Making Them Dance

Jae Woo Kim[1], Hesham Fouad[2] and James K. Hahn[1]

[1]Institute for Computer Graphics, The George Washington University
801 22nd St. NW Suite T720B Washington, DC 20052
{zoltar|hahn}@gwu.edu
[2]VRSonic Inc. 2533 Wilson Blvd. Suite 200 Arlington, VA 22201
hfouad@vrsonic.com

Abstract

In this paper, we provide a method to generate a perceptually appropriate dance motion for an input music sound track. Our solution extracts musical features from input music and searches for a sequence of perceptually correlated motion segments from a dance motion database. We suggest a set of mapping criteria as well as motion features which are necessary to perform the mapping from music to dance motion.

Introduction

Motion control is an important topic in both robotics and computer animation. Robots or virtual characters are controlled using motion control algorithms based on kinematics, dynamics and control theories. Recently, motion capture data has been used to generate realistic human motion by the computer animation community. Using motion capture data, highly realistic motion can be generated because the motion is recorded by capturing real performance of human beings (Arikan & Forsythe 2002, Kovar et al. 2002, Lee et al. 2002, Li et al. 2002). One of the important issues in motion capture research is how to specify a user's intention or constraints to get a desired result.

Robots have been widely used by industry in assisting and in some cases replacing human beings in difficult or dangerous work environments. More recently, robots have been developed for entertainment purposes. These robots have the intelligence to recognize human faces and voices and have the ability to communicate with human beings. Utilizing advanced control technologies, these robots are able to make smooth and dexterous movements and in some instance robots have been able to perform dance movements. Most of the research to date, however, has focused on developing techniques to address control problems such as balance control in response to environmental conditions. Little work has been done to create perceptually appropriate choreography in response to specific musical input. The approach we suggested in this paper can be applied to intelligent robots in order to generate pleasing dance choreography for entertainment purposes.

In this work, we are interested in making a robot or virtual human dance to recorded music. This is a problem of using musical cues, extracted through a music analysis process, as a specification for creating motion. In most cases humans dance not by improvising new motion, but by recombining several dance motions into a final dance routine (Shiratori et al. 2006). It is therefore possible to use a motion capture database consisting of prerecorded dance motions to synthesize a dance movement by searching and selecting a perceptually appropriate motion sequence from the database. By doing this, we can generate pleasing and entertaining dance motion that is synchronized to input music. One of the principal challenges in this approach is the problem of finding a mapping between musical sounds and corresponding motion.

This paper is organized as follows: In section 2, we briefly review related work. In section 3, we discuss our approach including music analysis, motion analysis, and the music to motion mapping scheme. In section 4 we discuss experimental results. Finally, in chapter 5 we provide our conclusions and future work.

Related Work

Much research on techniques to extract useful features from sound signals has been done in the fields of speech signal processing, non-speech sound signal processing, and musical signal processing. Linear Prediction Coefficients (LPC) and Mel Frequency Cepstral Coefficients (MFCC) are used in speech synthesis and recognition and they can also be useful features in representing musical signals (Davis et. al. 1980). Sound features related to the spectral shapes such as centroid, rolloff and flux have been used to perform content based audio classification and retrieval. These features are also useful to represent the timbre information of musical signals (Wold et al. 1996). Research on beat and tempo extraction for analyzing the rhythmic structure of music also has been done. Beat tracking has been done by estimating peaks and their strengths using autocorrelation techniques. Tzanetakis worked on musical feature extraction and genre classification of music. He used thirty musical features in analysis process to perform genre classification (Tzanetakis 2002).

Motion capture data are widely used in generating realistic motion because they guarantee the realism of the motion. Using stored motion data, motion graph techniques have been used to generate new sequences of motion by stitching many small motion segments according to input specifications and motion constraints. Motion constraints are specified as a set of key character poses, a path traveled by the character, or a user's motions in front of a video camera (Arikan & Forsythe 2002, Kovar et al. 2002, Lee et al. 2002, Li et al. 2002).

