MOTION SEGMENTATION OF 3D VIDEO USING MODIFIED SHAPE DISTRIBUTION

Toshihiko Yamasaki and Kiyoharu Aizawa

Department of Frontier Informatics, Graduate School of Frontier Sciences
The University of Tokyo

ABSTRACT

In this paper, temporal segmentation of 3D video based on motion analysis is presented. 3D video is a sequence of 3D models made for a real-world dynamic object. A modified shape distribution algorithm is proposed to realize stable shape feature representation. In our approach, representative points are generated by clustering vertices based on their spatial distribution instead of randomly sampling vertices as in the original shape distribution algorithm. Motion segmentation is conducted analyzing local minima in degree of motion calculated in the feature vector space. The segmentation algorithm developed in this paper does not require any predefined threshold values but rely on relative relationships among local minima and local maxima of the motion. Therefore, robust segmentation has been achieved. The experiments using 3D video of traditional dances yielded encouraging results with the precision and recall rates of 93% and 88%, respectively, on average.

1. INTRODUCTION

In recent years, great interests have been paid to dynamic three-dimensional (3D) modeling using synchronous multiple cameras [1]-[4]. Such sequential 3D models (we hereafter call them 3D video) can provide faithful and accurate 3D information of the real-world objects from the points of view of shape, color, and motion. In this regard, they are different from conventional 3D computer graphics and 3D motion capture data.

Our goal is to construct a large-scale database of 3D video and to develop efficient tools for their exploitation. One of the most essential functions in managing the database is motion segmentation of video sequence. It is the first step towards automatic annotation, indexing, browsing, retrieval, and so forth. In particular, segmenting 3D video into as small but meaningful pieces as possible is desired. Important to note here is that this motion segmentation is different from conventional segmentation for 2D video [5][6], which depends on scene changes.

Related works for motion segmentation can be found in 2D video [7][8] and in 3D motion capture data [9]-[13]. In 2D video, object segmentation to extract moving objects was usually conducted in the first step. Then, in [7], optical flow of moving objects was analyzed by singular value decomposition (SVD) and motion discontinuities in trajectories of the basis coefficients over time were detected. In [8], local minima in motion and local maxima in direction change were searched.

There are also a number of segmentation techniques for 3D motion capture data [9]-[13], since structural features such as motion of joints and other feature points are easily located and tracked. In [9], local minima in motion were analyzed. The idea of searching local minima in kinematic parameters was also employed in [10]. Some other approaches were proposed based on estimation error using SVD [11] and least square fitting [12]. In addition, model-based approaches were reported using Hidden Markov Model (HMM) [13] and Gaussian Mixture Model [11].

In contrast to motion segmentation of 3D motion capture data, that of 3D video is much more challenging because structural features are not available. 3D video is basically generated frame by frame independently. Therefore, it is hard to reveal the correspondence of joints and feature points of the object among frames. Therefore, the number of 3D video segmentation reported so far is quite limited [14][15]. In [14], a histogram of distance among vertices on 3D mesh model and three fixed reference points was generated for each frame, and segmentation was done when the distance between histograms of successive frames crossed threshold values. And, more efficient histograms based on spherical coordinate system were developed in [15]. The problem of these two approaches is that they strongly relied on “suitable” thresholding, which was defined only by empirical study (try and error) for each sequence. In addition, sensitivity to rotation and translation of the 3D models still remained in the histograms.

Therefore, in this paper, we propose a motion segmentation technique based on a 3D shape feature extraction algorithm called shape distribution [16]. We have modified the original shape distribution algorithm to obtain more stable histograms. In addition, we have developed a simple but effective segmentation criterion based on the motion. It does not require a pre-defined threshold value. In the experiments using 3D video of traditional dances, high precision and recall rates of 93% and 88%, respectively, have been achieved.

1-4244-0367-7/06/$20.00 ©2006 IEEE

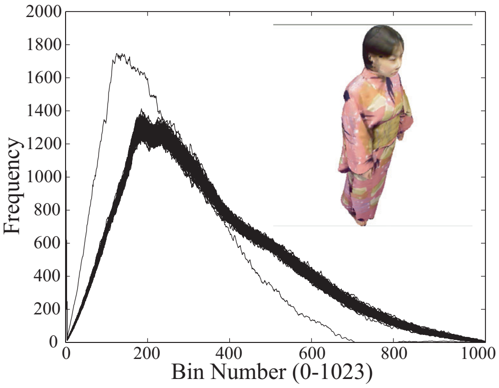

Fig. 1. Thirty histograms for the same 3D model (shown on the upper side) using the original shape distribution [16].

