Similar Motion Retrieval for Dynamic 3D Mesh Based on Modified Shape Distributions

T. Yamasaki[1] and K. Aizawa[1]

[1]Department of Frontier Informatics, Graduate School of Frontier Sciences, The University of Tokyo

Abstract

A similar motion search and retrieval system for dynamic 3D mesh is presented based on a modified shape distribution algorithm. In this paper, three fundamental functions for efficient retrieval have been developed: feature extraction, motion segmentation, and matching. Stable shape feature representation of 3D models has been realized by the modified shape distribution algorithm. Segmentation has been carried out by analyzing the degree of motion using the extracted feature vectors. Then, similar motion retrieval has been achieved employing the dynamic programming in the feature vector space.

Categories and Subject Descriptors (according to ACM CCS): I.5.4 [Pattern Recognition]: Applications

1. Introduction

Dynamic three-dimensional (3D) modeling of real-world objects using multiple cameras has been an active research area in recent years [KRN97, MWTW04, TOKI04]. Since such sequential 3D models, which we call dynamic 3D mesh, are generated employing many cameras, realistic representation of dynamic 3D objects is obtained. The objects' appearance such as shape and color is captured in dynamic 3D mesh. Similar to 2D video, dynamic 3D mesh consists of consecutive sequences of 3D models (frames). Each frame contains three kinds of data: vertices, connection, and color. As the amount of dynamic 3D mesh data increases, the development of efficient and effective retrieval systems is being desired.

Regarding content-based retrieval of 3D human motion, there has been a number techniques for so-called "motion capture" data [SKK04, CCW*04, MRC05] since joints and other feature points are easily located and tracked. In order to use the motion capture systems, however, users have to wear a special suit with optical or magnetic markers. On the other hand, feature tracking is difficult for dynamic 3D mesh because no markers nor sensors are attached to the users. In addition, each frame of dynamic 3D mesh is generated independently regardless of its neighboring frames due to the non-rigid nature of human body and clothes. This would result in un-regularized number of vertices and topology making the tracking problem more difficult.

The purpose of this work is similar motion retrieval for dynamic 3D mesh. For this purpose, we have developed three key components: feature extraction, motion segmentation, and matching. In particular, proper shape feature extraction from each frame and analysis of its temporal change are extra important tasks as compared to motion capture data retrieval since the correspondence of vertices among frames is not clear as mentioned above. In this regard, we have introduced a modified shape distribution algorithm we have developed in [YA06]. Motion segmentation is carried out to divide the whole dynamic 3D mesh sequence into small but meaningful and manageable clips. The clips are used as minimum units for retrieval. Then, a segmentation technique based on motion has been developed. Because motion speed and direction of feature points are difficult to track, the degree of motion is calculated in the feature vector space. The segmentation is conducted by searching local minima in the degree of motion accompanied with simple verification [YA06]. In retrieving, an example of dynamic 3D mesh clip is given to the system as a query. The similarity to each candidate clip is also computed in the feature vector space employing dynamic programming (DP) matching. In our experiments, four dynamic 3D mesh sequences of dances were utilized. In segmentation, high precision and recall rates of 91% and 90%, respectively, have been achieved. In addition, the system has also demonstrated very encouraging results by retrieving a large portion of the desired and related clips.

2. Shape Feature Extraction

A number of feature extraction techniques have been proposed aiming at static 3D model retrieval [TV04] employing histogram-based, graph-based, and view-based algorithms. Among them, a similarity measure called shape distribution [OFCD02] is one of the most powerful approaches. In the algorithm, a number of points (e.g., 1024) are randomly sampled on the 3D model surface. A feature vector for each frame is obtained by generating a histogram for distances of all possible combinations of the points. However, the original shape distribution cannot be directly applied to our dynamic 3D mesh data. Due to the random sampling of points, the generated histograms fluctuate even for the same 3D model. To reveal slight shape difference among frames, stable histograms are needed.

Therefore, we have developed a modified shape distribution algorithm using a clustering technique. In our model, it can be assumed that vertices of 3D models are mostly uniform on the surface. Therefore, they are clustered into 1024 groups using vector quantization (VQ) based on their spatial distribution. Then, the centers of mass of the clusters are used as representative points for distance histogram generation and the same histogram generation procedure as [OFCD02] follows.