There have been several research efforts focused on generating motion based on input music. Cardle et al. extracted musical features from raw MIDI data as well as an analog rendition of the MIDI soundtrack and then applied different motion editing operations, such as motion signal processing, infinite impulse response (IIR) filters, and stochastic noise functions to alter the given motion data according to the musical features. (Cardle et al. 2002)

Shiratori et al. suggested a technique to synthesize a dance performance that is perceptually matched to input music. To perform the mapping between musical features to motion features, they extracted musical rhythm, structure and intensity and motion rhythm and intensity. Finally they synthesized a dance performance by matching the rhythm and intensity between music and motion. The framework of their solution is similar to ours, but in our work we used many more musical features and defined several useful motion features to improve the mapping process. Also our mapping scheme has more criteria which can lead the mapping process to generate more pleasing and natural results. (Shiratori et al. 2006)

Music Driven Dance Motion Generation

In our approach, musical analysis as well as motion analysis processes are used to extract useful information from both the input musical data and a set of motion clips in a motion capture database consisting of a variety of recorded dance motions. The musical analysis process extracts thirty musical features including beat, pitch and timbre information. It also analyzes the global structure of input music using a clustering algorithm. Motion analysis extracts three motion features including motion intensity, span, and density which represent the characteristics of the dance motion. Finally mapping is performed between the musical and motion features using a novel mapping algorithm described below.

Music Analysis

The musical analysis process extracts thirty separate features categorized into three parts: beat, pitch and timbre information. In order to produce a perceptually appropriate mapping between music and motion, the extracted musical features have to correlate well to the listener's perception of the music. We used observational analysis to validate this correlation.

Musical analysis is carried out on musical segments. The size of each segment is based on musical beat information where each segment consists of sixteen beats. Once feature extraction is performed on all segments, the segments are clustered into n clusters using a k-means clustering algorithm where the number of clusters n is given by the user. The ideal number of clusters used in the analysis process is related to the number of distinct patterns found in the music. Fig 1 contains a workflow depiction of this process: Input music is segmented into segments at every sixteen beats. For each segment thirty features are extracted resulting in musical feature vector. Each segment is clustered into n clusters using the k-means clustering algorithm based on the distance among the musical feature vectors of each segment. A cluster consists of a set of musical segments sharing common musical features. Using observational analysis we have been able to validate that indeed perceptually similar segments are placed in the same cluster. Finally, once the musical segments are assigned to clusters, each musical segment is labeled according to its owning cluster resulting in a symbolic representation of the musical piece such as "DCBECEC…". This sequence represents the global repeating patterns in the music.

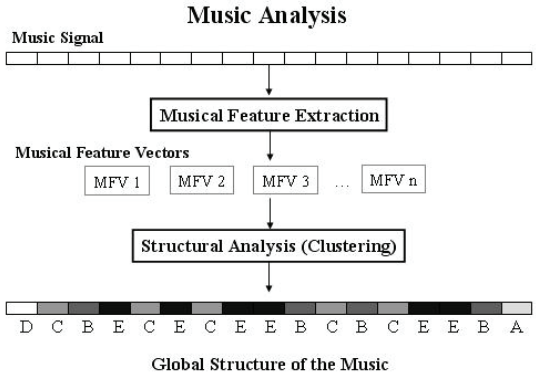
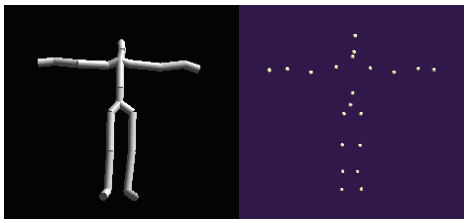


Fig. 1 Music Analysis

Motion Analysis

In this section we will discuss the motion representation used for motion analysis and we will define three motion features useful in representing the properties of motion.



(a) Human model (b) Point representation

Fig.2 Motion Representation

Motion Representation We used a Hierarchical Translation-Rotation (HTR) format for our motion capture

data. All of the motions used were captured from real dancers' performance. HTR format motion data consists of root (center of body) position, root orientation and the orientations of all other joint angles in a hierarchical fashion. While the HTR format is suitable for visualizing motion using graphics libraries, it is not well suited for performing motion analysis. We therefore transformed the motion data into a point representation.

