Phrase Structure Detection in Dance

Vidyarani M. Dyaberi Hari Sundaram Jodi James Gang Qian
Arts Media and Engineering Program
Arizona State University
e-mail: {vidyarani.dyaberi, hari.sundaram, jodi.james, gang.qian}@asu.edu

Abstract

This paper deals with phrase structure detection in contemporary western dance. Phrases are a sequence of movements that exist at a higher semantic abstraction than gestures. The problem is important since phrasal structure in dance, plays a key role in communicating meaning. We detect two fundamental dance structures – ABA and the Rondo, as they form the basis for more complex movement sequences. There are two key ideas in our work – (a) the use of a topological framework for deterministic structure detection and (b) novel phrasal distance metrics. The topological graph formulation succinctly captures the domain knowledge about the structure. We show how an objective function can be constructed given the topology. The minimization of this function yields the phrasal structure and phrase boundaries. The distance incorporates both movement and hierarchical body structure. The results are excellent with low median error of 7% (ABA) and 15% (Rondo).

Categories and Subject descriptors

I.5.1 [**Models**]: *Structural*, I.5.4 [**Applications**]: *Computer Vision*, I.5.2 [**Design Methodology**]: *Classifier design and evaluation, Feature evaluation and selection, Pattern analysis,* J.5 [**Arts and Humanities**]: *Performing arts (e.g., dance, music)*

General Terms

Algorithms, Design, Human Factors

Keywords

Dance, phrase, structure, topological graph

1. INTRODUCTION

This paper deals with the problem of detection of phrasal structures in dance. The problem is important since structure plays an important role in representing and synthesizing meaning in dance [1].

There has been prior research on gesture boundary segmentation and detection but limited work on phrase detection. In [5], the authors have developed a real–time system that can be used for posture recognition in dance. Work in [2], dealt with gesture segmentation using a dynamic hierarchical layered structure to model the human body and activity measures in human body segments to find gestures in a motion sequence. Note that *a phrase is a sequence of movements, that exists at higher level of semantic abstraction than gestures*.

MM'04, October 10–16, 2004, New York, New York, USA.

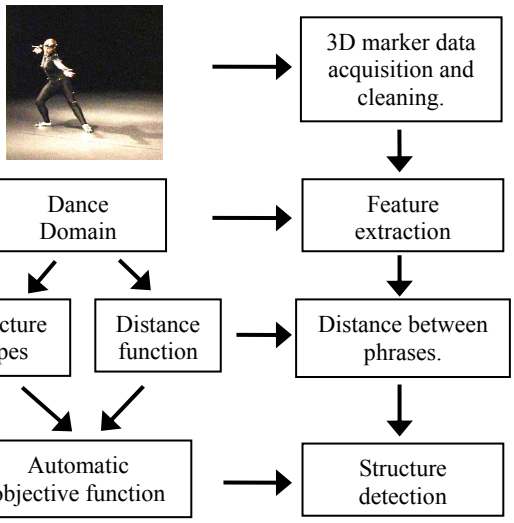

Figure 1: The system overview

We solve the problem of phrasal structure detection in the following way. First, we identify fundamental structures in cotemporary western dance – ABA and the Rondo. Then we determine robust features (kinetic energy, momentum, and force) that incorporates both the hierarchical body structure and segment movement. The distance between two phrases is computed using dynamic programming, since the phrases are of unequal length due to human variation.

We present the idea of a topological graph [6] for phrasal structure detection. Given the associated topological matrix, we show how to derive an objective function whose minimization shall determine the phrase change boundaries, thus confirming the presence of structure. The dances for the experiments, was created by an expert dancer (one of the authors) and was acquired using an eight camera VICON motion-capture setup. The results show that algorithm is robust – the structures are detected with very low median error 7% (ABA) and 15%. (Rondo). Figure 1 provides a system overview.

The rest of this paper is organized as follows. In the next section we review the role of compositional structures in dance. In section 3 discusses the process of data acquisition and cleaning, section 4 we discuss the creation of the composite feature while in section 5 we describe how distances between phrases are computed. Section 6 presents the topological matrix for structure detection. Section 7 presents the experimental results and section 8 summarizes the work and discusses future work.

