Automatic Sign Dance Synthesis from Gesture-based Sign Language

Naoya Iwamoto
iwamoto@toki.waseda.jp
Waseda University
Tokyo, Japan

Hubert P. H. Shum*
hubert.shum@northumbria.ac.uk
Northumbria University
Newcastle upon Tyne, UK

Wakana Asahina
2236-wakana@fuji.waseda.jp
Waseda University
Tokyo, Japan

Shigeo Morishima
shigeo@waseda.jp
Waseda University
Tokyo, Japan

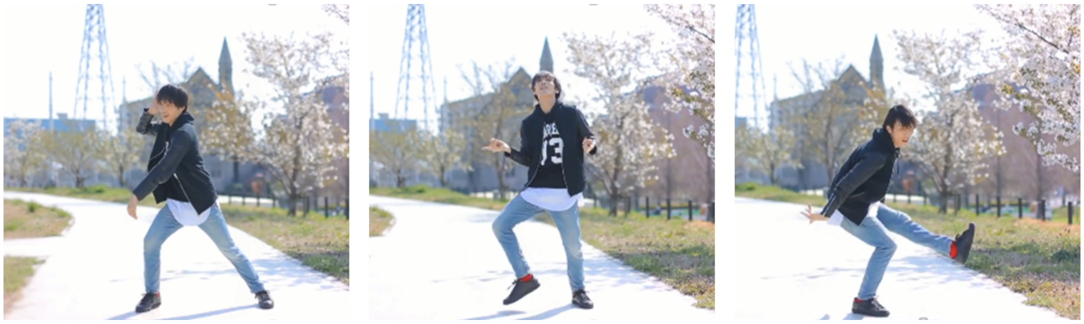

Figure 1: Sign dance - a genre of dance consisting of both sign gestures and body dancing motions.

ABSTRACT

Automatic dance synthesis has become more and more popular due to the increasing demand in computer games and animations. Existing research generates dance motions without much consideration for the context of the music. In reality, professional dancers make choreography according to the lyrics and music features. In this research, we focus on a particular genre of dance known as *sign dance*, which combines gesture-based sign language with full body dance motion. We propose a system to automatically generate sign dance from a piece of music and its corresponding sign gesture. The core of the system is a *Sign Dance Model* trained by multiple regression analysis to represent the correlations between sign dance and sign gesture/music, as well as a set of objective functions to evaluate the quality of the sign dance. Our system can be applied to music visualization, allowing people with hearing difficulties to understand and enjoy music.

*Correspondence author

MIG '19, October 28–30, 2019, Newcastle upon Tyne, United Kingdom
© 2019 Association for Computing Machinery.
ACM ISBN 978-1-4503-6994-7/19/10…$15.00
https://doi.org/10.1145/3359566.3360069

CCS CONCEPTS

• **Computing methodologies** → **Computer graphics**; Artificial intelligence; • **Artificial intelligence** → Computer vision.

KEYWORDS

Motion Synthesis, Dance, Sign Language, Multiple Regression Analysis

ACM Reference Format:
Naoya Iwamoto, Hubert P. H. Shum, Wakana Asahina, and Shigeo Morishima. 2019. Automatic Sign Dance Synthesis from Gesture-based Sign Language. In *Motion, Interaction and Games (MIG '19), October 28–30, 2019, Newcastle upon Tyne, United Kingdom.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3359566.3360069

1 INTRODUCTION

In recent years, thanks to the development of 3D animation editing tools such as the MikuMikuDance software [Higuchi 2014], the industry has become interested in producing dance animations with 3D characters. However, it is labor-intensive to choreograph high-quality dances from scratch. Such a process requires an understanding of body motion dynamics and music theories, which may be challenging to novice users. As a result, there is an increasing demand for dance synthesizing systems that automatically generate dance motion based upon a piece of input music. For example, Shiratori et al. [Shiratori et al. 2006] produced an automatic dance generation system utilizing the rhythm and intensity of dance motions. Since each segment is selected randomly from the database, the generated dance motion has no linguistic or emotional meaning. Takano et al. [Takano and Nakamura 2015] produced a human motion generation system which utilized motion labels. However, the labels used such as "running" or "jumping" are too simple to express complex emotions, which are important elements in music.

Professional dancers design choreography according to the context of the music. For music with lyrics (i.e. songs) that we focus on in this research, the context can be obtained from the lyrics and music features. For example, if the lyrics are about "big sky", many dancers would make a choreography that is looking up at the sky and opening their arms. This observation of the intrinsic relationship between dance motion and lyrics motivates us to generate dance motion that carries the contextual meaning of the music.

In this paper, we consider a particular genre of dance known as the *sign dance* as shown in Fig. 1. It is a type of dance combining gesture-based sign language and full-body dance moves. With sign dance, one can understand the context of the music from the sign gesture, and understand the impression of music like tempo and timbre from the dance move. This is particularly important for people with hearing difficulties, who rely on music visualization to understand music. Compared with traditional geometric shape based music visualization systems like those in iTunes, sign dance explicitly expresses lyrics and musical features, allowing the audience to understand the context and feel the emotion of a song. This can help people with hearing difficulties to enjoy music.

To implement an automatic sign dance generation system, we prepare a database including tuples of sign gesture, sign dance and music segments, which serve as prior knowledge. In the offline stage, we utilize multiple regression analysis to train a *Sign Dance Model* that combines sign dance and sign gesture/music. During the online stage, our system takes a piece of music and its corresponding sign gesture, which is translated from the lyrics of the music, as the input. With the trained Sign Dance Model and a set of objective functions that evaluate motion quality, our system synthesizes natural sign dance motion automatically.

To design the objective functions, we observed sign dances available in the music industry, and concluded some important criteria to generate high-quality sign dance. First, from the dance choreography point of view, there are implicit correlations between the hand and body motion. For example, a hand gesture of lowering the arms usually correlates to a squatting body motion to amplify the perception of the hand movement. Second, the body motion should match the impression of the music. High-intensity music usually implies a high level of emotion and should match with greater body movement. Third, the arms movement should align with the right stepping motion such that body balance is maintained. Finally, the generated sign dance should be smooth and natural. We formulate these criteria as objective functions to obtain the suitable sign dance motion.

Experimental results demonstrate that our system produces high-quality sign dances that are comparable to those performed by professional sign dancers. User studies show that our method performs better than the baseline method using random dancing motions. The computational complexity of our system is low, allowing it to be applied in the entertainment industry for interactive animation creation, dance choreography and console games.