A point representation is obtained by simple calculations on the limb lengths and orientation information obtained from HTR data format. (Figure 2 shows the human model and the corresponding point representation) This transformation transforms the orientations of each limb into a 3D location in the global Cartesian coordinates of each joint position. Using the point representation, it is simple to calculate the necessary motion features.

Motion Features Three motion features are used to describe motion properties. These features provide the cues necessary for mapping motion to music. In this section we will define each motion feature and explain its physical manifestation and how it is derived.

(1) Motion Intensity
Motion intensity represents the strength of a motion. It is obtained by approximating the linear summation of the momentum of each joint angle. (Shiratori et al. 2006) Motion intensity is calculated as follows:

$$I_{motion} = \sum_{i=2}^{n}\sum_{j=1}^{N} \| \mathbf{x_i}(j+1) - \mathbf{x_i}(j) \|$$

Here $x_i(j+1)$- $x_i(j)$ means the change of angle at joint i at frame j. n is the number of DOFs and N is the number of frames.

(2) Motion Span
Motion span represents the size of a motion based on the amount of space it spans. Motion span is calculated by considering how far each joint position travels in a motion segment. It is approximated by calculating the linear summation of the distance between each joint position relative to the calculated centroid of all the joint positions for that joint. Motion span is calculated as follows:

$$\mathbf{c_i} = \frac{1}{N}\left(\sum_{j=1}^{N} \mathbf{x_i}(j)\right)$$

$$S_{motion} = \sum_{i=2}^{n}\sum_{j=1}^{N} \| \mathbf{x_i}(j) - \mathbf{c_i} \|$$

where c_i is a vector representing the centroid of all the positions of joint i through the motion.

(3) Motion Density
Motion density represents how dense the motion is. When a joint moves a great deal in a relatively small region, we define this as dense motion. Conversely, if it moves little in a relatively large region we define this as sparse motion. Motion density can therefore be calculated as the ratio of motion intensity over motion span as follows:

$$D_{motion} = \frac{I_{motion}}{S_{motion}} = \frac{\sum_{i=2}^{n}\sum_{j=1}^{N} \| \mathbf{x_i}(j+1) - \mathbf{x_i}(j) \|}{\sum_{i=2}^{n}\sum_{j=1}^{N} \| \mathbf{x_i}(j) - \mathbf{c_i} \|}$$

Mapping

The mapping process is, in effect, a problem of searching and selecting the optimal motion sequence from the motion capture database which satisfies a set of mapping criteria. We define an evaluation function for each mapping criterion that is used to score how well a sequence of motion clips matches the input musical signal. The final score assigned to a sequence of motion clips is the sum of the scores obtained for each of mapping criterion. This evaluation process is performed on every combination of motion clips in the database for the input musical signal. The final motion sequence consists of the concatenation of the motion segments which has the highest score.

Our mapping scheme between music and motion is based on the following assumptions: First the musical beats and motion beats must be synchronized. Second, the analyzed music contains repeated patterns and when a specific pattern of the music repeats, the same corresponding pattern of motion which was previously matched to the music should be repeated. Third, there exists a perceptual correlation between musical features and motion features. Lastly, since human perception is especially sensitive to the changes of auditory or visual sensations, perceptual changes in the music should produce corresponding changes in the produced motion.

Music to motion mapping is performed based on a set of criteria that follow from the above assumptions. The first criterion is that musical beats must be matched to motion beats. In order to do this, we manually segmented the input musical data and all the motion clips in the database into segments containing sixteen beats each. In the future, this process can be automated by using the beat information extracted from the music and motion data. Once we have beat information, the music and motion beats are aligned by warping the motion data along to the time axis. The score assigned to a motion clip by the evaluation function for this criterion is directly related to the amount of warping necessary in order to align the beats. The reasoning for this is that extreme warping of the recorded motion can result in unnatural motion. If the warping exceeds more than 10% of the original motion, the evaluation function assigns the segment the minimum score **(**Shiratori et al. 2006).