2. COMPOSITIONAL STRUCTURES

In this work, we are focusing on the compositional structures that exist at the level of a phrase. Compositional structures are classical frameworks used to determine the overall structure of an

Figure 2: Movements showing phrase A (above) and movements showing phrase B (below)

entire dance piece [1]. It is possible to create dances without such compositional devices. However, successful dance composition usually requires structure of some kind, as does a good novel.

A phrase is the smallest and the simplest unit of form [1]. It is a short but complete unit in that it has a beginning, middle and an end. A phrase is to a dance as a sentence is to a book. Just as a sentence comprises of separate words, a phrase is made up of individual movements. The movements share common element of intent and communicate meaning (feelings / intentions etc).

The simplest form is AB (ref. Figure 2): a theme (phrase A) and a contrasting theme (phrase B). For example if A represents day, then B being a contrasting theme represents night. A and B share a common arena, but form opposing or contrasting perspectives. The going from A to B requires some sort of transition to provide the connecting link, the bridge.

In this paper we are interested in the following structures – (a) ABA : an extension of the AB form and (b) Rondo (ABACADA): A further extension where B, C and D are phrases distinct from phrase A. It has a basic theme A that keeps returning.

These are fundamental structures in ballet and contemporary western dance – they are taught in beginning dance composition classes as springboards from which to create more complex structures like a collage. For the detection of these structures we need to do data acquisition, data cleaning, feature extraction, distance calculation and objective function formulation. We discuss the above steps in the next few sections.

3. MOVEMENT DATA ACQUISITION

In this section we describe the data acquisition framework and discuss feature extraction. Data was acquired the 3D Marker-Based VICON motion capture system. This is an eight infra-red camera system with a capture frame rate of 120Hz. Forty-one 3D markers were placed on the dancer at specific locations (Figure 3), and were tracked. The captured marker data was then cleaned using a robust interpolation technique. Data-cleaning is an important problem since the three-dimensional nature of the human body causes marker self-occlusion of the 3D marker during motion capture. Features cannot be extracted correctly at all times, due to occlusion.

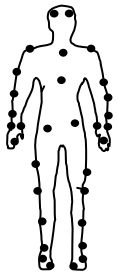

Figure 3: 3D marker positions

3.1 Feature Extraction

We use features that are invariant to the *marker location* – momentum, kinetic energy and force. This is important since the same phrase can be performed at different stage locations. The use of body mass of the dancer in all three features leads to more robust features. Figure 4 shows the plot of momentum for the a specific marker from a ABA sequence. The figure shows that that the momentum of the first and the third phrase is invariant to location. Note that the phrase lengths are different, due to variation in movement of the dancer.

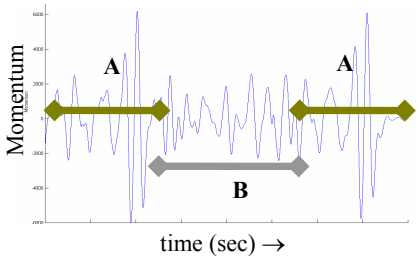

Figure 4: Plot of momentum vs. time, of Left Front Head marker showing the characteristic of ABA structure

4. FEATURE CONSTRUCTION

In this section we describe the *construction of the composite feature used for phrasal structure detection.* As described in section 3.1 we extract momentum, kinetic energy, and force using the motion of the 41 3D markers combined with the dancer body mass. We shall use a hierarchical body decomposition to derive a 42 dimensional composite feature vector that have the segment momentum, kinetic energy and force for 14 body segments. Next we describe the human body template and show how this knowledge is used to create the composite feature.

4.1 The Human-Body Template

The human body can be thought of as being composed of hierarchical physical segments, based on human anatomy [2]. We assume that each segment can move independent of one another. Segments in each layer (the parent segments) can have child segments that inherit the motion of their parent segment. For example the hand inherits the motion of the arm although it has motion of its own. We have considered the segments of the lowest layer of the hierarchy for the calculation of motion parameters as each of these segments have motion of their own and they also inherit motion of the segments in the layer above them.