This research presents the following contributions:

- We propose a data-driven system to automatically generate sign dance. Our system learns the correlations between gesture-based sign language and the corresponding sign dance motion/music in a database using multiple regression analysis. Such correlation is then used to synthesize new sign dance motion based on an input music and the corresponding sign language.
- We propose a set of objective functions that evaluate the quality of sign dance in order to synthesize high-quality dance motion. Such objective functions consider the matching quality between dance motion and sign gesture, the correlations between music and motion, body stability, as well as the motion connectivity.

The rest of the paper is organized as the following. In Section 2, we review works that are related to dance analysis and motion synthesis. In Section 3, we give an overview of the proposed system and point out the major system components. Section 4 explains the Sign Dance Model, which correlates sign dance segments with sign gesture/music segments. Section 5 presents a set of objective functions and explains how to synthesize high-quality sign dance. Section 6 presents both qualitative and quantitative results to support the quality of our system. Section 7 concludes the paper and discusses on limitations and future directions.

2 RELATED WORK

There is an increasing demand for dance synthesizing systems that can efficiently generate dance motion. Shiratori et al. [Shiratori et al. 2006] proposed an approach to synthesize dance performances that matched well with the input music by aligning the motion energy and the music intensity. However, the system did not generate dance motion according to the linguistic or emotional meanings of the music. Takano et al. [Takano and Nakamura 2015] proposed an approach to construct a space of motion labels such as "running" and "jumping". The system recognized a series of motions as motion symbols and projected the recognition results onto the motion label space. However, the structure of the proposed motion labels could not represent complex linguistic meaning and emotion, which are important elements in music. Xue et al. [Peng et al. 2016] proposed an automatic character motion generation system by using machine learning. They introduced a mixture of actor-critic experts approach that parameterized leaps and steps as the output actions. This approach, however, only targeted basic movements. Wilke et al. [Wilke et al. 2005] synthesized dancing characters using the Laban notation, which is a script to describe dance moves. Such a script is limited and cannot be used to denote complex hand gestures. Iwamoto et al. [Iwamoto et al. 2017] created a DJ interface for animators to interactively create dance movement.

In terms of dance analysis, Chen et al. [Chen et al. 2006] proposed a set of functions to evaluate the physical effort to perform a dance based on kinematics features. Dyaberi et al. [Dyaberi et al. 2004] proposed a topology graph structure to represent the phrasal structure in dance. Neave et al. [Neave et al. 2010] conducted a user study to find out the defining criterial of an attractive dance. Hoyet et al. [Hoyet et al. 2013] performed a similar research by showing dance moves to the subjects performed by virtual characters. Raptis et al. [Raptis et al. 2011] designed a set of features to classify a large number of dance moves in real-time. Chan et al. [Chan et al. 2011] presented a to render dancing characters for dance training using motion capture systems. Tang et al. [Tang et al. 2011] enhanced the system performance in dance moves recognition. However, such a system focused on analyzing user's dancing mvoes and display pre-recorded dance moves only.

Previous research has also looked into the possibility of generating multi-character interaction including dancing. Ho et al. [Ho et al. 2013] proposed a system to generate two people dancing from the motion of one dancing actor. The focus of the research was about inter-character relationship instead of the context of the dance. Tang et al. [Chan et al. 2013] identified a number of interaction types and embeded this idea to interaction synthesis. Shum et al. [Shum et al. 2012] simulated two character boxing using temporal tree expansion and games theory. Hyun et al. [Hyun et al. 2016] presented a set of high-level language known as motion grammar to describe interaction for multi-character interaction synthesis. Shen et al. [Shen et al. 2019] demonstrated a mesh structure to understand multi-character interaction. In this work, we consider the correlation between dance and sign gesture for a virtual characters. The research mentioned above could provide possible extensions for multi-character dance synthesis as a future work.

Deep learning based methods has produced promising results in kinematics motion synthesis. Holden et al. generated realistic, controllable locomotion by considering the different stepping phase of a locomotion patter. Zhang et al. [Zhang et al. 2018] synthesized quadruped motion by combining a motion prediction network that computes the character state, and a gating network that dynamically updates the weights of the motion prediction network. Holden et al. [Holden et al. 2016] adapted an autoencoder with a convolutional neural network focusing on the temporal aspect of human motion to generate a effective latent space for synthesizing new motion. Zhou et al. [Zhou et al. 2018] proposed a new recurrent neural network to synthesize arbitrary motions with highly complex styles. Wang et al. [Wang et al. 2019] included the concept of temporal prediction in recurrent neural network to create a more stable motion manifold. Lee et al. [Lee et al. 2018] synthesized human-object interaction by introducing motion grammar, a syntax to describe interaction, into a deep neural network. A common problem of deep learning is that it requires a significant amount of data, which is not suitable for specific problem such as ours, in which the training data is limited.

We focus on using gesture-based sign language to generate the dance motions that carry the meanings of the music lyrics. Hiruma et al. [Hirumura et al. 2015] proposed a framework to generate Japanese sign language animation automatically. Similarly, Baldassarri et al. [Baldassarri et al. 2009] implemented a system to translate written or spoken language into Spanish sign language. Gibet et al. [Gibet et al. 2011] presented a virtual character capable of French sign language. Efthimiou et al. [Efthimiou et al. 2009] presented an example-based sign language dictionary for language-sign translations. While these systems could create virtual characters performing sign language, the focus was on hand and arm movement only. Oshita et al. [Oshita and Senju 2014] proposed a method to generate hand motion from full-body motion. However, their output hand motion was merely reflective movement without context. Helge et al. [Rhodin et al. 2015] presented an approach to analyze hand movement and decompose it into wave parameters, including amplitude, frequency, and phase. The decomposed signals could then be used for character controls. However, due to the use of wave parameters, the considered hand motion involved only simple cyclic movement. Finally, Eugene et al. [Hsu et al. 2004] proposed a method for example-based performance control of human motion. They controlled a dance follower with the motion of a dance leader using stylized mappings. Hand gesture was not considered in this research. To the best of our knowledge, there is very limited work on considering the linguistic meaning of the music in dance synthesis, and there is no research on synthesizing full body dancing motion based on gesture-based sign language.

Our work is the most similar to a 2-pages poster focusing on generating sign dance [Asahina et al. 2016]. In this paper, we extend the work and implement a full system to synthesize sign dance. We design a set of new objective functions to construct the Sign Dance Model, which is a model that correlate dance moves and sign gestures. We synthesize multiple dance sequences based on different genes of music, and critically evaluate our system by conducting user studies.

3 SYSTEM OVERVIEW

The overview of our proposed system is shown in Fig. 2, which includes an offline stage and a run-time stage.

