of Subjects (Male/Female)	9 (5/4)
Age	$18-24$
Nationality	Chinese
Major	Music (3), Dance (3), Computer Science (3)

For Q1 and Q2, we calculate each participant's average score as the following:

$$PerSubject_{c,q,s} = \sum_{i,j} \frac{score_{i,j,c,q,s}}{scale \times N_i \times N_j}, \tag{1}$$

where $i \in \{1,2,3\}$ is the test index, $j \in \{1,2,3\}$ is the index for each of the three recommendations, c represents its category, $q \in \{1,2\}$ denotes the question, $s \in \{1,2,3,4,5,6,7,8,9\}$ is the subject index, $score_{i,j,c,q,s}$ is the score which participant s gives to the j-th recommendation of the i-th test data from the c-th category based on the q-th question, $scale$ is the scale of the scores ($=5$), N_i is the total number of the test data, and N_j is the number of recommendations ($=3$).

For Q3, $score_{i,j,c,q,s}$ is assigned the value 1 if the participant answers "yes", and 0 otherwise. Then, we apply Equation (1) to compute $PerSubject_{c,q,s}$ value. The $PerSubject$ value of 0.50 means that for every ten times, five of those times the participant considers the recommended music more suitable than the ground truth music. For a clearer understanding, we add a $+/-/\pm$ mark representing the overall performance of the recommendation. "+" means the participants consider the recommendation suits better than the ground truth, "−" means the ground truth suits better, and "±" means statistically they gives similar performances.

We filter out the $PerSubject$ records if they deviate from their group, e.g., when the participant 8 evaluates the Tango category, the $PerSubject$ value for Q3 ($+(0.83)$) is 58% more than the average $PerSubject$ value ($-(0.25)$) of the participant 7 ($-(0.33)$) and the participant 9 ($-(0.17)$).

Afterwards, the mean score for a major is calculated as the following:

$$PerMajor_{c,q,m} = \sum_{s\_m} \frac{PerSubject_{c,q,s}}{N_{s\_m}}, \tag{2}$$

where $PerSubject_{c,q,s}$ is computed according to Equation (1), m is the major index, and N_{s_m} is the total number of participants in a major.

TABLE 3. Qualitative evaluations on the test data. All recruited participants are asked to answer three questions enquiring the recommendation qualities. For each category, we provide the evaluation for each subject, the evaluation for each major, and the average over all subjects. Note, "GT" stands for the ground truth data.

Category	Subject Id	Style Quality (Q1)			Tempo Quality (Q2)			Quality Compared with the GT (Q3)		
		Per Subject	Per Major	Average	Per Subject	Per Major	Average	Per Subject	Per Major	Average
Cha-cha	1	0.98	0.90	0.79	0.98	0.90	0.78	+(1.00)	+(0.85)	+(0.67)
	2	0.84			0.77			+(0.67)		
	3	0.87			0.91			+(0.89)		
	4	0.93	0.89		0.91	0.89		+(1.00)	+(0.85)	
	5	0.91			0.93			+(0.67)		
	6	0.82			0.82			+(0.89)		
	7	0.76	0.57		0.69	0.57		−(0.44)	−(0.30)	
	8	0.44			0.47			−(0.33)		
	9	0.51			0.56			−(0.11)		
Rumba	1	0.87	0.81	0.74	0.91	0.77	0.72	+(1.00)	+(0.81)	+(0.65)
	2	0.76			0.73			+(0.56)		
	3	0.80			0.67			+(0.89)		
	4	0.93	0.81		0.91	0.81		+(0.89)	+(0.78)	
	5	0.71			0.76			+(0.67)		
	6	0.78			0.76			+(0.78)		
	7	0.73	0.61		0.67	0.59		−(0.33)	−(0.37)	
	8	0.51			0.56			−(0.33)		
	9	0.60			0.56			−(0.44)		
Tango	1	0.73	0.78	0.77	0.87	0.80	0.76	+(0.83)	+(0.83)	+(0.60)
	2	0.70			0.77			+(0.67)		
	3	0.90			0.77			+(1.00)		
	4	0.93	0.84		0.87	0.84		±(0.50)	+(0.61)	
	5	0.83			0.83			+(0.83)		
	6	0.77			0.83			±(0.50)		
	7	0.70	0.63		0.63	0.57		−(0.33)	−(0.25)	
	8[1]	~~0.83~~			~~0.80~~			~~+(0.83)~~		
	9	0.57			0.50			−(0.17)		
Waltz	1	0.73	0.81	0.78	0.69	0.72	0.74	+(0.67)	+(0.81)	+(0.63)
	2	0.87			0.83			+(0.75)		
	3	0.82			0.67			+(1.00)		
	4	0.96	0.84		0.93	0.83		+(0.78)	+(0.67)	
	5	0.82			0.73			+(0.67)		
	6	0.76			0.82			+(0.56)		
	7	0.67	0.62		0.69	0.64		−(0.33)	−(0.33)	
	8[2]	~~0.76~~			~~0.78~~			~~+(0.67)~~		
	9	0.58			0.60			−(0.33)		