4.2 Composite Feature

Here we present the composite feature derived using the hierarchical relationships between the different parts of the body. A total of 14 segments have been considered. The momentum, kinetic energy and force of every segment is calculated as the average of the markers that make up the segment. The relative mass of every segment is derived from equations extracted from standard ergonomic work [3]. Momentum, kinetic energy and the force of each segment were calculated of the body using the following equations:

$$P_k = \frac{1}{k}\sum_{i=1}^{k} m\,|\vec{v_i}|, \quad KE_k = \frac{1}{2*m}\sum_{i=1}^{k} P_i^2, \quad F_k = \frac{1}{k}\sum_{i=1}^{k}\frac{dP_i}{dt}, \quad <1>$$

where m is the segmental mass, v_i is the segmental velocity, P_k, KE_k and F_k are the segmental momentum, segmental kinetic energy and segmental force.

5. DISTANCE FUNCTIONS

In this section we shall show how the distances between the features and the phrases is computed. We shall use dynamic programming used to compute the distance between the phrases. The distance shall also incorporate the hierarchical layered template of human body. As seen in the previous section, 14 segments have been considered. The distance between the features is calculated using the Mahalanobis distance function.

$$d(x,y) = \sqrt{(x-y)^T \Sigma^{-1}(x-y)} \quad , \qquad <2>$$

where x and y are two composite features (both x and y are 14 dimensional vectors) and Σ is the covariance matrix.

5.1 Phrase distance

The distance between two phrases is computed using dynamic programming method. We use the Mahalanobis to calculate the distance between the composite features. *The two dance phrases are usually of unequal length even when it is repeated as in an ABA phrase.* We use the following constraints – (a) every sample in the phrase is used to compute the distance. (b) the distance calculation always goes forward in time. Our algorithm is as follows: Let is assume that n is the length of one phrase, m is the length of the other phrase, and G_i is the global distance upto *(i-1,j-1)*and the local distance at *(i,j)* is given by *d(i,j)*, that is calculated using the Mahalanobis distance, (ref. equation <2>). The global distance (i.e. the phrasal distance) is then calculated as:

$$G_{k+1} = \min[d(i,j), d(i-1,j), d(i,j-1)] + G_k \quad , \qquad <3>$$

where, k is the index for computing the temporary phrasal distance. The initial condition is $G_1 = d(1,1)$ and final condition $G_{k-1} = d(n,m)$ gives the smallest distance between the two phrases. All distances discussed here are normalized between 0 and 1.

6. STRUCTURE DETECTION

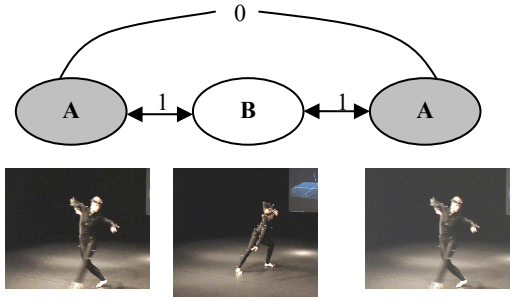

Figure 5: The ABA sequence represented using a topological graph

In this section we discuss the key idea behind the structure detection – the topological graph [6]. Then we shall show the automatic determination of objective function, given the topological graph.

6.1 The topological graph

Central to our structure detection algorithms is the idea of the topological graph. This structure has phrases at the nodes and the edge strengths are simply the distances between the nodes. Associated with each topological graph is a topological matrix constructed using the edge strengths of the graph. The topological graphs and matrices are domain dependent.

Let $\mathbf{S_d}$ be the metric space induced on the set of all phrases **P** in the dance sequence, by the distance function **d**. Then, the topological graph $T_G = \{V, E\}$ of a sequence of k phrases, is a fully connected graph, with the phrases at the vertices (V) and where the edges (E) specify the metric relationship between the phrases. The graph has associated with it, the topological matrix T_{MAT}, which is the k by k matrix where the entry $T_{MAT}(i, j)$ contains the strength (i.e. the distance between the two phrases corresponding to the nodes) of the edge connecting node i to node j in the graph. Note that the idea of the topological graph is distinct from directed acyclic graph used in macro-structure detection [4].