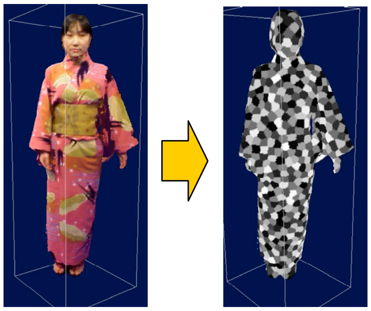

Fig. 2. Concept of modified shape distribution. Vertices of 3D model are firstly clustered into 1024 groups by vector quantization in order to scatter representative vertices uniformly on 3D model surface and to generate histogram stably.

2. MODIFIED SHAPE DISTRIBUTION

For 3D video segmentation, two steps of processing are needed: feature extraction and motion analysis. With regard to feature extraction from 3D models, a number of techniques have been developed aiming at static 3D model retrieval [17]. In this study, features of each frame are extracted and their temporal change is analyzed.

Among the feature extraction algorithms, shape distribution [16] is known as one of the most effective methods. In [16], a number of points (e.g., 1024) were randomly scattered on the 3D model surface and distance between all possible combination of points was calculated. Then, the histogram of distance distribution was generated as a feature vector to express the shape characteristics of a 3D model. The shape distribution algorithm has a virtue of robustness to objects' rotation, translation, and so on.

However, histograms using the original shape distribution cannot be generated stably because of the random sampling of the 3D surface. Fig. 1 shows 30 histograms generated for the same 3D model selected from our 3D video. The histograms were generated by scattering 1024 vertices and setting the bin number as 1024 (divided the range between maximum and minimum values in distance into 1024). It is observed that the shapes of the histograms fluctuate and sometimes a totally different histogram is obtained. In [16], deviation in the histograms was not so significant because rough feature extraction was pursued for similar shape retrieval of static 3D models. On the other hand, in our case, it is required to clarify a slight shape difference among frames in 3D video.

Therefore, we have modified the original shape distribution algorithm for more stability. Since vertices are mostly uniform on the surface in our 3D models, they are firstly clustered into 1024 groups based on their 3D spatial distribution employing vector quantization as shown in Fig. 2. The code vectors are regarded as representative vertices for calculating distance. Although such clustering process is computationally expensive, it needs to be carried out only once in advance. Therefore, the computational cost can be neglected. As a result of the clustering, representative points are distributed uniformly and generation of more stable histograms has been made possible. In our algorithm, the bin number is set to 1024. After obtaining histograms, smoothing is applied to them to remove noise.

The changes in the shape of generated histograms correspond to those in posture or shape of 3D models. Therefore, in our algorithm, the distance between histograms are utilized to express the degree of motion.

3. MOTION SEGMENTATION

In motion segmentation, for dance sequences in particular, motion speed is an important factor. When human change motion type or motion direction, the motion speed becomes small. More importantly, motion is shortly paused at segmentation points to make the dance look elegant.

Searching the points when the motion speed becomes small is achieved by looking for local minima in distance between the histograms of the successive frames. In this regard, our approach is similar to [8][9]. The difference is that since movement of feature points of human body in 3D video is not clear like motion capture data, the degree of motion is calculated in the feature vector space.

In [9], the extracted local minima in motion speed were verified by thresholding whether they were truly segmentation points or not. The local minimum values should be lower than a predefined threshold value and the local maximum values between the local minima should be higher than another threshold. In this respect, threshold optimization depending on input data was still required in [9].

In our scheme, local minima are regarded as segmentation points when the two local maxima on both sides of the local minimum value (D_{lmin}) are greater than 1.1 x D_{lmin}. Since the verification is relative, it is robust to data variation and no empirical decision is required. In addition, one dimensional data of motion speed goes thorough smoothing process with a Gaussian-like filter.

Table 1. Summary of 3D video utilized in experiments.

Sequence	# 1	# 2-1	# 2-2	# 3
# of frames	173	613	612	1,981
# of vertices (average)	83k	17k	17k	17k
# of patches (average)	168k	34k	34k	34k
Resolution	5 mm	10 mm	10 mm	10 mm
Frame rate	10 frames/s			

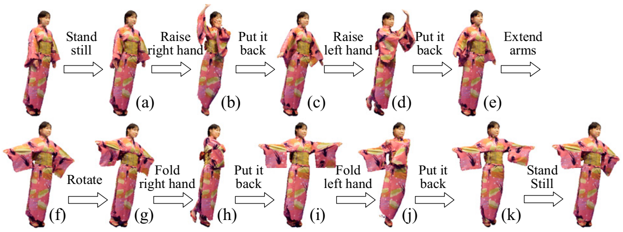

Fig. 3. Subjective segmentation results for sequence #1 by eight volunteers.

4. EXPERIMENTAL RESULTS

In our experiments, four 3D video sequences generated by the system developed in [4] were utilized. The parameters of the data are summarized in Table 1. The sequences #1 and #2 are Japanese traditional dances called *bon-odori* and the dance #3 is a Japanese warm-up dance. The sequences #2-1 and #2-2 are identical but performed by different persons.