[1] This record is considered as an outlier and the corresponding data are crossed out. The noisy data are not taken into consideration for calculating the "Per Major" value and the "Average" value.
[2] The same as [1].

Finally, the average value for a category is calculated as the following:

$$Average_{c,q} = \sum_{m} \frac{PerMajor_{c,q,m}}{N_m}, \tag{3}$$

where N_m is the total number of majors.

Table 3 shows the qualitative evaluation results. The ratings are subjective, e.g., participants major in Music or Dance tend to rate higher scores and they consider the recommended background music piece more suitable than the ground truth music piece. While participants major in Computer Science tend to rate lower scores and they consider the ground truth music piece more suitable for the dance motions. Yet some trends can be observed. For example, all the average ratings on style are higher than the average ratings on tempo. This indicates that although the recommended music is consistent with the dance style, more efforts should be spent on tempo coherence.

From the table, we observe that the music recommendation method achieves the best performance for the category "Cha-cha" and performs the worst for the "Rumba". When compared with the ground truth data, the recommendation method achieves the best performance for the category "Cha-cha" and performs the worst for the "Tango". Furthermore, the largest discrepancy between the style and the tempo is from the category "Waltz" and the smallest discrepancy is from the "Cha-cha" and the "Tango".

The qualitative evaluation is consistent with the results of the quantitative evaluations, e.g., the confusion matrix in Fig. 8 shows that the category "Rumba" and the category "Waltz" are the two most commonly confused categories, while in qualitative evaluation, "Rumba" performs the worst on style and tempo.

On the other hand, the "Tango" category (the most discriminative category according to Fig. 8) performs the worst when compared with the ground truth data in the qualitative evaluation. We consider this failure as

the result of tempo information loss in the recommendation procedure.

VI. DISCUSSION AND CONCLUSION

This work proposes a deep learning based method of music recommendation according to dance motions. In precedent work, we tried to establish a direct connection between motion and music through a LSTM-AE based dance motion generation from music, but from the experimental results, we can see that the LSTM-AE based solution is not feasible. So we resort to search-based approaches. To take advantage of the current advances of the deep learning algorithms, we use action recognition models to extract dance motion features, and recommend background music by classifying dance motions. This work also proposes a recommendation accuracy based quantitative measurement for evaluation. As far as we know, there is very few deep learning based music recommendation methods according to motion analysis. It is also the first time to introduce a quantitative measurement into evaluation for both motion synthesis and music recommendation. This work is concerned with the overall match between the music track and the whole dance motion sequences.

From the experimental results, we can conclude that the deep learning based music recommendation method is feasible. Through effective feature extraction of the deep learning framework and the correspondence between the dance motions and the music tracks, it is possible to build a music recommendation system that gives good performance. Research in this direction will have a profound impact on applications such as automatic choreography, automatic background music generation, and human-computer interaction. For future research, our work will be divided into two directions. The first is to further explore the characteristics of music and dance, and extract more representative feature to establish the connection between the two. The second is to continue to explore the deep learning models to build an end-to-end system from motion to music.

ACKNOWLEDGMENT

(Wenjuan Gong and Qingshuang Yu contributed equally to this work.)

REFERENCES

[1] C. L. Krumhansl and D. L. Schenck, "Can dance reflect the structural and expressive qualities of music? A perceptual experiment on Balanchine's choreography of Mozart's divertimento No. 15," *Musicae Scientiae*, vol. 1, no. 1, pp. 63–85, Mar. 1997.

[2] Y.-H. Su, "Rhythm of music seen through dance: Probing music–dance coupling by audiovisual meter perception," in *Proc. OSF*, Nov. 2017, pp. 1–20.