The offline process is shown in Fig. 2a. We build a database consisting of tuples of a captured sign gesture segment (light yellow shaded), the corresponding captured sign dance segment (dark red shaded) and music segment. The sign gesture and sign dance segments are full body motions captured from a professional sign dancer and segmented based upon the beat pattern of the music. Using the database, we train a Sign Dance Model that expresses the correlations between sign dance motion and sign gesture/music using multiple regression analysis.

During run-time, the input of the system is a piece of music (Fig. 2b) and the corresponding sign gesture (Fig. 2c). The inputs are segmented based on the beat pattern. Each of the music and sign gesture pairs (Fig. 2d) is used as the input of the Sign Dance Model to evaluate an expected sign dance motion. Together with some objective functions and hard constraints, we can obtain the optimal sign dance motion for each segment (Fig. 2e). Finally, we apply joint fusion to combine the corresponding sign gesture segments and the sign dance motion segments to synthesize the final sign dance (Fig. 2f).

4 THE SIGN DANCE MODEL

In this section, we introduce the Sign Dance Model, which acts as an oracle to estimate an appropriate sign dance based on a sign gesture and the corresponding music segment. We first explain how we construct the sign dance database that consists of sign gesture, sign dance and music segments. Then, we explain how we train the Sign Dance Model using multiple regression analysis.

4.1 Sign Dance Database

Here, we explain how we capture the sign gesture and the sign dance motion required to build the sign dance database.

We define the term sign gesture as the full body motion of performing gesture-based sign language. The actor stands still when

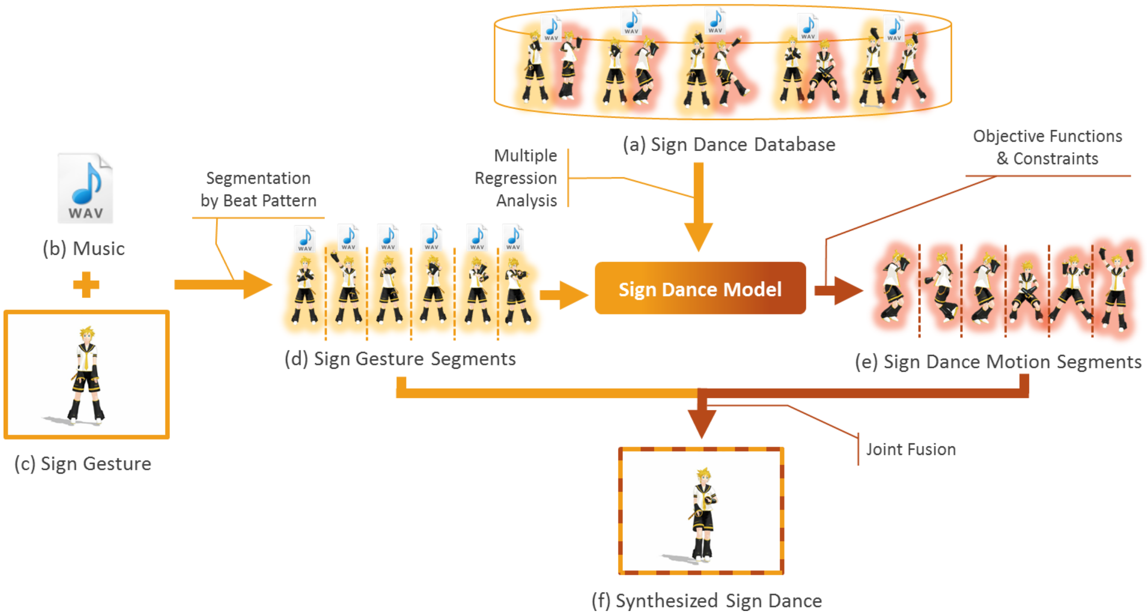
(a) Sign Dance Database
WAV
(b) Music
(c) Sign Gesture
Segmentation by Beat Pattern
Multiple Regression Analysis
(d) Sign Gesture Segments
Sign Dance Model
Objective Functions & Constraints
(e) Sign Dance Motion Segments
Joint Fusion
(f) Synthesized Sign Dance

Figure 2: The overview of our sign dance synthesis system.

performing the gesture, and the focus of the motion is the hand and arm joints. In this research, since we focus on songs, the sign gesture is translated based on the lyrics of a piece of music. For other types of music such as that in theatrical plays, the sign gesture can be translated from the storyline. We also define the term sign dance as the full body motion of performing sign dance motions. The actor combines gesture-based sign language with dance moves that are suitable for expressing the emotion and context of the music.

While the sign dance motion implicitly consists of the sign gesture, due to the dance movement, some of the gestures are exaggerated or altered. It is, therefore, ineffective to extract accurate sign gestures from sign dances. As a solution, we capture the two motions separately and synchronize them with the music.

The synchronized sequences of captured sign gestures, captured sign dance and music for each song are cut into shorter segments. These tuples of segments, including sign gesture, sign dance and music, are the building blocks of our sign dance database. The segmentation is done based on the beat pattern of the song. In particular, each musical bar (or measure) of the music is considered as one segment. Typically, a bar is around 2 - 3 seconds long depending on the beat per minute of the music. The justification for this segmentation method is that dance moves are usually synchronized to the beat pattern of the music, and each bar represents a short but meaningful duration. By cutting the dance motion according to bars, we can minimize the chance of cutting a continuous dance move into multiple parts. Finally, we normalize the duration of the segments by the mean number of frames of all segments in the database, which is 3.09 seconds in our database.

During the capture, we employed a professional sign dancer, who had extensive experience in the choreography of sign dance, to perform both sign gesture and sign dance motions for a number of songs. We utilized a VICON 12-cameras optical motion capture system to capture the motions as illustrated in Fig. 3. The captured 42 marker positions were converted into a 26 joints skeletal orientation format using the inverse kinematics function provided in the software Autodesk MotionBuilder. Each joint was represented by a 4D quaternion vector. Finger movement was not included due to capturing difficulties. We obtained synchronized sign gesture and sign dance for 10 different songs, which covered different music impressions according to the music tempo and timbre. The data was cut into 259 tuples of sign gesture, sign dance and music segments.

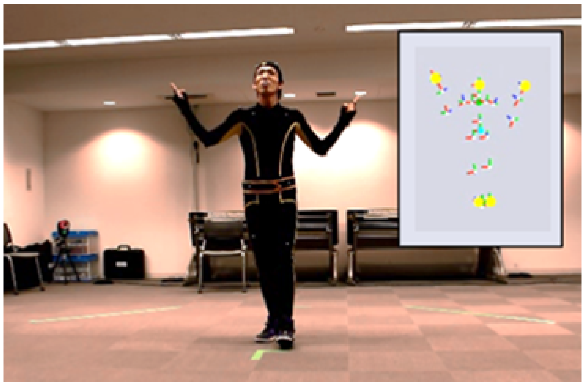

Figure 3: Sign dance motion captured from a professional sign dancer.

