Robust Pause Detection Using 3D Motion Capture Data For Interactive Dance

Yi Wang, Gang Qian and Thanassis Rikakis
Arts, Media and Engineering Program
Arizona State University
Tempe, AZ 85287-8709

Abstract

In this paper, we present a robust pause detection algorithm for an interactive dance system using 3D motion capture data. The algorithm uses joint angle and shape context as two types of feature vectors obtained from both labeled and unlabeled data stream to detect pauses of the mover in the space. By looking at variances of the feature vectors over a time window, the probability of pause is computed. Satisfactory interactive dance performances have been successfully created and presented using the reported system.

1. Introduction

Interactivity is attracting tremendous attention in performing arts [1,2,3,4] For example, in interactive dance, movers can interact with a central control system through local and global body movements to control lights, trigger various background music as well as visual effects. We have previously reported an interactive dance system, where static poses (i.e. body shapes) were used as communication cues between the dancer and the interactive system, where a marker-based motion capture system was deployed as the major sensing equipment [4]. When the mover stops moving and pauses into a specific pose; and the recognition system sees this pose, a corresponding event will be triggered. Although the system reported in [4] has been working quite successfully, there are still many challenges in the pose recognition system. One of them is false alarms, which can happen when the mover is moving and temporarily go through a body shape which is similar to one the special poses. Realizing the fact that a true pose that the mover is using to talk with the system can only happen when the mover stops, detection of the pause can be performed before pose recognition to reduce false positives. Because a pose is the subset of a pause, through pause detection, stillness and moving data can be sent individually so that the gesture recognition engine can recognize poses more exactly without looking at misleading data in the period of movement. In addition to the reduction of false positives, pauses themselves actually can be used as another set of cues for the mover to interact with the system. Therefore, pause detection using data from marker-based motion capture systems needs to be addressed. In this paper, we report a simple, but robust ,pause detection algorithm.

2. Feature Selection and Extraction

Given a series of data frames with marker positions in the 3D space, the pause detection results can be obtained by looking at the variations of the feature vectors extracted from the frames over certain time period. The traditional thought about features is to use the absolute coordinates of current marker position in 3D space as feature vectors. It's simple and easy to implement, however, it is not appropriate in the application of pause detection in a dance performance. The reason is that some of the poses while dancing still contain small motion of some of the body parts. Thus, those pauses can not be detected robustly if absolute marker positions are used. Some unstable shaking of the body while holding poses could also cause false negatives.

Another thought that needs to be taken into account in feature selection is that the features needs to be robust to errors produced by the motion capture system during the motion capture session, such as unlabelled or mislabeled markers, or even missing markers caused by occlusion. Good labeling means that the markers' identities can be correctly recovered.

In our approach, we used both joint angle space and shape context as feature vectors, which can be extracted from both labeled and unlabeled data streams. When the labeling of the motion capture system is good, we calculate the 22 joint angles between different body parts, using a total number of 33 labeled markers. When the system labeling is poor, a feature vector in the shape context space is used. In this case, we require only three major torso markers to be correctly labeled in order to construct a torso-based local coordinate system. Then two virtual planes are generated on this coordinate system for the projections of unlabeled markers from unlabeled data stream. The shape contexts of two special local points in the torso coordinate system are mapped using histograms. In the joint angle space, the unstable shaking noise can be filtered out after smoothing. In the shape context, given a pause the projection shape between small time intervals should be the constant, combined with unlabeled data, the noise can be effectively reduced. Also, in our algorithm,

we clean the marker data by fixing the wrong data from the current labeled frame via comparison with the pervious frames in labeled stream. In the detection part, the timing windows are used to calculate the means and variances of feature vector over a certain time period. The matching probability of pause is computed.