The second criterion is that there should be a correlation between repeated musical patterns and corresponding motion patterns. In other words, if a musical pattern 'A' at time t_1 repeats again at time t_2, then we expect that the motion pattern 'a' which was assigned to music pattern 'A'

at t_1 repeats at time t_2. The evaluation function for this criterion calculates a similarity metric between the motion segments which are assigned to the occurrences of the same musical pattern.

The third criterion is that each music segment must have a perceptual correlation with the motion segment assigned to it. For example, a high intensity music segment should be assigned a motion clip which has high motion intensity (Shiratori et al. 2006). We currently only use sound and motion intensity in evaluating this criterion, but in the future other perceptually significant features will be added.

The last criterion is that the difference in auditory sensation at the boundary between two music segments must be matched to the difference in the visual sensation between the two motion segments. When music progresses from a pattern 'A' to another pattern 'B', we can get the difference in auditory sensation between the two patterns by calculating the distance between the seed positions of the two clusters in feature space. We can also calculate the difference of the visual sensation between two motion segments which have been assigned to each music segment by calculating the distance between two positions in motion feature space. The score assigned to a motion segment is therefore based on how closely the visual transition it produces matches the auditory transition between the music segment under consideration and the previous music segment.

Finally, we can define the evaluation function E of the mapping as follows:

$$E = \sum_{i=1}^{4} k_i E_{C_i}$$

Here, E_{Ci} represents the evaluation function for the criterion C_i (for $1 \le i \le 4$), and k_i is the weight value for each evaluation function. A user can control the importance of each evaluation function by adjusting these weight parameters.

Experimental Results

We performed experiments by applying the suggested approach to a set of music data in a database we constructed for the experiment. Our motion database consisted of 40 dance motion sequences. We applied the suggested evaluation functions to evaluate the mapping from the input music to dance motion. We could generate quite interesting and pleasing dance motion for the input music by selecting the motion sequence which has the highest score. We compared the result of our solution with (1) dance motion which has the lowest score, (2) dance motion which is randomly mapped, and (3) dance motion which is mapped using only beat and intensity matching. The experimental results were promising where the motion generated using the suggested approach produced dance motion that was more natural and pleasing than the motion produced using any of the alternative scenarios (1-3).

Discussion and Future Work

In this paper, we defined motion features which can represent the properties of dance motion and suggested the mapping criteria for generating pleasing and perceptually appropriate dance motion sequence for input music.

Our motion database is simply constructed with motion clips of choreographic motions but we can construct a motion graph by cutting the motion sequences into small pieces and stitching them by generating transitions among the pieces. We can generate a larger variety of motions by using a motion graph where the problem is transformed into a graph search problem which finds the path along the motion graph which produces the highest evaluation score given input musical cues.

When motion graph approaches are used, the connectivity among the motion segments will be an important issue so that a continuous motion sequence is generated. This will require an additional criterion that evaluates the connectivity of adjacent motion clips.

The complexity of a motion graph is an important factor which will affect the performance of the mapping algorithm. A large number of motion segments and transitions amongst them will generally produce better performance of the mapping algorithm. However, as the size of the graph increases, the time complexity of most graph search algorithms increases exponentially. One possible solution is the use of genetic algorithms to search the optimal path along the motion graph. Genetic algorithms are much less sensitive to graph size than other graph search techniques. We should mention that there does not exist an optimal solution to this problem. However, given the subjective nature of the problem, near optimal solutions may suffice. Genetic algorithms are very good at searching near optimal solutions in reasonable computing times. We can encode a sequence of motion as a chromosome and evolve the genetic algorithm to search the optimal path in the motion graph and the evaluation function we defined can be used in the evolution process.

References

Arikan, O., and Forsythe, D. 2002. *Interactive motion generation from examples*. In Proceedings of ACM SIGGRAPH 2002, Annual Conference Series, ACM SIGGRAPH.

Cardle, M., Barthe, L., Brooks, S., and Robinson, P. 2002. *Music-Driven Motion Editing: Local Motion Transformations Guided by Music Analysis*. In Proceedings of Eurographics 2002.

Davis, S., and Mermelstein, P. 1980. *Experiments in syllable-based recognition of continuous speech*. IEEE Transactions on Acoustics, Speech and Signal Processing, 28:357–366.