4.2 Multiple Regression Analysis

Here, we explain how we train the Sign Dance Model with multiple regression analysis using the data in our sign dance database.

The model expresses the correlation between the sign dance and the sign gesture with music. In particular, we use the features of the sign gesture and the audio as the explanation variables, and use the features of the sign dance as the objective variables. Given one segment, we formulate the following equation:

$$d_i = \sum_{j=1}^{j_{total}} a_{ij} g_j + \sum_{k=1}^{k_{total}} b_{ik} m_k + c_i, \quad i \in [1, i_{total}], \tag{1}$$

where d_i denotes the i^{th} feature of the sign dance segment, i_{total} is the total number of sign dance features, g_j denotes the j^{th} feature of the sign gesture segment, j_{total} is the total number of sign gesture features, m_k denotes the k^{th} feature of the music segment, m_{total} is the total number of music features, a_{ij}, b_{ik} and c_i are regression coefficients. To train the Sign Dance Model, we substitute the sign dance, sign gesture and music tuples from the database into the equation, and estimate the regression coefficients that can minimize the combined error of all tuples.

Designing the best features for the sign dance, sign gesture and music is more of an art than a science. In our experience, if the behaviours of the features between sign dance and sign gesture were more similar, the regression results tended to be better. Therefore, we hand-picked a subset of joints that exhibit similar movement behaviours between sign dance and sign gesture, including the neck, arms, forearms and hips. We serialize the temporal series of the quaternions of these joints into a feature vector. Since sign dance and sign gesture use the same feature representation, i_{total} is equal to j_{total}. For the music features, we use the spectral centroid that indicates the timbre and brightness of the music, and the pulse clarity that indicates the strength of the beat [Lartillot et al. 2008a]. This 2-dimensional audio feature is acquired using MIR Toolbox [Lartillot et al. 2008b].

The multiple regression analysis works very well in modelling our relatively small sign database. This is because there are strong correlations between the sign dance and the sign gesture/music. The model is linear and therefore is efficient to train while avoiding overfitting. For larger databases consisting of more complicated data, such as similar sign gestures mapping to significantly different sign dances, non-linear regression models can be considered.

5 SIGN DANCE SYNTHESIS

In this section, we explain how to synthesize the sign dance using the Sign Dance Model and a set of sign dance evaluation criteria. We first explain the framework for synthesizing sign dance. Then, we go into detail to explain individual objective functions and constraints.

5.1 The Synthesis Framework

Given an input sign gesture and the corresponding music, we first perform segmentation based on the beat pattern, which is the same process as mentioned in Section 4.1. We then synthesize the sign dance for each segment individually.

For each sign gesture and music segment pair, we first utilize the Sign Dance Model to obtain the expected sign dance features. We then search for the best sign dance segment in the database with reference to the expected sign dance features and a set of evaluation criteria and constraints. These include evaluating (1) how well the sign dance segment matches the expectation obtained from the Sign Dance Model, (2) how well the synthesized sign dance segment matches the feeling of the music segment, (3) how dynamically stable the synthesized sign dance segment is, and (4) how smoothly the synthesized sign dance segment connects to the previously synthesized segment. Detailed implementation information for each of these four criteria will be given in the next subsection.

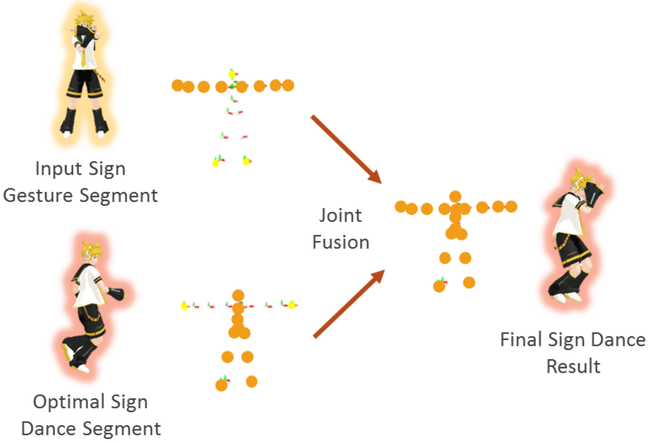

Figure 4: Combining different joints from sign gesture and sign dance segments to obtain the final result.

While the optimal sign dance segment obtained from the database should have a similar hand gesture to the input one, there may be some minor gesture variations. Therefore, we propose to utilize *joint fusion* (also known as *degree of freedom replacement*) for combining the arm-related joints from the two motion segments [Mousas et al. 2013]. Our implementation is visualized in Fig. 4, in which we fuse the arm and hand degrees of freedom from the input sign gesture with the body degrees of freedom from the optimal sign dance segment to generate the output sign dance.

5.2 Evaluation Criteria

Here, we explain the criteria to evaluate if a candidate sign dance segment is suitable for the input sign gesture and music segment. These criteria are applied when we search for the best sign dance segment in the database.

The first criterion is the sign dance expectation, f_e. With the trained Sign Dance Model, we extract the features from the input sign gesture, g_j, and music, m_k, and substitute them into Eq. 1. This gives us the expected sign dance features, d_i', which represent the sign dance that matches with the input sign gesture and music the best. We apply the following equation to evaluate the Euclidean distance between a candidate sign dance segment and the expected one suggested by the Sign Dance Model:

$$f_e = \sum_{i}^{i_{total}} \sqrt{(d_i - d_i')^2}, \tag{2}$$

where d_i' is the i^{th} feature of the expected sign dance segment, d_i is that of the sign dance segment to be evaluated. A smaller result indicates a better match.

The second criterion is the music-motion intensity, f_i. From the sign dance motion we captured, we observe that larger movement is usually performed when the music is louder. Therefore, we propose to evaluate the normalized difference between music intensity, which is measured as root mean square intensity, and motion intensity, which is measured as weight effort according to the Laban movement analysis [Wilke et al. 2005]:

$$f_i = \left| \frac{\sqrt{\frac{1}{f_{total}} \sum_{f=1}^{f_{total}} u_f^2}}{u_{norm}} - \frac{\frac{1}{f_{total}} \sum_{f=1}^{f_{total}} \sum_{n=1}^{n_{total}} \left| v_{n,f} \right|}{v_{norm}} \right|, \qquad (3)$$

where f_{total} is the total number of frames in the considered segment, u_f is the music power at frame f, n_{total} is the total number of joints, $v_{n,f}$ is the joint velocity of joint n at frame f, u_{norm} and v_{norm} are constants to normalize the music intensity part and the motion intensity part of the equation.