3. Motion Segmentation

In contrast to motion segmentation techniques for 3D motion capture data, the number of those for dynamic 3D mesh is quite limited [XYA05, YA06]. Dynamic 3D mesh has been focused only on its acquisition so far, and it is in its infancy. In [XYA05], histograms based on distance of vertices from a fixed reference point were utilized. Segmentation boundaries were defined when the distance between the histograms of successive frames crossed threshold values. This procedure corresponded to dividing the whole sequence into high/low motion activity spans. However, proper threshold setting was left unsolved. In [XYA05], threshold values were decided only by empirical study.

Our segmentation model is also based on the degree of motion. When motion type or direction changes, motion speed decreases temporarily. In human motion such as dances, in particular, motion is paused at segmentation points for a moment to make the dance look lively. Therefore, in this paper, segmentation is carried out by finding local minima in the degree of motion as in [SNI03], which has been developed for segmenting motion capture data. The difference of our approach is that the degree of motion is expressed as the distance between feature vectors of successive frames. In addition, thresholding as employed in [XYA05, SNI03] is not required in our verification process. The verification is based on relative relationships among local minima and neighboring local maxima. Only when the local maxima values on both sides of the local minimum point are greater than 110% of the local minimum value, it is regarded as a segmentation point. Please refer to [YA06] for detail.

Table 1: *Parameters of our dynamic 3D mesh.*

Sequence	#1	#2-1	#2-2	#2-3
# of frames	173	613	612	616
# of vertices	83k	17k	17k	17k
# of patches	168k	34k	34k	34k
Resolution (mm)	5	10	10	10
Frame rate (frames/s)	10	10	10	10

4. Matching Between Motion Clips

In this paper, example-based queries are employed. A clip from a certain dynamic 3D mesh is given as a query and similar motion is searched from the other clips in the database. DP matching [CLRS01] is utilized to calculate the dissimilarity between the query and candidate clips.

A dynamic 3D mesh sequence in a database (Y) is divided into segments in advance according to Section 3. Assume that the feature vector sequences of the query (Q) and the i-th clip in Y, $Y^{(i)}$, are denoted as follows:

$$Q = \{q_1, q_2, \cdots, q_s, \cdots, q_l\} \quad (1)$$

$$Y^{(i)} = \{y_1^{(i)}, y_2^{(i)}, \cdots, y_t^{(i)}, \cdots, y_m^{(i)}\} \quad (2)$$

where q_s and $y_t^{(i)}$ are the feature vectors of the s-th and t-th frame in Q and $Y^{(i)}$, respectively. Besides, l and m represent the number of frames in Q and $Y^{(i)}$. Let us define $d(s,t)$ as the Euclidean distance between q_s and $y_t^{(i)}$ as in (2):

$$d(s,t) = ||q_s - y_t^{(i)}|| \quad (3)$$

Then, the dissimilarity (D) between the sequences Q and $Y^{(i)}$ is calculated as:

$$D(Q, Y^{(i)}) = c(l,m)/\sqrt{l^2+m^2} \quad (4)$$

where the cost function $c(s,t)$ is defined as in the following equation:

$$c(s,t) = \begin{cases} d(1,1), & for\ l = m = 1 \\ d(s,t) + min\{c(s,t-1), c(s-1,t), c(s-1,t-1)\}, & otherwise \end{cases} \quad (5)$$

Here, symbols of Q and $Y^{(i)}$ are omitted in $d(s,t)$ and $c(l,m)$ for simplicity. Since the cost is a function of the sequence lengths, $c(l,m)$ is normalized by $\sqrt{l^2+m^2}$. The lower the D is, the more similar the sequences are.

5. Experimental Results

In our experiments, four dynamic 3D mesh sequences of Japanese traditional dances called “bon-odori” were uti-