[3] R. Fan, S. Xu, and W. Geng, "Example-based automatic music-driven conventional dance motion synthesis," *IEEE Trans. Vis. Comput. Graphics*, vol. 18, no. 3, pp. 501–515, Mar. 2012.

[4] M. Lee, K. Lee, and J. Park, "Music similarity-based approach to generating dance motion sequence," *Multimedia Tools Appl.*, vol. 62, no. 3, pp. 895–912, Feb. 2013.

[5] F. Ofli, E. Erzin, Y. Yemez, and A. M. Tekalp, "Learn2Dance: Learning statistical music-to-dance mappings for choreography synthesis," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 747–759, Jun. 2012.

[6] H.-Y. Lee, X. Yang, M.-Y. Liu, T.-C. Wang, Y.-D. Lu, M.-H. Yang, and K. Jan, "Dancing to music," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 3581–3591.

[7] F. J. Karpati, C. Giacosa, N. E. V. Foster, V. B. Penhune, and K. L. Hyde, "Dance and music share gray matter structural correlates," *Brain Res.*, vol. 1657, pp. 62–73, Feb. 2017.

[8] Y. Qi, Y. Liu, and Q. Sun, "Music-driven dance generation," *IEEE Access*, vol. 7, pp. 166540–166550, 2019.

[9] A. Omid, F. Jules, and P. Philippe, "GrooveNet: Real-time music-driven dance movement generation using artificial neural networks," in *Proc. Workshop Mach. Learn. Creativity, ACM SIGKDD Conf. Knowl. Discovery Data Mining*, Halifax, NS, Canada, 2017.

[10] J. Lee, S. Kim, and K. Lee, "Listen to dance: Music-driven choreography generation using autoregressive encoder-decoder network," 2018, *arXiv:1811.00818*. [Online]. Available: http://arxiv.org/abs/1811.00818

[11] M. Cardle, L. Barthe, S. Brooks, and P. Robinson, "Music-driven motion editing: Local motion transformations guided by music analysis," in *Proc. 20th Eurographics UK Conf.*, 2012, pp. 38–44.

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

[13] A. Manfrè, I. Infantino, F. Vella, and S. Gaglio, "An automatic system for humanoid dance creation," *Biologically Inspired Cognit. Archit.*, vol. 15, pp. 1–9, Jan. 2016.

[14] M. Schedl, H. Zamani, C.-W. Chen, Y. Deldjoo, and M. Elahi, "Current challenges and visions in music recommender systems research," *Int. J. Multimedia Inf. Retr.*, vol. 7, no. 2, pp. 95–116, Jun. 2018.

[15] X. Zhang, "Impact of music on comprehensive quality of students in sports dance teaching," in *Informatics and Management Science VI*. London, U.K.: Springer, 2013, pp. 679–683.

[16] W.-T. Chu and S.-Y. Tsai, "Rhythm of motion extraction and rhythm-based cross-media alignment for dance videos," *IEEE Trans. Multimedia*, vol. 14, no. 1, pp. 129–141, Feb. 2012.

[17] M. Rubinstein, "Analysis and visualization of temporal variations in video," Ph.D. dissertation, Massachusetts Inst. Technol., Cambridge, MA, USA, Feb. 2014.

[18] W. Zhu, C. Lan, J. Xing, W. Zeng, Y. Li, L. Shen, and X. Xie, "Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks," 2016, *arXiv:1603.07772*. [Online]. Available: http://arxiv.org/abs/1603.07772

[19] 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.

[20] C. Schörkhuber and A. Klapuri, "Constant-Q transform toolbox for music processing," in *Proc. 7th Sound Music Comput. Conf.*, 2010, pp. 3–64.

[21] P. Grosche, M. Müller, and F. Kurth, "Cyclic tempogram—A mid-level tempo representation for musicsignals," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Dallas, TX, USA, Mar. 2010, pp. 5522–5525.

[22] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," 2016, *arXiv:1612.05065*. [Online]. Available: http://arxiv.org/abs/1612.05065

[23] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2017, pp. 2392–2396.

[24] J. Pons, J. Serra, and X. Serra, "Training neural audio classifiers with few data," in *Proc. ICASSP-IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2019, pp. 16–20.