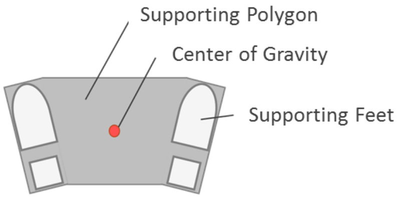

Figure 5: Utilizing the center of gravity to evaluate body balance.

The third criterion is body balance, f_b. We fuse the arms and hand joints from the input sign gesture segment into a given candidate sign dance segment. Then, we evaluate the dynamic stability according to [Ho and Shum 2013]. We first calculate the center of mass:

$$COM = \sum_{t=0}^{l_{total}} w_l p_l, \qquad (4)$$

where p_l is the 3D position of the l^{th} body part, l_{total} is the total number of parts, w_l is the estimated weight of the l^{th} body part, which is obtained according to [Armstrong 1988]. The center of gravity is obtained by setting the y component of the center of mass as 0 (i.e. ground height). We check if the center of gravity is within the supporting polygon, which is bounded by the supporting feet/foot, as visualized in Fig. 5. If the center of gravity goes beyond the supporting polygon, the posture is considered to be dynamically unstable. Finally, the body balance criterion is defined as:

$$f_b = \frac{F_{unstable}}{F_{total}}, \qquad (5)$$

where $F_{unstable}$ is the number of unstable frames, F_{total} is the total number of frames in the sign dance segment.

The final criterion is the motion connectivity, f_c. Again, we generate the fused sign dance segment as before, and evaluate if the first frame of the motion is similar to the final frame of the last synthesized sign dance segment. In our implementation, we evaluate first order connectivity using joint angles according to [Shiratori et al. 2006]. The motion connectivity criterion is defined as the sum of joint angle difference between the two frames:

$$f_c = \sum_{n=1}^{n_{total}} 2 \arccos \left| {q_{n,1}}^{-1} \cdot q_{n,2} \right|, \qquad (6)$$

where $q_{n,1}$ is the n^{th} joint represented in the quaternion of the last synthesized sign dance frame, $q_{n,2}$ is that of the first frame of the fused sign dance segment considering, n_{total} is the total number of joints.

The aforementioned criteria can be formulated as either optimization functions or hard constraints. A criterion can be modelled as an optimization function by maximizing/minimizing the evaluated value. It can also be modelled as a hard constraint by assigning a hard threshold. In our experience, inappropriate body balance and music/motion intensity tend to generate significant visual artifacts, and therefore are modelled as hard constraints. The sign dance expectation and motion connectivity criteria are softer, and therefore are modelled as optimization functions:

$$\arg\min(w_e f_e + w_c f_c) \quad \text{subjected to } f_c < T_c, f_i < T_i, \qquad (7)$$

where $w_e = 1.5$ and $w_c = 1.0$ are empirically set weights, $T_c = 0.5$ and $T_i = 0.1$ are empirically set thresholds.

6 EXPERIMENTAL RESULTS

In this section, we evaluate our proposed system with both qualitative and quantitative experiments.

All the experiments were performed on a computer with an Intel i7-3770 CPU, an AMD Radeon HD 7770 graphics card and 8GB of RAM. The sign dance database consists of 10 pieces of music. Each piece of music was 89.7 seconds long on average. It took 220 seconds to train the Sign Dance Model, and 57.4 seconds on average to synthesize the sign dance of one piece of music. In other words, our system ran faster than real-time on a low-end computer, with a slight delay when generating the first sign dance segment.

We used the MMD character model following the guideline set by Crypton Future Media, Inc. For more information, please refer to http://piapro.net/en_for_creators.html

6.1 Qualitative Evaluations

To evaluate our system, we synthesized sign dance based on the input songs and sign gestures.

To assess how well the Sign Dance Model generalizes the knowledge in the sign dance database, when synthesizing the sign dance motion for a music, we excluded the tuples related to that music in the database. In other words, both the Sign Dance Model and the database had no knowledge of the piece of music and its corresponding sign dance. We set up the experiment as such to ensure that the system could not simply re-display the captured sign dance for a given song. Instead, as the system had no prior information for the input music, it had to rely on knowledge generalized from all the other pieces of music to synthesize the current one.

Fig. 6 shows some examples of the synthesized sign dance, in which the orange arrows indicate overall body movement. In general, the style of the dance follows that of the captured sign dance very well. The character utilizes body movement to emphasize the

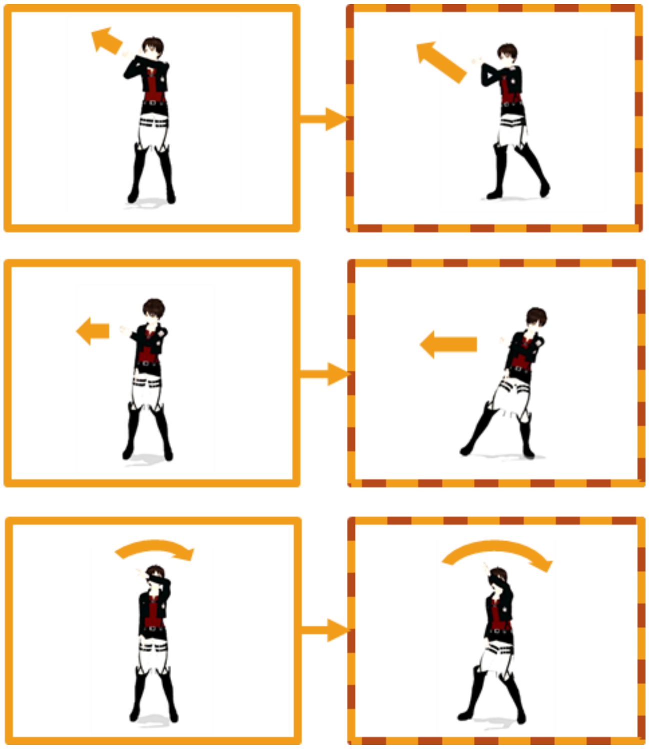

Figure 6: Examples of input sign gestures (yellow outline) and the corresponding synthesized sign dance (yellow and red dotted outline).

sign language. For example, when the sign language is about lowering the arms, the character picks a squatting motion and create the sign dance. Overall, our system creates natural dance movement in which the sign gesture and the body dance movement align very well.

