GESTURE SEGMENTATION IN COMPLEX MOTION SEQUENCES

Kanav Kahol, Priyamvada Tripathi, Sethuraman Panchanathan, Thanassis Rikakis*

Research Center for Ubiquitous Computing, Arts and Media Engineering
Department of Computer Science and Engineering
Arizona State University
Tempe, Arizona, 85287, USA
*kanav@asu.edu

ABSTRACT

Complex human motion sequences (such as dances) are typically analyzed by segmenting them into shorter motion sequences, called *gestures*. However, this segmentation process is subjective, and varies considerably from one human observer to another. In this paper, we propose an algorithm called *Hierarchical Activity Segmentation*. This algorithm employs a dynamic hierarchical layered structure to represent the human anatomy, and uses low-level motion parameters to characterize motion in the various layers of this hierarchy, which correspond to different segments of the human body. This characterization is used with a naïve Bayesian classifier to derive *creator profiles* from empirical data. Then those profiles are used to predict how creators will segment gestures in other motion sequences. When the predictions were tested with a library of 3D motion capture sequences, which were segmented by 2 choreographers they were found to be reasonably accurate.

1. INTRODUCTION

Human gesture analysis is important in many different application areas, such as human-computer interaction, surveillance, dance analysis and general motion analysis. To perform gesture analysis a continuous motion sequence must first be segmented into a sequence of discrete gestures. However, segmentation of gestures from motion sequences is not an easy task because: (1) gesture boundaries are subjective (2) boundaries are sequence dependent and, (3) it is impossible to enumerate all possible gestures

1.1Gesture boundaries are subjective

Gesture boundaries vary from one observer to the next for a given motion sequence. For example, Figures 1.1a and 1.1b show Body Force Graphs of a motion sequence as segmented by two different observers. As can be seen by comparing these two segmentations, the boundaries vary considerably. Observer A sees 10 discrete gestures and Observer B sees 21 discrete gestures.

1.2 Boundaries are sequence dependent

The perception of boundaries between sequential gestures depends upon the sequence of those gestures. Re-ordering a sequence of gestures can radically alter where the boundaries are perceived. Apparently, gesture segmentation is heavily influenced by higher-level cognitive processes, and by the context surrounding each gesture [2].

1.3 It is impossible to enumerate all possible gestures

Given the continuous nature of motion, there are an unlimited number of motion sequences that can be performed. Therefore it is impossible to enumerate a complete set of gestures. Gesture boundaries must often be arbitrarily defined, making it difficult to automate the gesture segmentation process.
Badler *et.al* [7] cited this ambiguity of gesture boundaries and hence questioned the usefulness of gestures as elemental units. However humans tend to perceive motion in terms of discrete gestures and so gesture segmentation has still been found to be useful for some applications, including the design of human computer interfaces, the notation of motion sequences, and simple animation.

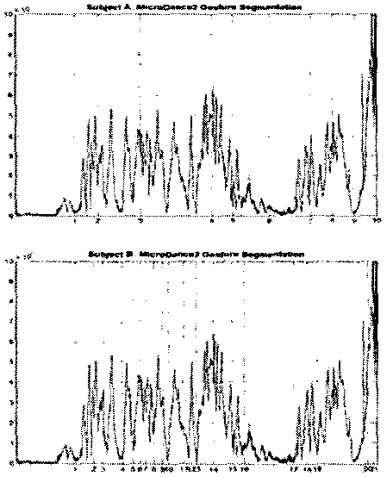

Figure 1.1 Segmentation Difference from Subject to Subject (Vertical lines depict gesture boundaries)
In the past gesture segmentation has typically been done using one of two methods: (1) Spatial Space Segmentation (2) Template matching techniques. However, both of these techniques are tedious, error-prone processes and produce results that might be interpreted as good or bad, depending upon the evaluator. Hence, there is a need for an adaptable, dynamic user-configurable algorithm that produces continuous estimates (in terms of probabilities) of gesture boundaries at all points in a motion sequence. In this paper we propose a novel segmentation

0-7803-7750-8/03/$17.00 ©2003 IEEE

algorithm that can dynamically segment meaningful gesture units by the use of an "activity" measure of the human body. This activity measure is calculated as a function three motion primitives: (1) momentum, (2) kinetic energy, and (3) force. A hierarchical layered approach is used to represent the human body, and activity measures are computed for various segments of the human body.