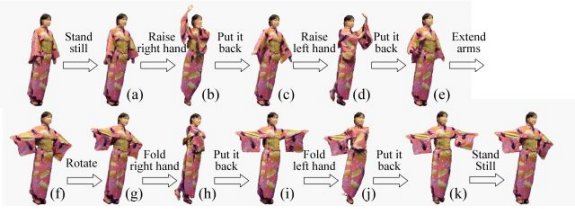

Figure 1: *Subjective segmentation results for sequence #1 by eight volunteers.*

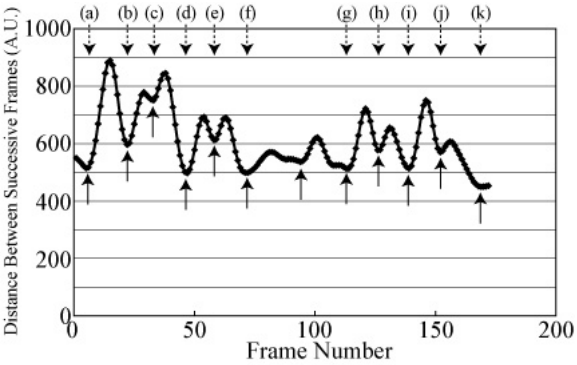

Figure 2: *Comparison of subjectively defined segmentation points and results of our system for sequence #1.*

lized, which were generated using the system developed in [TOKI04]. The parameters of the data are shown in Table 1. Sequences #2-1 $\sim$ #2-3 are identical but performed by different persons. In sequences #2-1 $\sim$ #2-3, about ten types of motion are included and they are combined and repeated several times.

In the experiments, the number of clusters in VQ was set to 1024. The error (distance from the correct feature vector) in the feature vectors coincided with that of the original shape distribution [OFCD02] with 16384 point sampling. Therefore, the stability and correctness of the feature vector is enhanced by 16 by our modified shape distribution. In addition, the computational time is reduced by 256 ($= 16^2$).

An example of segmentation results is shown in Figs. 1 and 2. Fig. 1 shows the subjective motion segmentation results for sequence #1. In the experiment, eight volunteers were involved in the evaluation. The segmentation points were defined when four or more subjects voted for the same points. In Fig. 2, the extracted segmentation points by our system for sequence #1 are illustrated as solid arrows. The dotted arrows from (a) to (k) are the ground truth defined in Fig. 1. There was only one over-segmentation and no miss-segmentation for this sequence.

The performance of segmentation is summarized in Table 2. Both precision and recall rates are larger than 90% in

Table 2: *Parameters of our dynamic 3D mesh.*

Sequence	#1	#2-1	#2-2	#2-3
A: # of relevant records retrieved	11	40	42	34
B: # of irrelevant records retrieved	1	2	3	6
C: # of relevant records not retrieved	0	3	3	8
Precision: A/(A+B)	92	95	93	80
Recall: A/(A+C)	100	93	93	85

Figure 3: *Example of similar motion retrieval: (a) query clip from sequence #2-1; (b) The most similar clip in sequence in sequence #2-2.*

most of the cases. The mean precision and recall rates for all the sequences were 91% and 90%, respectively. Most of the miss-segmentations were caused because the dancer did not pause properly even when the motion type changed. On the other hand, over-segmentation arose when the motion speed was decreased for un-recognizable motion transition such as changing pivoting foot.

Fig. 3 shows an example of similar motion retrieval results. A motion clip of "drawing a big circle by hands" in sequence #2-1 was used as a query and similar motion was searched from clips in sequence #2-2. Fig. 3(b) demonstrates the most similar clip retrieved from #2-2. It has been confirmed that our retrieval system performs quite well for other queries.

Table 3 summarizes the retrieval performance using sequences #2-1 $\sim$ #2-3. In the experiment, each clip from sequences shown in the column was used as a query. And the clips from the sequences shown in the row were used as candidates. The query itself was not included in candidates. The number of query clips was 127 in total. The performance was evaluated by the method employed in [OFCD02]. The "first tier" in Table 3(a) demonstrates the averaged percentage of the correctly retrieved clips in the top-k highest similarity score clips, where k is the number of the ground truth of similar motion clips judged by the authors. An ideal matching would give no false positives and return a score of 100%. The "second tier" in Table 3(b) gives the same type of result, but for the top $2 \times k$ highest similarity score clips. The "nearest neighbor" in Table 3(c) shows the percentage of the test

Table 3: *Retrieval performance: (a) first tier; (b) second tier; (c) nearest neighbor. The row and column represent the candidate clips and the query clip, respectively. The query itself was not included in the candidate clips.*

(a)

Seq.	#2-1	#2-2	#2-3
#2-1	80% (298/372)	63% (252/397)	70% (292/420)
#2-2	62% (247/394)	63% (220/350)	63% (259/414)
#2-3	57% (242/421)	56% (232/414)	85% (346/408)

(b)