We also compared the synthesized sign dance with the captured sign dance. As mentioned before, the experiment setup here excluded the captured sign dance motion of the music that was currently considered to access the system performance with "unseen" music. In such a setup, it was impossible to generate exactly the same sign dance motion as captured, as those motion segments did not exist in the database. However, in most situations, our system successfully picked alternative sign dance motions that were similar to the captured ones, and generated sign dances that were similar to the captured motions. Fig. 7 shows some comparisons between the captured sign dances and the synthesized ones. The orange arrows indicate overall body movement.

The readers are referred to the supplementary video for more results of synthesized sign dance.

6.2 Quantitative Evaluations

To quantitatively evaluate our system, we conducted user studies and set up an accuracy analysis experiment. We invited 12 participants to evaluate our system. They were between 22 and 26 years old. 5 of them had dance experience, and the other 7 had no experience in. We compared our synthesized sign dances (i.e. generated

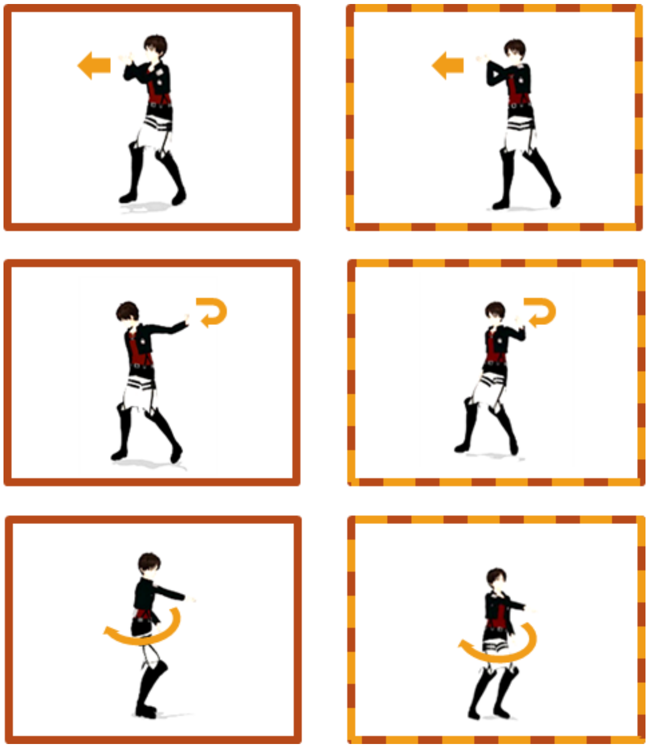

Figure 7: Comparison between captured sign dance (red outline) and synthesized sign dance (yellow and red dotted outline) for the same music segments.

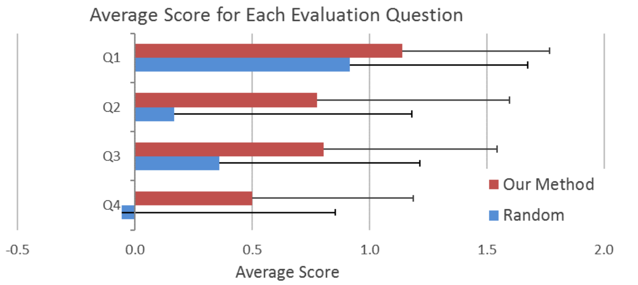

Figure 8: Comparison between our method and random dance generation. Red and blue bars represent the score of our method and the random synthesis respectively for different questions. Solid lines represent standard derivations.

using the same experiment set up as the experiment in Section 6.1) with a baseline random system (i.e. generated by randomly picking sign dance segments in the database). We showed both results one by one in random order to the participants, and asked them to answer the following questions with a five-points scale (2: yes, 1: maybe yes, 0: not sure, -1: maybe no, 1: no).

Q1: Does the synthesized sign dance match the hand gesture?

Q2: Does the intensity of the synthesized sign dance match that of the music?

Table 1: Accuracy analysis with leave-one-out cross-validation

Music	1	2	3	4	5	6	7	8	9	10	Average
Accuracy	36.0%	78.6%	71.4%	67.7%	78.3%	50.0%	73.1%	81.5%	23.3%	75.0%	**63.5%**

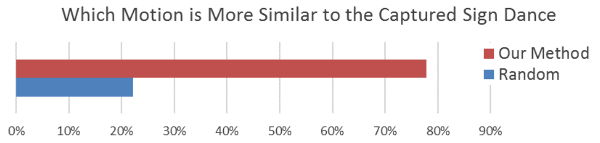

Figure 9: Evaluating which motion is more similar to the captured sign dance. Red and blue bars represent the percentage of participant choosing our method and the random synthesis respectively.

Q3: Is the synthesized sign dance natural?

Q4: Is the synthesized sign dance similar to the captured sign dance (both synthesized and captured sign dance are displayed together for this question)?

Fig. 8 shows the results of the questionnaire. Our method consistently outperforms the baseline method. The average score of our method across all questions (0.81 points) is significantly higher than that of the baseline method (0.34 point). Furthermore, our method achieves a smaller average standard deviation across all questions (0.72 points) compared to the baseline method (0.88 points).

We also showed the sign dance synthesized by our method, the sign dance synthesized by the baseline method, as well as the captured sign dance all together to the participants, and asked the participants to indicate which of the two synthesized sign dances was more similar to the captured one. Fig. 9 shows the result. Without knowledge on how the sign dance is synthesized, 78% of the participants believe that the one synthesized by our method was more similar to the captured sign dance, in contrast to the 22% selecting the one synthesized by the baseline method. This shows that our method successfully learns the dancing style from captured motion data.

To further verify that our Sign Dance Model can model the dancing style from captured sign dances, we performed leave-one-out cross-validation on our system. We set up the experiment as follows. When training the Sign Dance Model, we excluded all information of the piece of music we wished to synthesize from the database. When synthesizing the sign dance, we utilized the full database consisting of the tuples from all pieces of music. We then checked if the Sign Dance Model trained by excluding the considered song could pick the sign dance motion segments of such a song. The accuracy was calculated as the percentage of sign dance segments selected that are exactly the same as those in the captured sign dance.

Tab. 1 shows the accuracy of individual pieces of music. We achieve a high average accuracy of 63.5%, indicating that our Sign Dance Model generalize the style of the captured dance motion very well. Music 1 and 9 performs relatively poor. We believe that this is because those two pieces of music had a high tempo, while the rest of the music did not. Therefore, when excluding the music in the leave-one-out cross-validation, there may not be enough music of similar style to train the Sign Dance Model. We believe that a larger database could solve this problem.