The gesture segmentation algorithm proposed in this paper uses (1) a dynamic hierarchical layered structure to model the human body and (2) activity measures in human body segments to find gestures in a motion sequence. The algorithm was tested on a library of 3D motion capture data. Twenty-five content rich motion sequences were segmented by five human subjects to provide ground truth. The algorithm achieved 86.5% average recognition accuracy per user when tested against this ground truth data

2. RELATED WORK

Human motion has been analyzed in terms of the body parts and trends in motion primitives over time. These motion primitives are then used to represent continuous motion sequences. Two types of approaches are usually used to characterize human motion sequences. (1) State-space approaches, which use points, line and 2D blobs, and (2) template matching approaches, which use overlaid meshes to characterize particular movement patterns. Some research has also been done using Hidden Markov Models (HMMs) [1], Neural networks, and particle filtering. [3] Attempts have also been made to automatically segment motion sequences into atomic units and then cluster them into meaningful sequences [10]. All of these methods look for local minima in low-level motion descriptors (such as velocity and acceleration) within continuous motion sequences. Motion segments defined by these local minima's are used to construct a *motion alphabet*. The individual segments that comprise this motion alphabet by themselves carry little meaning. However, gestures are composed of sequences of the alphabet segments, and *gesture specific recognizers* (such as HMMs or Neural Networks), which behave like state machines, and are used to find the boundaries between gestures. However, motion alphabets derived from one motion sequence generally do not produce satisfactory results when used to segment a different sequence. Also when motion sequences contain complex gestures (consisting of a long string of alphabetic sequences) the number of different alphabetic combinations can become astronomical, making a state machine approach impractical.

Attempts have been made to resolve this problem of linking low-level movements with cognitively coherent gestures. Piker *et.al.* [4] described activity in terms of two motion activity descriptors: (1) monotonous (steady) motion descriptors and (2) non-monotonous (unsteady) motion descriptors.

Research has also been done in the classification and representation of gestures. Ihara *et.al.* [6] provides a gesture-description model comprised of a semantic layer and a physical layer, and they classify a single semantic gesture as a 'fundamental gesture'. They base their research on the assumption that "Most of the gestures made by humans can be expressed by combining simple low level semantic gestures." Ihara *et.al.* used Labanotation to characterize higher-layer motion factors like shape, effort and weight from lower layer motion factors like velocity, which were derived from motion, capture data. These approaches are however based on (1) defining every single gesture in terms of a computer understandable form, which is an intensive task and (2) developing mapping models that can map low-layer data to high layer concepts of gestures. Also semantics of gesture may vary from one motion sequence to the next, and the idea of having low level gestures combining to form higher level gestures is not applicable to all motion sequences. It is arguable that a complex motion sequence is not merely a linear sequence of the atomic units. The various segments interact in layers analogous to the layers in computer animation. The depth at which meaningful units can be obtained is flexible and adaptable to the given sequence thus providing a better representation of the said motion.

In this paper, we propose a simpler approach that is based on a hierarchical layered template of human body, and use measurements of movement within each of these layers to find gesture boundaries.

3. HIERARCHICAL LAYERED TEMPLATE OF HUMAN BODY

The human body can be thought of as being composed of physical segments [3]. Each segment can move independently, and hence can exhibit an independent degree of activity. These segments can be represented in a hierarchy that is based on human anatomy. Figure 3.1 shows the hierarchical structure that has been used in conducting the research described in this paper. The lowest layer of this hierarchy is called layer 1, the next higher layer is called layer 2, etc.

Parent segments in each layer can have child segments in the layer below theirs. A parent segment inherits the aggregate characteristics of all of its child segments. For example, the momentum of a leg will be a vector sum of the momentum of upper leg, the lower leg and the foot. Child segments inherit the motion of the parent segment. Thus the foot inherits the motion of the leg, although it might also have a motion of its own.

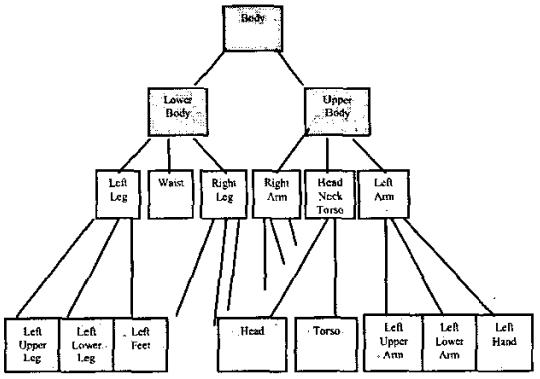

Figure 3.1: Human Body Hierarchy Model

In motion sequences there might be some time periods during which two adjacent segments have very similar motion vectors. When this happens these two segments are perceived as a single

segment. Also if the relative *orientation* of two adjacent segments does not change over a period of time, they are perceived as a single segment. However, most adjacent segments will move with some perceptible degree of independence, and will thus be perceived as separate segments. As an example consider the motion of walking. The upper arm, the lower arm and the hand will typically coalesce together, and be perceived as one segment: namely the arm. In contrast, the three leg segments (the upper leg, lower leg, and foot) will behave as separate segments and will thus be perceived as different segments, each with their own separate set of motion characteristics.