Seq.	#2-1	#2-2	#2-3
#2-1	98% (366/372)	84% (335/397)	70% (292/420)
#2-2	85% (335/394)	82% (287/350)	87% (360/414)
#2-3	89% (374/421)	94% (390/414)	96% (392/408)

(c)

Seq.	#2-1	#2-2	#2-3
#2-1	98% (40/41)	76% (31/41)	90% (36/40)
#2-2	57% (24/42)	62% (26/42)	62% (26/42)
#2-3	67% (28/42)	69% (29/42)	90% (36/40)

in which the retrieved clip with the highest score was correct. It is demonstrated that 56% $\sim$ 85% of similar motion clips are included in the first tier and more than 80% (82% $\sim$ 98%) of clips are correctly retrieved in the second tier. Besides, precision rates of nearest neighbor is 57% $\sim$ 98%.

Some false positives were detected due to the fact that shape distribution is designed for global shape features. For instance, the difference in shape between "standing still" and "clapping hands in front of stomach" motions is positions of arms and hands, which takes only a small portion of surface area of 3D models. In such a case, little difference in extracted feature vectors can be observed. Slight difference in shape should be analyzed by other methods.

6. Conclusions

In this paper, key technologies for dynamic 3D mesh retrieval have been developed. The modified shape distribution algorithm has been employed for stable feature representation of 3D models. Segmentation has been conducted analyzing the degree of motion calculated in the feature vector space. The similar motion retrieval has been realized by DP matching scheme using the feature vectors. We have demonstrated effective segmentation with precision and recall rates of over 90% on average. In addition, reasonable retrieval results have been demonstrated by experiments.

7. Acknowledgements

This work is supported by Ministry of Education, Culture, Sports, Science and Technology of Japan under the "Development of fundamental software technologies for digital archives" project.

References

[CCW*04] CHIU C., CHAO S., WU M., YANG S., LIN H.: Content-based retrieval for human motion data. *Journal of Visual Communication and Image Representation 15*, 3 (2004), 446–466.

[CLRS01] CORMEN T., LEISERSON C., RIVEST R., STEIN C.: *Introduction to Algorithms*. MIT Press & McGraw-Hill, 2001.

[KRN97] KANADE T., RANDER P., NARAYANAN P.: Virtualized reality: constructing virtual worlds from real scenes. *IEEE Multimedia 4*, 1 (Jan./Mar. 1997), 34–47.

[MRC05] MULLER M., RODER T., CLAUSEN M.: Efficient content-based retrieval of motion capture data. In *Proc. SIGGRAPH2005* (2005), pp. 677–685.

[MWTW04] MATSUYAMA T., WU X., TAKAI T., WADA T.: Real-time dynamic 3-d object shape reconstruction and high-fidelity texture mapping for 3-d video. *IEEE Trans. Circuit and System for Video Technology 14*, 3 (Mar 2004), 357–369.

[OFCD02] OSADA R., FUNKHOUSER T., CHAZELLE B., DOBKIN D.: Shape distributions. *ACM TOG 21*, 4 (2002), 807–832.

[SKK04] SAKAMOTO Y., KURIYAMA S., KANEKO T.: Motion map: image-based retrieval and segmentation of motion data. In *Proc. 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation* (2004), pp. 259–266.

[SNI03] SHIRATORI T., NAKAZAWA A., IKEUCHI K.: Rhythmic motion analysis using motion capture and musical information. In *Proc. IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems* (2003), pp. 89–92.

[TOKI04] TOMIYAMA K., ORIHARA Y., KATAYAMA M., IWADATE Y.: Algorithm for dynamic 3d object generation from multi-viewpoint images. In *Proc. SPIE* (2004), vol. 5599, pp. 153–161.

[TV04] TANGELDER J., VELTKAMP R.: A survey of content based 3d shape retrieval methods. In *Proc. Shape Modeling International 2004* (2004), pp. 145–156.

[XYA05] XU J., YAMASAKI T., AIZAWA K.: Effective 3d video segmentation based on feature vectors using spherical coordinate system. In *Meeting on Image Recognition and Understanding (MIRU) 2005* (2005), pp. 136–143.

[YA06] YAMASAKI T., AIZAWA K.: Motion segmentation of 3d video using modified shape distribution. In *IEEE 2006 International Conference on Multimedia & Expo (accepted)* (2006).