2.1. Joint Angle Space

The torso orientation and joint angles between adjacent body parts are extracted to represent poses. We use 22 angles between nine body parts, including torso, upper arms, forearms, upper legs and low legs, excluding head. Table 1 shows the number of angles related to different body parts. Considering the errors introduced by mislabeled and unlabeled marker, the quality of current marker frame is evaluated using the marker positions of pervious frames. Since the frame rate is 60 frames/second, even very fast movement will not create large gaps between two successive frames. When a large gap is observed, the markers causing the gaps will be replaced by their values in the previous frame. After the labeled markers are cleaned, the angles between different markers are calculated. Hence, the joint angle feature vectors $V = \{V_1, V_2, V_3 \ldots V_{22}\}$ are constructed by marker feature vectors. Please note that although the marker cleaning algorithm is very rough and it can introduce errors in the joint angle calculation, it is not harmful to the detection of pauses. The reason is that when the mover is in a pause, the marker cleaning method is valid. When the mover is in motion, the joint angle calculated might be wrong, but it will not usually lead to a detection of pause, which is acceptable in this particular application of pause detection.

In the joint angle feature extraction procession, 33 global labeled markers $\{M_1^{(G)}, M_2^{(G)} \ldots M_{33}^{(G)}\}$ are transformed to torso coordinate $\{M_1^{(T)}, M_2^{(T)} \ldots M_{33}^{(T)}\}$ using $M_i^{(T)} = T_{T/G} M_1^{(G)} + O^{(T)}$, where $T_{T/G}$ is the transformation matrix. The feature vectors $\{V_1^{(T)}, V_2^{(T)} \ldots V_{22}^{(T)}\}$ of 22 angles are gained as the input of detection parameter.

Table 1. Joint Angle Space

Body Part	Upper Arms	Forearms	Upper Legs	Low Legs	Torso
Angle Number	6	6	4	4	2

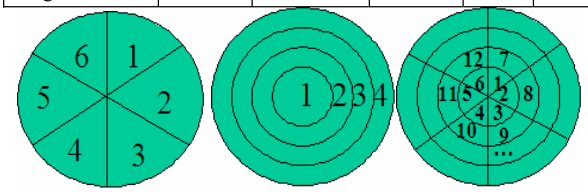

Figure 1. From left to right, the sector model, the shell model and the combined model are three basic space decompositions for shape histograms. Every bin is marked in the 2D spaces.