It should be noted that this process takes place dynamically. Thus, the number of perceived segments might change many times during a single motion sequence, and the corresponding body hierarchy diagram will also be dynamic. Because these dynamics are based on perceived motion, they provide important clues for gesture segmentation. In other words the dynamics of the constantly changing body segmentation provide a basis for predicting how human will perform gesture segmentation over time.

4. HIERARCHICAL ACTIVITY SEGMENTATION

Hierarchical Activity Based segmentation is an adaptable tool to find gesture boundaries based on activity measures in the various segments of the body hierarchy while executing a motion sequence. The *SegmentActivity* of each segment can be computed from three motion parameters associated with each segment – namely momentum, Kinetic energy and Force.

F (SegmentKE, SegmentForce, SegmentMomentum) (4.1)

The algorithm used for gesture segmentation consists of three steps: (a) Spatial Segmentation Derivation, (b) Segment Activity Derivation, and (c) Gesture boundary detection, which is based on the results of steps (a) and (b). The following sub-sections explain these three steps.

4.1 Spatial segmentation derivation

The marker set used for motion capture consists of at least three markers per segment in the lowest layer of hierarchy defined in Figure 3.1. Based on these three points, a unique plane can be computed for each of the segments in Layer 1, which is then used to compute the *orientation*, *velocity*, and *acceleration* of that segment. Frames are captured at a rate of 120 frames/sec, and angles between the planes of adjacent segments are calculated for every frame. Empirically determined ranges of angles specify bounds within which segments are deemed to coalesce into a single segment. In addition, if the angles between two adjacent segments do not change for more than 5 frames, then the segments are said to coalesce into a single segment.

4.2 Segment activity derivation

In every frame the velocity and the acceleration of every segmental plane is calculated. The relative mass of every segment is derived from equations extracted from the ergonomic literature [5]. Table 4.1 shows the equations that were used to compute the relative weights of various body parts.

The Segmental Mass computed with the equations in Table 4.1 is used to calculate the momentum, the kinetic energy and the force of each body segment in 3D using the following formulae:

*SegmentForce=SegmentMass*SegmentAcceleration (4.2)*
*SegmentMomentum=SegmentMass*SegmentVelocity (4.3)*
*SegmentKE=0.5*SegmentMass *SegmentVelocity (4.4)*

Segment	Segment mass [kg] as a function of the total body mass M [kg]
Head	0.0307 M + 2.46
Neck and torso	0.75(0.5640 M - 4.66)
Upper arm	0.0274 M - 0.01
Lower arm and hand	0.85*(0.0233 M - 0.01)
Hand	0.15*(0.0233 M - 0.01)
Thigh	0.1159 M - 1.02
Shank	0.0452 M + 0.82
Foot	0.0069 M + 0.47
Waist	0.25(0.5640 M - 4.66)

Table 4.1 estimated body segment masses (adapted from [5])

The idea behind using momentum, energy and force metrics is to employ a weighting mechanism for the propagation of motion parameter values from the lower layers to the upper layers of the hierarchy. For example, torso should contribute more to the upper body activity than hands should. If unweighted velocity or acceleration parameters are used to compute activity as has been done by some researchers [6] the resulting calculated activity does not correspond well to the perceived activity.

Higher Layer Segmental Force, Momentum, and Kinetic Energy of a segment are computed by taking the vector sum (or in some cases the scalar sum) of all of its child segments. Through this methodology a measure of activity in every segment of human body can be defined. During our experiments it was found that the higher frequency components of the body motion were not perceived by observers, so gaussian smoothing (low-pass filtering) was used to eliminate these components from the motion data, thus producing a smoother measure of activity that correlated better with human perception.

4.3 Gesture segmentation

In order to determine how humans segment dance motion human observers were shown video and 3D motion capture sequences (in the form of connected points of light) on a video display. Each sequence was of 2 to 3 minutes duration, and was called a *microdance*. Observers were asked to define the gesture boundaries within each microdance.

The 3D motion capture data for same microdance sequences were analyzed to compute the local minima in the force of the body (i.e. the highest layer in the hierarchy). Each of these local minima was then considered to be a potential gesture boundary.

At the moment of each of these local minima, the force, momentum, and kinetic energy parameters were examined for each of the lower body segments. For each segment a binary

triple (corresponding to whether each of these 3 parameters was or was not at a local minimum) was computed. For example, if the segment's force was at a local minimum, but the momentum and kinetic energy were not, the triple would be (100). Since there were 16 possible ways in which adjacent body segments could coalesce to form single segments, an additional 16 elements were included. This represents possible combinations of segments in the hierarchy and are set if the corresponding segments are behaving as one unit as determined by spatial segmentation derivation algorithm.