6.2 Objective function for ABA

The topological graph formulation helps us to automatically generate the objective function for detecting phrasal structure. For example the ABA structure that is a sequence of three phrases can be represented by a topological graph as shown in Figure 5.

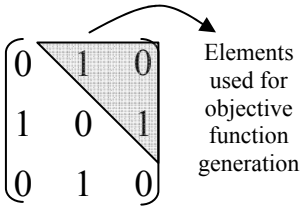

Figure 6

The ABA structure thus can be explicitly stated by the topological relationship: $d(A,B) = d(B,A) = 1$, $d(A,A) = 0$, where d is the distance between the phrases A and B. The distance '1' implies that phrases are distinct and distance '0' implies that the phrases are identical. Both the distance metric and the relationship between phrases are determined by the domain expert.

This topological graph has a topological matrix associated with it. The topological matrix is then a 3 by 3 matrix where the entry $T_{MAT}(i, j)$ contains the strength (i.e. the distance between the two phrases) of the edge connecting node i to node j in the graph. Figure 6 shows the topological matrix of the ABA structure and the upper triangular elements of the matrix that are considered for the generation of the objective function. Since the topological matrix is symmetric, we need consider only the upper triangular portion of the matrix. The topological matrix helps us formulate the objective function for the ABA sequence as follows:

$$O_{ABA}(\vec{t}) = d(A,A,\vec{t}) * (1 - d(A,B,\vec{t})) * (1 - d(B,A,\vec{t})),$$
$$\vec{t}^{\,*} = \arg\min_{\vec{t}} O_{ABA}(\vec{t}) \quad 0 < t_1 < t_2 < T, \quad , \qquad <4>$$

where $\vec{t}$ is just the vector (t_1,t_2). t_1 is the transition time from phrase A to phrase B and t_2 is the transition time from phrase B to phrase A, and where *the phrase boundaries* (t^*_1, t^*_2) *are detected by objective function minimization.* T is the duration of the entire phrase sequence. Similarly we can generate the topological matrix for Rondo which is the structure ABACADA, where A, B, C and D are all distinct phrases. Hence for a general case n phrase sequence, $T_{MAT}(i, j)$ entry of a n by n topological matrix is:

$$T_{MAT}(i,j) = d(i,j) = \alpha, \quad i,j : 1.....n, \qquad <5>$$

where $0 \le \alpha \le 1$ depends on domain and is provided by the expert of the domain.

6.3 Automatic Objective function generation

Given a topological matrix of an arbitrary structure, we can automatically generate the objective function whose minimization determines the phrase boundaries. The algorithm for setting up the objective function is thus as follows:

```
O = 1; Objective function initialization
for i = 1:(n-1)
    for j = 2:n
        O ← O * | T_MAT(i,j)- d(i,j)|
    end;
end;
```

where n is the number of phrases in the sequence, T_{MAT} (i, j) entry of a n by n topological matrix and $d(i,j)$ is the distance to be calculated. For the general case, number of elements in O is given by: $n(n-1)/2$. Since Rondo has 7 phrases, the objective function will have 21 elements. As in the ABA case, a constrained minimization of the objective function is used for structure detection; this also detects the phrase boundaries.

7. EXPERIMENTAL RESULTS

This section presents the experimental results and evaluations of our approach. The structure detection algorithm was implemented using MATLAB. The motion capture system was used to acquire the data for our experiments. It is an eight-camera VICON system with the frame rate of 120Hz. In our analysis, we have used five sequences of ABA and four sequences of Rondo that were choreographed and performed by one of the authors, who is an expert dancer. The phrase boundaries were marked by the dancer by looking at the motion captured data. Time was extracted at these phrase boundaries and thus determining the ground truth.

Table 1 shows the experimental results of ABA structure, where l_A and l_B are the lengths of the phrases, Δt_a is the average absolute time difference between the estimated and the ground truth values and E_a is the average error.