Fig. 3 demonstrates the subjective segmentation results by eight volunteers. They were asked to define motion boundaries without any instruction nor others' segmentation results. In this experiment, when results of four (50%) or more subjects voted for the same points, the segmentation boundaries were defined. The results were used for evaluation. For the sequences #2 and #3, the segmentation boundaries were defined by the authors.

The segmentation results for the sequence #1 are illustrated in Fig. 4. The ordinate represents the distance between histograms of successive frames. The dotted arrows from (a) to (k) represent the subjectively defined segmentation points shown in Fig. 3. The solid arrows are the results of our system. There was only one over-segmentation. In addition, no miss-segmentation was detected. The over-segmentation between (f) and (g) was due to the fact that the pivoting foot was changed while the dancer was rotating and motion speed decreased temporarily.

As another example, the first 15 segmentation points (approximately, out of 21 seconds) obtained from sequence #2-1 are shown in Fig. 5. It is observed that the 3D video sequence is divided into small but meaningful segments. There was only one over-segmentation and no miss-segmentation for the period, which is not shown in the figure.

In our algorithm, only the distance between two successive frames are considered. Fig. 6 shows the precision and recall rates when more neighboring frames are involved in the distance calculation using sequence #2-1. As the number of frames involved in the calculation increases, recall rate is slightly improved while precision rate declines. This is because involving more neighboring frames in calculating the degree of motion corresponds to neglecting small or quick motion. Our 3D video was captured at 10 frames/s due to the system constraint. In such a low frame rate case, calculating the distance between only the successive frames yields the best performance.

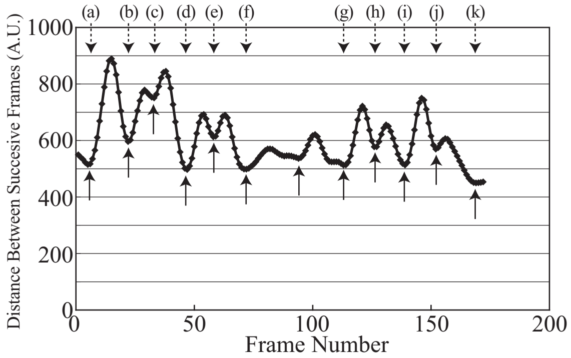

Fig. 4. Comparison of subjectively defined segmentation points and results of our system for sequence #1.

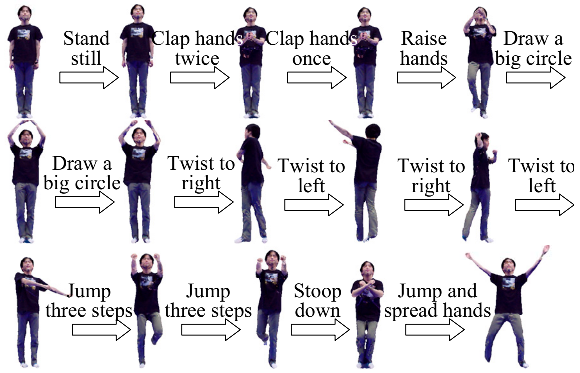

Fig. 5. First 15 segmentation points in sequence #2-1.

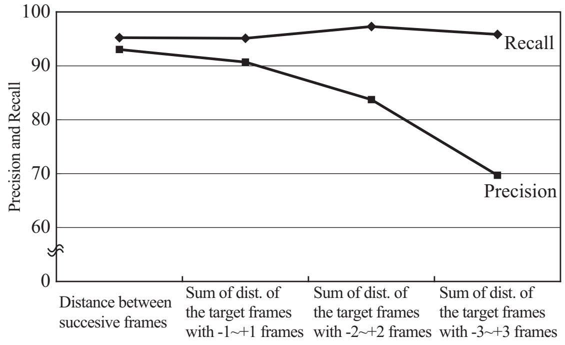

Fig. 6. Precision and recall rates when the number of neighboring frames involved in calculation of degree of motion was changed. Sequence #2-1 was used.

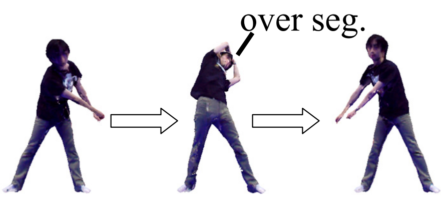

Fig. 7. Example of over-segmentation.

Table 2. Performance summary of segmentation.