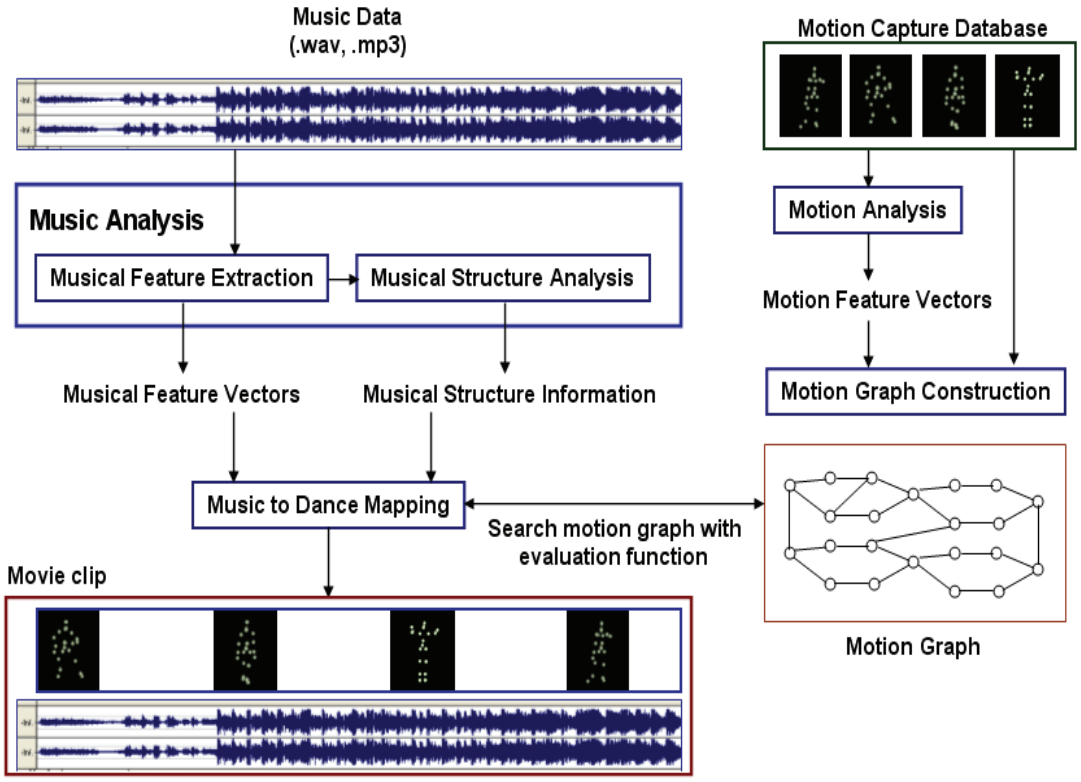


Fig 3. System Overview

Goldberg, D. 1989. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Boston, Massachusetts.: Addison-Wesley.

Kovar, L., Gleicher, M., and Pighin, F. 2002. *Motion graphs*. In Proceedings of SIGGRAPH 2002, Annual Conference Series, ACM SIGGRAPH.

Lee, J., Chai, J., Reitsma, P. S. A., Hodgins, J. K., and Pollard, N. S. 2002. *Interactive control of avatars animated with human motion data*. In Proceedings of ACM SIGGRAPH 2002, Annual Conference Series, ACM SIGGRAPH.

Li, Y., Wang, T., and Shum, H.-Y. 2002. *Motion texture: A two-level statistical model for character motion synthesis*. In Proceedings of ACM SIGGRAPH 2002, Annual Conference Series, ACM SIGGRAPH.

Makhoul, J. 1975. *Linear prediction: A tutorial overview*. Proceedings of the IEEE, 63:561–580.

Shiratori, T., Nakazawa, A., and Ikeuchi, K. 2006. *Dancing-to-Music Character Animation*. In Proceedings of Eurographics 2006.

Tzanetakis, G. 2002. *Manipulation, Analysis and Retrieval Systems for Audio Signals*. Ph.D dissertation Princeton University.

Wold, E., Blum, T., Keislar, D., and Wheaton, J. 1996. *Content-based classification, search and retrieval of audio*. IEEE Multimedia, 3(2).