Table 1: Experimental results for ABA. The first two columns are phrase length in seconds and the last two show temporal error in seconds and percentage error.

Seq #	l_A	l_B	l_A	Δt_a	E_a
ABA1	12.39	13.17	15.20	1.02	**7.38%**
ABA2	18.3	15.13	15.03	0.87	**5.52%**
ABA3	12.9	16.32	13.9	1.03	**6.90%**
ABA4	13.44	13.67	13.50	0.61	**4.49%**
ABA5	12.29	11.42	15.31	1.82	**13.8%**

The error percentage is calculated as

$$E = \frac{100 * \Delta t * 2}{l_A + l_B}, \qquad <6>$$

where E is the percentage error for time calculation t_1. Δt_1 is the absolute difference between the ground-truth value and the estimated value. The experimental results for the Rondo (ABACADA) sequence are as follows:

Table 2: Experimental results for Rondo. The first seven columns are phrase lengths in seconds and the last two show temporal error in seconds and percentage error.

R	l_A	l_B	l_A	l_C	l_A	l_D	l_A	Δt_a	E_a
1	14.5	14.8	11.3	10.9	15	12.9	10	2	**14.9%**
2	11.2	12.8	11.4	12.0	12.9	13.0	12.4	1.9	**15.8%**
3	11.3	11.3	11.3	11.3	11.3	12.5	11.2	2.7	**23.6%**
4	12.1	12.1	12.1	12.1	12.1	12	12	1.9	**15.7%**

where l_A , l_B , l_C and l_D are the length of the phrases in seconds, and Δt_a and E_a are as explained before.

The performance of the algorithm is very good – the median error to detect the ABA phrase is around 7%, and the median error for the rondo sequence is around 15%. We believe that the difference is due to the higher complexity of the Rondo sequence. We believe that the results can be improved using more robust statistical methods, both for structure detection and data cleaning.

8. CONCLUSIONS AND FUTURE WORK

In this paper we have presented a framework for detecting the phrasal structures in dance. We specifically detect the ABA and the Rondo sequence – two fundamental structures in western contemporary dance.

We first discussed the importance of using structures in contemporary western dance – they are key to generating meaning in dance movement. We acquired movement data using 3D markers using the VICON motion-capture system. The data was subsequently cleaned, and the movement of each marker was extracted. This was combined with the hierarchical human body decomposition and three features were extracted per body segment – momentum, kinetic energy and force. We used 14 body segments. The distances between phrases were calculated using dynamic programming.

We proposed a topological graph framework to detect structure – this has phrases at each node, and the edges specify the relationships between the phrases. We showed how to compute an objective function based on the graph, and this is minimized to yield the phrase boundaries. We have excellent experimental results with low median error for ABA (7%) and Rondo (15%).

The algorithm presented in this paper leaves much room for improvement: (a) using statistical methods for data cleaning (b) we shall take into account the structural relationships in the human body. We are also planning to create a real-time implementation of our current framework.

9. REFERENCES:

[1] L. A. BLOM and L. T. CHAPLIN (1982). The intimate act of choreography. Pittsburgh, Pa., University of Pittsburgh Press.

[2] K. KAHOL, P. TRIPATHI and S. PANCHANATHAN (2003). *Gesture Segmentation in Complex motion sequences*, Proc. IEEE International Conference on Image Processing 2003, Barcelona, Spain, Sep. 2003.

[3] K. H. E. KROEMER, H. B. KROEMER and K. E. KROEMER-ELBERT (1997). Ergonomics: how to design for ease and efficiency, Prentice Hall.

[4] I.-J. LIN and S.-Y. KUNG (1997). *Coding and comparison of DAG's as a novel neural structure with applications to on-line handwriting recognition.* IEEE Transactions on Signal Processing **45**(11): 2701-2708.

[5] G. QIAN, F. GUO, T. INGALLS, et al. (2004). *A Gesture Driven Multimodal Dance System*, IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, June 2004.

[6] H. SUNDARAM (2002). *Segmentation, structure detection and summarization of multimedia sequences*: xxviii, 331 leaves, bound., Thesis Ph D --Columbia University 2002.