7 CONCLUSIONS AND DISCUSSIONS

This paper presented a data-driven system to generate sign dance automatically, which combines gesture-based sign language and full body dance motion. Our system utilizes multiple regression analysis to model the correlations between dance motion and hand gestures/music. This information is applied to synthesize a new sign dance from a piece of input music and the corresponding sign gesture, utilizing a set of objective functions that evaluate motion quality. Our system is robust and computationally inexpensive, making it suitable for interactive animation creation and real-time dancing characters in games. It opens up a new direction of music visualization that allows people with hearing difficulties to understand and enjoy music.

Our implementation of multiple regression analysis implicitly assumes that there is a globally optimal mapping between hand gesture and body dancing motion, which may not necessary be true as the same hand gesture may be represented by different dance moves depending on the dancing style. One direction for future work is to research a proper representation of dancing style, and explore possible regression or machine learning algorithms that can handle multiple local mappings. This allows the system to map hand gestures and dancing motion optimally in the corresponding local space according to the dancing style.

Our system input includes a stream of gesture-based sign language performed by a human sign language translator. One possible extension of this work is to combine automatic sign gesture synthesis systems such as [Baldassarri et al. 2009; Hirumura et al. 2015], such that the system can synthesize both sign language and body dancing motion from the input music and the corresponding lyrics.

Due to capturing difficulties, our gesture database does not include finger information. Only a few high-end data gloves in the market can capture the fine finger movement required for sign language, and the data processing involved is labour intensive. In theory, finger information could enhance the mapping quality between gesture and dance motion under our multiple regression analysis model. We are interested in augmenting an existing sign gesture database into our captured motion.

Currently, we handcraft the sign gesture, sign dance and music features used for the Sign Dance Model. Such handcrafted features require domain knowledge and may not be optimally selected. We are interested in exploring feature extraction algorithms and dimensional reduction algorithms to obtain a better set of features.

The results of the work demonstrate the possibility of introducing hand gesture into other domains such as music and dancing. The traditional entertainment industry does not consider people with hearing difficulty heavily. We hope that this research can cause the interest in the field and enhance the quality of life for people in need.

ACKNOWLEDGMENTS

The project was supported in part by the Royal Society (Ref: IES\R2\181024), JST ACCEL (Grant number: JPMJAC1602), and JSPS KAKENHI (Grant number: JP17H06101 and JP19H01129).

REFERENCES

Harry G. Armstrong. 1988. Anthropometry and Mass Distribution for Human Analogues. Volume 1. Military Male Aviators.

Wakana Asahina, Naoya Iwamoto, Hubert P. H. Shum, and Shigeo Morishima. 2016. Automatic Dance Generation System Considering Sign Language Information. In *Proceedings of the 2016 ACM SIGGRAPH (SIGGRAPH '16)*. ACM, New York, NY, USA, 23:1–23:2. https://doi.org/10.1145/2945078.2945101

Sandra Baldassarri, Eva Cerezo, and Francisco Royo-Santas. 2009. Automatic Translation System to Spanish Sign Language with a Virtual Interpreter. In *Human-Computer Interaction – INTERACT 2009: 12th IFIP TC 13 International Conference, Uppsala, Sweden, August 24-28, 2009, Proceedings, Part I*. Springer Berlin Heidelberg, Berlin, Heidelberg, 196–199. https://doi.org/10.1007/978-3-642-03655-2_23

Jacky C.P. Chan, Jeff K.T. Tang, and Howard Leung. 2013. Synthesizing Two-character Interactions by Merging Captured Interaction Samples with their Spacetime Relationships. *Computer Graphics Forum* 32, 7 (2013), 41–50. https://doi.org/10.1111/cgf.12210

J. C. P. Chan, H. Leung, J. K. T. Tang, and T. Komura. 2011. A Virtual Reality Dance Training System Using Motion Capture Technology. *IEEE Transactions on Learning Technologies* 4, 2 (April 2011), 187–195. https://doi.org/10.1109/TLT.2010.27

Yinpeng Chen, Hari Sundaram, and Jodi James. 2006. A Computational Estimate of the Physical Effort in Human Poses. In *Advances in Multimedia Modeling: 13th International Multimedia Modeling Conference, MMM 2007, Singapore, January 9-12, 2007. Proceedings, Part II*. Springer Berlin Heidelberg, Berlin, Heidelberg, 393–404. https://doi.org/10.1007/978-3-540-69429-8_40

Vidyarani M. Dyaberi, Hari Sundaram, Jodi James, and Gang Qian. 2004. Phrase Structure Detection in Dance. In *Proceedings of the 12th Annual ACM International Conference on Multimedia (MULTIMEDIA '04)*. ACM, New York, NY, USA, 332–335. https://doi.org/10.1145/1027527.1027604

Eleni Efthimiou, Stavroula-Evita Fotinea, Christian Vogler, Thomas Hanke, John Glauert, Richard Bowden, Annelies Braffort, Christophe Collet, Petros Maragos, and Jérémie Segouat. 2009. Sign Language Recognition, Generation, and Modelling: A Research Effort with Applications in Deaf Communication. In *Universal Access in Human-Computer Interaction. Addressing Diversity*, Constantine Stephanidis (Ed.). Springer Berlin Heidelberg, Berlin, Heidelberg, 21–30.

Sylvie Gibet, Nicolas Courty, Kyle Duarte, and Thibaut Le Naour. 2011. The SignCom System for Data-driven Animation of Interactive Virtual Signers: Methodology and Evaluation. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 6 (Oct. 2011), 23 pages. https://doi.org/10.1145/2030365.2030371

Yu Higuchi. 2014. *MikuMikuDance*. http://www.geocities.jp/higuchuu4/index_e.htm

Toshihiro Hirumura, Nobuyuki anbd Shimizu, Shuichi Umeda, Naoto Kato, Taro Miyazaki, Seiki Inoue, Hiroyuki Kaneko, and Yuji Nagashima. 2015. Automatic Generation System of Japanese Sign Language (JSL) with CG Animation of Fixed Pattern Weather Infor-mation. *ABU Technical Journal* 264 (oct 2015), 2–5.

Edmond S. L. Ho, Jacky C. P. Chan, Taku Komura, and Howard Leung. 2013. Interactive Partner Control in Close Interactions for Real-time Applications. *ACM Trans. Multimedia Comput. Commun. Appl.* 9, 3, Article 21 (July 2013), 19 pages. https://doi.org/10.1145/2487268.2487274

Edmond S. L. Ho and Hubert P. H. Shum. 2013. Motion Adaptation for Humanoid Robots in Constrained Environments. In *Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA '13)*. IEEE, 3813–3818. https://doi.org/10.1109/ICRA.2013.6631113

Daniel Holden, Jun Saito, and Taku Komura. 2016. A Deep Learning Framework for Character Motion Synthesis and Editing. *ACM Trans. Graph.* 35, 4, Article 138 (July 2016), 11 pages.