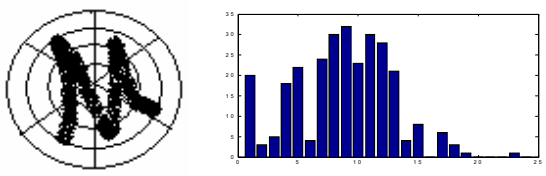

Figure 2. A 2D shape histogram

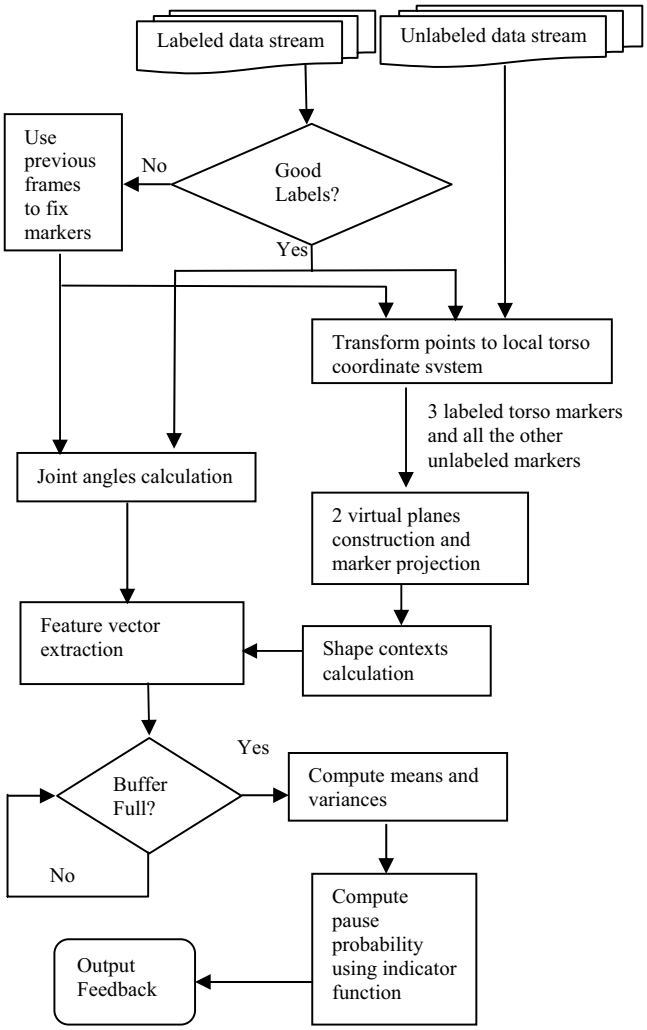

Figure 3. Pause Detection

2.2. Shape Context

Shape context is a shape descriptor to describe the coarse distribution of the rest of the points of the shape with respect to one point on that shape [5, 6, 7]. It is one of the most robust feature vectors in the feature-based shape-matching algorithm. The general idea is that, for each point in the shape, a descriptor is generated, which is called shape context, to express the configuration of the entire shape relative to that reference point. In the 2D case, the space is divided in a form of log-space, using

shells and sectors (See Figure 1 for examples), centered at a reference point. Then the numbers of the rest points fall in the bins are represented by histogram which is defined to be the shape context of this reference point. In a 2D shape of n points, when combined bins are used, the total number of bins is $N=\{\#shell \times \#sector\}$. For a point p, the corresponding histogram (or shape context) vector $\mathbf{h}(p)=(h_1(p), h_2(p),..., h_N(p))$ can be constructed so that $h_k(p)$ is the number points reside in the k^{th} bin of the shape context of point p. Figure 1 shows three types of bin configuration. In Figure 2, the histogram of shape '*M*' with the center as the reference point is given.

In the shape context feature extraction for pause detection, we first use C7, LBWT and RBWT (see Figure 4 for the location of the markers.) on the torso to construct local torso coordinate system. We then define two virtual planes $P^{(f)}$ (the frontal plane) and $P^{(s)}$ (the sagittal plane). The orientation and position of each plane is constant in the local torso coordinate system so that marker projections on the plane are constant shapes for a particular gesture. All 33 global unlabeled points $\{P_1^{(G)}, P_2^{(G)} ... P_{33}^{(G)}\}$ are transformed into torso coordinate $\{P_1^{(T)}, P_2^{(T)} ... P_{33}^{(T)}\}$ and their projections on the two virtual planes can be obtained. Let the projections in the local 2D space coordinate systems be $\{P_1^{(f)}, P_2^{(f)} \ldots P_{33}^{(f)}\}$(for frontal plane) and$\{P_1^{(s)}, P_2^{(s)} \ldots P_{33}^{(s)}\}$ (for sagittal plane).(See Figure 5 for the two projection planes.)

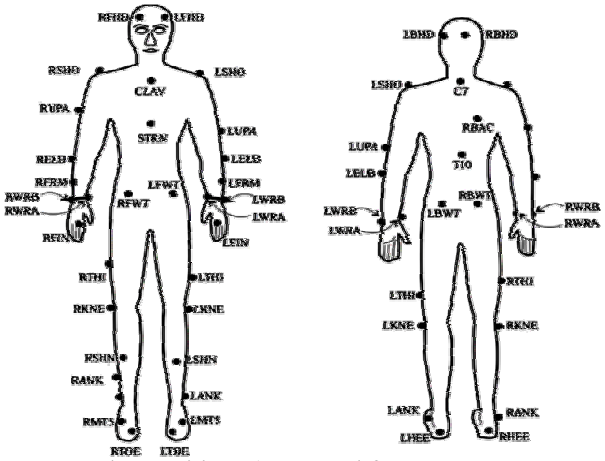

Figure4. Marker positions (excerpted from VICON user's manual)