The resulting set of triples provides a complete characterization of all of the body segments at each instant when the body acceleration was at a local minimum. In other words, it provides a characterization at each potential gesture boundary.

This sequence of potential gesture boundaries was then compared with the gesture boundaries indicated by each of the human observers, and a naïve Bayesian classifier determined which of the potential gesture boundaries were chosen by each of the human observers. This provided an *observer profile* for each of these observers. This profile would then be used (in conjunction with the set of triples for a dance sequence) to predict where each observer would place gesture boundaries within that dance sequence.

5. RESULTS

In order to measure the accuracy of the prediction algorithm it was tested on a 3D data capture library. This library consists of 25 short motion sequences (2 to 3 minutes each) of dance and other normal day-to-day motions. Five observers were shown three of these library sequences, and marked the gesture boundaries in each sequence. Each of these sequences contains three iterations at different tempos. Using the observer profile generated from these 3 sequences, the prediction algorithm attempted to predict where these same observers would segment the remaining set of 20 motion sequences from the library. Table 5.1 shows results obtained. Training set was increased to 5 motion sequences to obtain an average accuracy of 87.9%.

Subject	Gesture Segmentation Accuracy (minimum)	Gesture Segmentation Accuracy (maximum)	Gesture Segmentation Accuracy (average)
1	79.80	89.00	86.70
2	82.50	91.10	87.40
3	77.60	87.90	84.60
4	80.10	89.40	86.24
5	81.80	90.20	87.60
Average	**80.36**	**89.52**	**86.58**

Table 5.1 results summary

5. CONCLUSIONS AND FUTURE WORK

The results of these experiments indicate that although different people segment motion sequences based on different criteria, each does so in a largely predicable manner. Thus, by analyzing an observer's segmentation of a small set of sequences, it is possible to predict how he/she will segment other sequences. The dynamic Hierarchical Activity Segmentation algorithm presented in this paper provides a characterization of body motion that is adequate to perform such a prediction. Incorrect predictions are currently being analyzed to find patterns that suggest improvements to the prediction algorithm. Early analysis indicates that performance could be improved by using a more elaborate body model that allows for more layers in the hierarchy. The current prediction algorithm only uses *scalar* values of force, momentum and energy for each body segment in its prediction of gesture boundaries. This simple method of prediction ignores the changes of direction in these parameters, which may be perceived as gesture boundaries. To take these changes into account, the model will be modified to predict segment boundaries based on *vector* values.

6. ACKNOWLEDGEMENTS

We would like to thank Raphael Motto and Siew Kong Wong who provided us with the library of 3D motion capture sequences. We would like to acknowledge Mary Fitzgerald, Jennifer Tsukayama and Todd Ingalls who helped us capture 3D motion capture sessions of dance pieces.

7. REFERENCES

[1] J. Hoey, and J.J. Little, "Representation and recognition of complex human motion", Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on , Volume: 1 , Page(s): 752 -759 vol.1, 2000

[2] J. Rose, *A multilevel approach to the study of motor control and learning*, Allyn and Bacon, 1997

[3] J.K., Aggarwal and Q. Cai, "Human motion analysis: a review", Nonrigid and Articulated Motion Workshop, 1997. Proceedings., IEEE , Page(s): 90 –102, 16 Jun 1997

[4] K.A Peker, A.A Alatan and A.N.Akansu, "Low-level motion activity features for semantic characterization of video", IEEE International Conference on , Volume: 2 , Page(s): 801 -804 vol.2, 2000

[5] K.H.E. Kroemer, H.B. Kroemer, K.E. Kroemer-Elbert. *Ergonomics: how to design for ease and efficiency*, Prentice Hall, 1997

[6] M Ihara, M Watanabi and K Nishimura, "A Gesture Description Model based on synthesizing Fundamental Gestures", Southeastcon '99. Proceedings. IEEE , Page(s): 47 – 52, 1999

[7] N. Badler, M. Costa, L. Zhao and D. Chi, "To gesture or not to gesture: what is the question?", Computer Graphics International, 2000. Proceedings , Page(s): 3 –9, 2000

[8] T Mori and K Uehara, "Extraction of Primitive Motion and Discovery of Association Rules from motion data", Robot and Human Interactive Communication, 2001. Proceedings. 10th IEEE International Workshop on , Page(s): 200 –206, 2001

[9] T. Darrell and A. Pentland, ``Robust Estimation of a Multi-Layer Motion Representation", in Proc. IEEE Workshop on Visual Motion, IEEE Computer Society Press, Princeton, NJ, 1991.

[10] T. Wang, H. Shum, Y. Xu, and N. Zheng, "Unsupervised Analysis of Human Gestures", IEEE Pacific Rim Conference on Multimedia 2001: 174-181, 2001