Ludovic Hoyet, Kenneth Ryall, Katja Zibrek, Hwangpil Park, Jehee Lee, Jessica Hodgins, and Carol O'Sullivan. 2013. Evaluating the Distinctiveness and Attractiveness of Human Motions on Realistic Virtual Bodies. *ACM Trans. Graph.* 32, 6, Article 204 (Nov. 2013), 11 pages. https://doi.org/10.1145/2508363.2508367

Eugene Hsu, Sommer Gentry, and Jovan Popović. 2004. Example-based Control of Human Motion. In *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '04)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 69–77. https://doi.org/10.1145/1028523.1028534

Kyunglyul Hyun, Kyungho Lee, and Jehee Lee. 2016. Motion Grammars for Character Animation. *Comput. Graph. Forum* 35, 2 (May 2016), 103–113. https://doi.org/10.1111/cgf.12815

Naoya Iwamoto, Takuya Kato, Hubert P. H. Shum, Ryo Kakitsuka, Kenta Hara, and Shigeo Morishima. 2017. DanceDJ: A 3D Dance Animation Authoring System for Live Performance. In *Proceedings of the 2017 International Conference on Advances in Computer Entertainment Technology (ACE '17)*.

Olivier Lartillot, Tuomas Eerola, Petri Toiviainen, and José Fornari. 2008a. Multi-Feature Modeling of Pulse Clarity: Design, Validation and Optimization.. In *ISMIR* (2009-12-28), Juan Pablo Bello, Elaine Chew, and Douglas Turnbull (Eds.). 521–526.

Olivier Lartillot, Petri Toiviainen, and Tuomas Eerola. 2008b. A matlab toolbox for music information retrieval. *Data analysis, machine learning and applications* (2008), 261–268.

Kyungho Lee, Seyoung Lee, and Jehee Lee. 2018. Interactive Character Animation by Learning Multi-objective Control. *ACM Trans. Graph.* 37, 6, Article 180 (Dec. 2018), 10 pages. https://doi.org/10.1145/3272127.3275071

Christos Mousas, Paul Newbury, and Christos-Nikolaos Anagnostopoulos. 2013. Splicing of Concurrent Upper-body Motion Spaces with Locomotion. *Procedia Computer Science* 25 (2013), 348 – 359. https://doi.org/10.1016/j.procs.2013.11.042 2013 International Conference on Virtual and Augmented Reality in Education.

Nick Neave, Kristofor McCarty, Jeanette Freynik, Nicholas Caplan, Johannes Hönekopp, and Bernhard Fink. 2010. Male dance moves that catch a woman's eye. *Biology Letters* (2010). https://doi.org/10.1098/rsbl.2010.0619

Masaki Oshita and Yuta Senju. 2014. Generating Hand Motion from Body Motion Using Key Hand Poses. In *Proceedings of the Seventh International Conference on Motion in Games (MIG '14)*. ACM, New York, NY, USA, 147–151. https://doi.org/10.1145/2668084.2668095

Xue Bin Peng, Glen Berseth, and Michiel van de Panne. 2016. Terrain-adaptive Locomotion Skills Using Deep Reinforcement Learning. *ACM Trans. Graph.* 35, 4, Article 81 (July 2016), 12 pages. https://doi.org/10.1145/2897824.2925881

Michalis Raptis, Darko Kirovski, and Hugues Hoppe. 2011. Real-time Classification of Dance Gestures from Skeleton Animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA '11)*. ACM, New York, NY, USA, 147–156. https://doi.org/10.1145/2019406.2019426

Helge Rhodin, James Tompkin, Kwang In Kim, Edilson de Aguiar, Hanspeter Pfister, Hans-Peter Seidel, and Christian Theobalt. 2015. Generalizing Wave Gestures from Sparse Examples for Real-time Character Control. *ACM Trans. Graph.* 34, 6, Article 181 (Oct. 2015), 12 pages. https://doi.org/10.1145/2816795.2818082

Yijun Shen, Longzhi Yang, Edmond S. L. Ho, and Hubert P. H. Shum. 2019. Interaction-based Human Activity Comparison. *IEEE Transactions on Visualization and Computer Graphics* (2019), 16. https://doi.org/10.1109/TVCG.2019.2893247

Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. 2006. Dancing-to-Music Character Animation. *Computer Graphics Forum* 25, 3 (2006), 449–458. https://doi.org/10.1111/j.1467-8659.2006.00964.x

Hubert P. H. Shum, Taku Komura, and Shuntaro Yamazaki. 2012. Simulating Multiple Character Interactions with Collaborative and Adversarial Goals. *IEEE Transactions on Visualization and Computer Graphics* 18, 5 (May 2012), 741–752. https://doi.org/10.1109/TVCG.2010.257

Wataru Takano and Yoshihiko Nakamura. 2015. Construction of a space of motion labels from their mapping to full-body motion symbols. *Advanced Robotics* 29, 2 (2015), 115–126. https://doi.org/10.1080/01691864.2014.985611

Jeff K. T. Tang, Jacky C. P. Chan, and Howard Leung. 2011. Interactive Dancing Game with Real-time Recognition of Continuous Dance Moves from 3D Human Motion Capture. In *Proceedings of the 5th International Conference on Ubiquitous Information Management and Communication (ICUIMC '11)*. ACM, New York, NY, USA, Article 50, 9 pages. https://doi.org/10.1145/1968613.1968674

He Wang, Edmond S. L. Ho, Hubert P. H. Shum, and Zhanxing Zhu. 2019. Spatio-temporal Manifold Learning for Human Motions via Long-horizon Modeling. *IEEE Transactions on Visualization and Computer Graphics* (2019), 12. https://doi.org/10.1109/TVCG.2019.2936810

Lars Wilke, Tom Calvert, Rhonda Ryman, and Ilene Fox. 2005. From dance notation to human animation: The LabanDancer project. *Computer Animation and Virtual Worlds* 16, 3-4 (2005), 201–211. https://doi.org/10.1002/cav.90

He Zhang, Sebastian Starke, Taku Komura, and Jun Saito. 2018. Mode-adaptive Neural Networks for Quadruped Motion Control. *ACM Trans. Graph.* 37, 4, Article 145 (July 2018), 11 pages. https://doi.org/10.1145/3197517.3201366

Yi Zhou, Zimo Li, Shuangjiu Xiao, Chong He, Zeng Huang, and Hao Li. 2018. Auto-Conditioned Recurrent Networks for Extended Complex Human Motion Synthesis. In *International Conference on Learning Representations*. https://openreview.net/forum?id=r11Q2SlRW