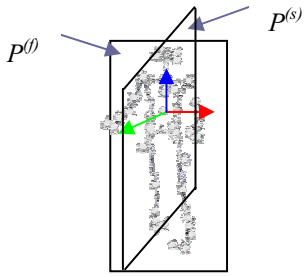

Figure 5. shape context features

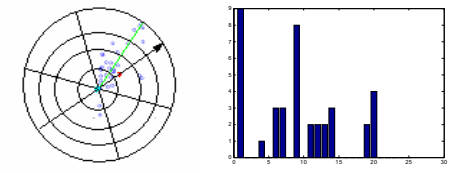

Figure 6. An example of shape context and the related histogram. The points in the left plot are the projections of makers onto the frontal plane.

The 2D local coordinate systems on the two planes are constructed as follows: first, the mean of the projections of the three torso markers on the two planes are used as the origin of the two planes, i.e. $O^{(f)}$ and $O^{(s)}$. Then the centroid of all marker projections on these two planes are computed and denoted as $C^{(f)}$ and $C^{(s)}$. Vectors $<O^{(f)}C^{(f)}>$ and $<O^{(s)}C^{(s)}>$ are then defined as the new x-axes of the two local plane coordinate systems. The bins are indexed in the 2D local coordinate systems and the coordinates of the marker projections on the two planes need to be computed. Taking the frontal plane for example, a point $P_i^{(f)}$ is transformed to $P_i'^{(f)}$ by $P_i'^{(f)} = R(\phi)\,(P_i^{(f)} - O^{(f)})$, where $R(\phi)$ is the rotation matrix in 2D space and ϕ is the angle between old and new x-axes. Figure 6 shows the related shape context and the corresponding histogram.

$$R(\phi) = \begin{bmatrix} \cos\phi & \sin\phi \\ -\sin\phi & \cos\phi \end{bmatrix} \quad (1)$$

When the combined model with four shells and six sectors was used, there were in a total of 24 bins. The radius of the shell is given by $|P_L'^{(f)}O^{(f)}|$, where $L=arg\ \max_k\{|\ P_k'^{(f)}\ O^{(f)}|\ ,\ k=1...33\}$). The shape context vector with respect to the plane centers $O^{(f)}$ and $O^{(s)}$ are gained for each plane and represented by one histogram. Finally, two histogram of shape contexts are represented as the feature vectors of detection input parameters $\{V_1^{(f)}, V_2^{(f)} \ldots V_{24}^{(f)}\}$ and$\{V_1^{(s)}, V_2^{(s)} \ldots V_{24}^{(s)}\}$.

3. Pause Detection

Once the feature vectors are extracted, the pauses are detected. Figure 3 shows the diagram of the algorithm. Although both joint angle feature vectors and shape context feature vectors are present in the diagram, in our experiments, we only used one type of feature vector at a time. However, the following pause detection scheme can be applied to both cases. Consider a data window of m frames (Figure 7).

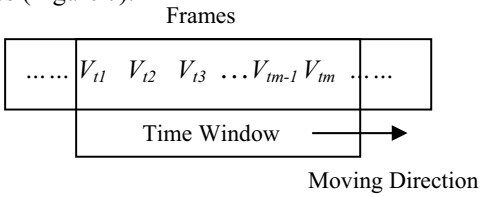

Figure 7. Time Window

Assume that the elements of the feature vectors are statistically independent. At the current time instant t_m, the mean $\mu_i^{(tm)}$ and variance $D_i^{(m,tm)}$ of each component of the feature vectors from t_1 to t_m are computed, where i is the feature vector component index, ranging from 1 to the dimensionality of the feature vector, which would be 22 for joint angle feature vector and 24 for the shape context feature vector. For each vector component, define an indicator function $I_i^{(m,t_m)} \in \{0,1\}$, to represent the moving and stillness of each feature vector at current frame such that $I_i^{(m,t_m)}$ is set to 0 (moving) when $D_i^{(m,tm)}$ is greater than a pre-chosen threshold d, or 1 otherwise (pause).