[25] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, and T. Jambor, "Auralist: Introducing serendipity into music recommendation," in *Proc. 5th ACM Int. Conf. Web Search Data Mining (WSDM)*, 2012, pp. 13–22.

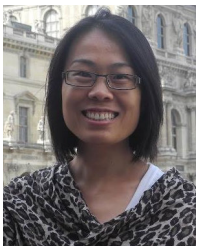
[26] S. Oramas, V. C. Ostuni, T. D. Noia, X. Serra, and E. D. Sciascio, "Sound and music recommendation with knowledge graphs," *ACM Trans. Intell. Syst. Technol.*, vol. 8, no. 2, pp. 1–21, Jan. 2017.

[27] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[28] S. Kamal, A. Jalal, and D. Kim, "Depth images-based human detection, tracking and activity recognition using spatiotemporal features and modified HMM," *J. Electr. Eng. Technol.*, vol. 11, no. 6, pp. 1857–1862, Nov. 2016.

[29] J. W. Kim, H. Fouad, and K. James Hahn, "Making them dance," in *Proc. AAAI Fall Symp., Aurally Informed Perform.*, 2006, pp. 1–5.

[30] T. Shiratori, A. Nakazawa, and K. Ikeuchi, "Synthesizing dance performance using musical and motion features," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, Dec. 2006, pp. 3654–3659.
[31] T. Tang, J. Jia, and H. Mao, "Dance with melody: An LSTM-autoencoder approach to music-oriented dance synthesis," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 1598–1606.
[32] W. Zhuang, C. Wang, S.-Y. Xia, J. Chai, and Y. Wang, "Music2Dance: Music-driven Dance Generation using WaveNet," 2020, *arXiv:2002.03761*. [Online]. Available: https://arxiv.org/abs/2002.03761
[33] R. Bellini, Y. Kleiman, and D. Cohen-Or, "Dance to the beat: Synchronizing motion to audio," *Comput. Vis. Media*, vol. 4, no. 3, pp. 197–208, Sep. 2018.
[34] N. T. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. ICLR*, 2017, pp. 1–13.
[35] S. Böck and G. Widmer, "Maximum filter vibrato suppression for onset detection," in *Proc. 16th Int. Conf. Digit. Audio Effects*, Maynooth, Ireland, 2013, pp. 1–7.
[36] Z. Hasan and S. J. Thomas, "Kinematic redundancy," *Prog. Brain Res.*, vol. 123, pp. 379–387, 1999. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0079612308628721
[37] D. E. Rumelhart and J. L. McClelland, "Learning internal representations by error propagation," in *Parallel Distributed Processing. Vol 1: Foundations*. Cambridge, MA, USA: MIT Press, 1986.
[38] L. Shi, Y. Zhang, J. Cheng, and H. Lu, "Two-stream adaptive graph convolutional networks for skeleton-based action recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12018–12027.
[39] S. Tsuchida, S. Fukayama, and M. Goto, "Query-by-dancing: A dance music retrieval system based on body-motion similarity," in *MultiMedia Modeling, 2019* (Lecture Notes in Computer Science), vol. 11295. Cham, Switzerland: Springer, 2019.
[40] Z. Cao, G. H. Martinez, T. Simon, S.-E. Wei, and Y. A. Sheikh, "OpenPose: Realtime multi-person 2D pose estimation using part affinity fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021.
[41] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of online learning and an application to boosting," in *Computational Learning Theory (EuroCOLT), 1995* (Lecture Notes in Computer Science: Lecture Notes in Artificial Intelligence), vol. 904. Berlin, Germany: Springer, 1995, doi: 10.1007/3-540-59119-2_166.
[42] B. Burger, J. London, M. R. Thompson, and P. Toiviainen, "Synchronization to metrical levels in music depends on low-frequency spectral components and tempo," *Psychol. Res.*, vol. 82, pp. 1195–1211, Jul. 2017, doi: 10.1007/s00426-017-0894-2.
[43] H. Ohkushi, T. Ogawa, and M. Haseyama, "Music recommendation according to human motion based on kernel CCA-based relationship," *EURASIP J. Adv. Signal Process.*, vol. 2011, no. 1, pp. 1–14, Dec. 2011.
[44] F. Korzeniowski and G. Widmer, "Feature learning for chord recognition: The deep chroma extractor," in *Proc. 17th Int. Soc. Music Inf. Retr. Conf.*, New York, NY, USA, 2016, pp. 37–43.