Sequence	# 1	# 2-1	# 2-2	# 3
A: # of relevant records retrieved	11	40	42	124
B: # of irrelevant records retrieved	1	2	3	11
C: # of relevant records not retrieved	0	3	3	25
Precision: A/(A+B)	92	95	93	92
Recall: A/(A+C)	100	93	93	83

Table 2 summarizes the segmentation performance. There are only a few miss- and over-segmentations per minute. Since the sequence #3 contains complicated motion, which is hard to detect, the number of miss-segmentations is larger than the other sequences.

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 occurred when the motion speed decreased even when the meaning of the motion did not change. An example is shown in Fig. 7. The dancer is drawing a big circle by arms, and the motion speed decreases at the top of the circle. To resolve the problem, high-level motion observation may be needed.

5. CONCLUSIONS

In this paper, motion segmentation for 3D video was discussed. A robust and stable shape feature extraction has been realized by the modified shape distribution algorithm to scatter representative points uniformly on the 3D model surface. Then, the degree of motion was expresses as the distance between histograms of successive frames. Segmentation boundaries were defined by searching the local minima in motion speed with a simple verification process which does not require pre-defined threshold values. As a result, such high precision and recall rates as 93% and 88%, respectively, on average have been achieved. In our future work, we are planning to apply our feature representation and segmentation results to similar motion retrieval of 3D video.

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. All the 3D video data were take by the courtesy of Japan Broadcasting Corporation (NHK).

REFERENCES

[1] T. Kanade, P. Rander, and P. Narayanan, “Virtualized reality: constructing virtual worlds from real scenes,” IEEE Multimedia, vol. 4, no. 1, pp. 34–47, Jan./March 1997.

[2] S. Wurmlin, E. Lamboray, O.G. Staadt, and M. H. Gross, “3D video recorder,” Proc. Pacific Graphics’02, pp. 325-334, 2002.

[3] T. Matsuyama, X. Wu, T. Takai, and T. Wada, “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, vol. 14, no. 3, pp. 357–369, March 2004.

[4] K. Tomiyama, Y. Orihara, M. Katayama, and Y. Iwadate, “Algorithm for dynamic 3D object generation from multi–viewpoint images,” Proc. SPIE, Vol. 5599, pp. 153–161, 2004.

[5] F. Idris and S. Panchanathan, “Review of image and video indexing techniques,” Journal of Visual Communication and Image Representation, vol. 8, no. 2, pp. 146-166, June 1997.

[6] I. Koprinska and S. Carrato, ”Temporal video segmentation: A survey,” Signal Processing: Image Communication 16, pp. 477-500, 2001.

[7] Y. Rui and P. Anandan, “Segmenting visual actions based on spatio-temporal motion patterns,” Proc. IEEE Computer Vision and Pattern Recognition, pp. 1111-1118, 2000.

[8] T.S. Wang, H.Y. Shum, Y.Q. Xu, and N.N. Zheng, “Unsupervised analysis of human gestures,” Proc. IEEE Pacific Rim Conference on Multimedia, pp. 174-181, 2001.

[9] T. Shiratori, A. Nakazawa, and K. Ikeuchi, “Rhythmic motion analysis using motion capture and musical information,” Proc. IEEE Conf. on Multisensor Fusion and Integration for Intelligent Systems, pp. 89-92, 2003.

[10] K. Kahol, P. Tripathi, and S. Panchanathan, “Automated gesture segmentation from dance sequences,” Proc. 6th IEEE Int. Conf. on Automatic Face and Gesture Recog., pp. 883-888, 2004.

[11] J. Barbie, A. Safonova, J.Y. Pan, and C. Faloutsos, “Segmenting motion capture data into distinct behaviors,” Proc. of Graphics Interface 2004 (GI'04), pp. 195-194, 2004.

[12] C.M. Lu and N.J. Ferrier, “Repetitive motion analysis: segmentation and event classification,” IEEE TPAMI, vol. 26, no. 2, pp. 258-263, 2004.

[13] W. Takano and Y. Nakamura, “Segmentation of human behavior patterns based on the probabilistic correlation,” Proc. 19th Annual Conf. of the Japanese Society for Artificial Intelligence, 3F1-01, 2005.

[14] J. Xu, T. Yamasaki, and K. Aizawa, “3D video segmentation using point distance histograms,” Proc. 2005 IEEE Int. Conf. on Image Processing (ICIP2005), pp. I–701–I–704, 2005.

[15] J. Xu, T. Yamasaki, and K. Aizawa, “Effective 3D video segmentation based on feature vectors using spherical coordinate system,” Proc. Meeting on Image Recognition and Understanding (MIRU) 2005, pp. 136–143, 2005.

[16] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, “Shape distributions,” ACM Transactions on Graphics (TOG), vol. 21, issue 4, pp. 807-832, 2002.

[17] J. Tangelder and R.C. Veltkamp, “A survey of content based 3D shape retrieval methods,” Proc. Shape Modeling International 2004, pp. 145-156, 2004.