$$\mu_i^{(t_m)} = \frac{\sum_{k=t_1}^{t_m} V_{i,k}}{t_m - t_1} \quad (2)$$

$$D_i^{(m,t_m)} = \frac{\sum_{j=1}^{m} (V_{i,j} - \mu_i^{(t_m)})^2}{m-1} \quad (3)$$

$$I_i^{(m,t_m)} = \begin{cases} 1, D_i^{(m,t_m)} \le d \\ 0, D_i^{(m,t_m)} > d \end{cases} \quad (4)$$

The probability of pause of the entire data frame $P^{(m,n,t_m)}$ is obtained as the sample mean of the pause indicator functions related to all the feature vector elements as follows:

$$P^{(m,n,t_m)} = \frac{\sum_{i=1}^{n} I_i^{(m,t_m)}}{n} \quad (5)$$

For the shape context feature, two probabilities are computed from two planes. The smaller one is used as the final probability of pause.

4. Experimental Results

Pause detection algorithm is implemented and runs in real time inside the interactive dance system using Visual C++ in the .net environment on a Pentium IV PC with 2.3 GHz CPU and 1GByte memory.

Table 2. Pause Detection Rate %

Exposition	1		2		3	
	P	M	P	M	P	M
Angle	96.4	92.4	97.6	89	100	94.3
Shape Context	92.6	97.6	89.2	98.2	96.4	93.5

*P represents pause detection rate (1-False Negative Rate)
*M represents moving detection rate (1-False Positive Rate)

The data consists of three data sets. Both joint angle and shape context feature vectors were tested. The ground-truth of the data were obtained manually. The detection rates are presented in terms of false negative and false positive. It can be seen from the results that joint angle feature vector performs better with fewer false negatives while shape context feature vector produces fewer false positives.

5. Conclusion

In this paper, we used and compared both the joint angle space and the shape context as feature vectors to detect pauses for interactive dance. Pause detection can improve the performance of pose recognition, and the pause itself can be used independently as cues for movers to interact with the system. The performance of the algorithm is satisfactory through testing against the ground-truth data.

6. References

[1] Camurri, A., Hashimoto, S., Ricchetti, M., Ricci, A.,Suzuki, K., Trocca, R. and Volpe, G., "EyesWeb: Toward Gesture and Affect Recognition in Interactive Dance and Music Systems". Computer Music Journal. 24(1), 57-69, 2000

[2] Moore, C.-L., Yamamoto, K. *Beyond Words: Movement Observation and Analysis*. Gordon and Breach Science Publishers, New York, 1988.

[3] Woo, W., Kim, N., Wong, K., and Tadenuma M. "Sketch on Dynamic Gesture Tracking and Analysis Exploiting Vision-based 3D Interface". Proc. SPIE PW-EI-VCIP'01, vol. 4310, pp. 656-666, 2000

[4]. G. Qian, F. Guo, T. Ingalls, L. Olson, J. James and T. Rikakis, "A Gesture-Driven Multimodal Interactive Dance System," in Proceedings of the International Conference on Multimedia and Expo, Taibei, Taiwan, China, June 27-30, 2004

[5]. Mihael Ankerst, Gabi Kastenmüller, Hans-Peter Kriegel, Thomas Seid, "3D Shape Histograms for Similarity Search and Classification in Spatial Databases," Proc. 6th International Symposium on Spatial Databases (SSD'99), Hong Kong, China, July 1999.

[6]. Marcel K¨ortgen, Gil-Joo Park, Marcin Novotni, Reinhard Klein "3D Shape Matching with 3D Shape Contexts," 7th Central European Seminar on Computer Graphics, Budmerice castle, Slovakia; April 22nd - 24th, 2003

[7]. Serge Belongie, Jitendra Malik and Jan Puzicha "Matching Shapes," in Proceedings of the Eighth IEEE International Conference on Computer Vision, Vancouver, BC, Canada, July, 2001.