[45] W. Mrabti, K. Baibai, B. Bellach, R. O. H. Thami, and H. Tairi, "Human motion tracking: A comparative study," *Procedia Comput. Sci.*, vol. 148, pp. 145–153, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1877050919300183
[46] H. Zhou and H. Hu, "Human motion tracking for rehabilitation—A survey," *Biomed. Signal Process. Control*, vol. 3, no. 1, pp. 1–18, Jan. 2008.
[47] P. Wang, W. Li, P. Ogunbona, J. Wan, and S. Escalera, "RGB-D-based human motion recognition with deep learning: A survey," *Comput. Vis. Image Understand.*, vol. 171, pp. 118–139, Jun. 2018.
[48] P. Wang, H. Liu, L. Wang, and R. X. Gao, "Deep learning-based human motion recognition for predictive context-aware human-robot collaboration," *CIRP Ann.*, vol. 67, no. 1, pp. 17–20, 2018.
[49] R. Lun and W. Zhao, "A survey of applications and human motion recognition with microsoft Kinect," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 29, no. 5, Aug. 2015, Art. no. 1555008.
[50] D. Zhou, X. Feng, P. Yi, X. Yang, Q. Zhang, X. Wei, and D. Yang, "3D human motion synthesis based on convolutional neural network," *IEEE Access*, vol. 7, pp. 66325–66335, 2019.
[51] A. C. Fang and N. S. Pollard, "Efficient synthesis of physically valid human motion," *ACM Trans. Graph.*, vol. 22, no. 3, pp. 417–426, Jul. 2003.
[52] A. van den Oord, S. Dieleman, and B. Schrauwen, "Deep content-based music recommendation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 26, 2013, pp. 2643–2651.
[53] M. Schedl, P. Knees, B. McFee, D. Bogdanov, and M. Kaminskas, "Music recommender systems," in *Recommender Systems Handbook*, F. Ricci, L. Rokach, and B. Shapira, Eds. Boston, MA, USA: Springer, 2015.
[54] M. Kaminskas and F. Ricci, "Contextual music information retrieval and recommendation: State of the art and challenges," *Comput. Sci. Rev.*, vol. 6, nos. 2–3, pp. 89–119, May 2012.
[55] M. Schedl, E. Gómez, and J. Urbano, "Music information retrieval: Recent developments and applications," *Found. Trends Inf. Retr.*, vol. 8, nos. 2–3, pp. 127–261, 2014.
[56] M. Müller, *Fundamentals of Music Processing: Audio, Analysis, Algorithms, Applications*. Cham, Switzerland: Springer, 2015.
[57] M. Müller, *Information Retrieval for Music and Motion*. Berlin, Germany: Springer-Verlag, 2007.
[58] C. Si, Y. Jing, W. Wang, L. Wang, and T. Tan, "Skeleton-based action recognition with spatial reasoning and temporal stack learning," in *Proc. ECCV*, vol. 11205, 2018, pp. 103–118.
[59] K. Choi, G. Fazekas, K. Cho, and M. Sandler, "A tutorial on deep learning for music information retrieval," 2017, *arXiv:1709.04396*. [Online]. Available: http://arxiv.org/abs/1709.04396
[60] C. Sminchisescu, B. Rosenhahn, R. Klette, and D. Metaxas, "3D human motion analysis in monocular video: Techniques and challenges," in *Human Motion*. Amsterdam, The Netherlands: Springer, 2008, pp. 185–211.



WENJUAN GONG (Member, IEEE) received the Ph.D. *(cum laude)* degree from the Autonomous University of Barcelona, in 2013. She was a Postdoctoral Research Assistant with Oxford Brookes University, in 2014. She is currently a Lecturer with the China University of Petroleum. She participated in the European Project Consolider Ingenio 2010 and the EPSRC Project Tensorial Modeling of Dynamical Systems for Gait and Activity Recognition. She has led the CCF-Tencent Funds and the Natural Science Foundation of China of Shandong Province. Her research interests include computer vision and machine learning.

QINGSHUANG YU was born in Shandong, China, in 1996. He received the B.S. degree from Qufu Normal University, China, in 2019. He is currently pursuing the master's degree with the College of Computer Science and Technology, China University of Petroleum, China. His research interests include deep learning and multimedia information fusion processing.