CONTENT-BASED CROSS SEARCH FOR HUMAN MOTION DATA USING TIME-VARYING MESH AND MOTION CAPTURE DATA

Toshihiko Yamasaki and Kiyoharu Aizawa

Department of Information and Communication Engineering, The University of Tokyo

ABSTRACT

This paper describes a content-based cross search scheme for two kinds of three-dimensional (3D) human motion data: time-varying mesh (TVM) and motion capture data. TVM is a sequence of 3D mesh models made for real-world 3D objects. TVM is generated using multiple-view images taken with multiple cameras. Since TVM can record shape, color, and motion of the real-world 3D objects, it has been drawing a lot of attention these days. In order to realize practical archiving systems for TVM, efficient retrieval systems are indispensable. The retrieval systems for TVM developed so far are based on query-by-example. This means additional TVM generation is required for constructing queries, which is computationally demanding and time consuming. On the other hand, motion capture systems are widely used to capture 3D human motion. However, the data structure is very different from that of TVM. Therefore, the two kinds of 3D human motion data are incompatible to each other. In this paper, we present a retrieval system that enables retrieving TVM using motion capture data as queries and vice versa using the modified shape distribution algorithm.

1. INTRODUCTION

3D motion capture and recording of real-world objects has been one of the ultimate goals of computer vision. One of the successful and matured systems is a "motion capture system [1]-[3]." Although there are still some problems in motion capture systems such as how many sensors are to be used and where they should be attached, motion capture systems are widely used in commercial applications such as movies and video games. Also, there are some free motion capture data libraries on the Internet [4]-[6].

In motion capture systems, performers wear special suits with optical or magnetic markers on their feature points such as joints. Thanks to these markers, it is possible to obtain structural information (i.e., the location and motion of the feature points) correctly, However, the markers or special suits often limits the degree of freedom of the performers' motion. In addition, motion capture systems can obtain only the skeletal motion: dynamics of the clothes and creases on it are out of the scope of the systems.

In this point of view, some researches to capture the whole 3D scenes as time-varying mesh (TVM) has been conducted in the last decade [7]-[11]. TVM is a sequence of 3D mesh models that are generated from multiple-view images taken with multiple cameras using shape-from-silhouette approaches. TVM can capture and reproduce not only the shape and the color but also the motion of the 3D objects. Although the TVM generation is still an emerging technology, efficient and effective retrieval systems for TVM will be required in the future for managing a large-scale database of TVM. To our best knowledge, the retrieval systems for TVM reported so far are only by the authors [12][13]. On the other hand, we can find a number of papers on retrieval systems for motion capture data [14]-[16]. The difficulty in achieving TVM retrieval systems lies not only on the lack of TVM data available to the researchers but also on the fact that it is difficult to locate and track the feature points in TVM. Due to the non-rigid nature of human body and clothes, TVM frames are generated independently regardless of their neighboring frames. Therefore, the number of vertices and connection change every frame, which makes it difficult to extract structural information. Although there is a paper to regularize the 3D mesh using a deformation algorithm [9], it is necessary to refresh the 3D model every few frames.

In our previous TVM retrieval systems [12][13], a query-by-example approach was employed. Although the system demonstrated promising results, query sequences were selected from the database and matched with the other sequences. It is computationally expensive and time consuming to generate new TVM sequences as queries. Therefore, how to generate queries with lower cost is still an open problem.

The purpose of this paper is to develop a retrieval system that combines TVM with motion capture data for low cost and efficient content-based 3D motion search. The system enables us to retrieve TVM sequences using motion capture data as queries. In addition, it is also possible to retrieve motion capture data using TVM. Since the data structure is different from each other, we first convert motion capture data to 3D mesh sequences, which can be easily conducted using 3D computer graphics (CG) rendering software. Then, feature vectors are extracted from the rendered mesh models using our modified shape distribution algorithm [12][13]. In

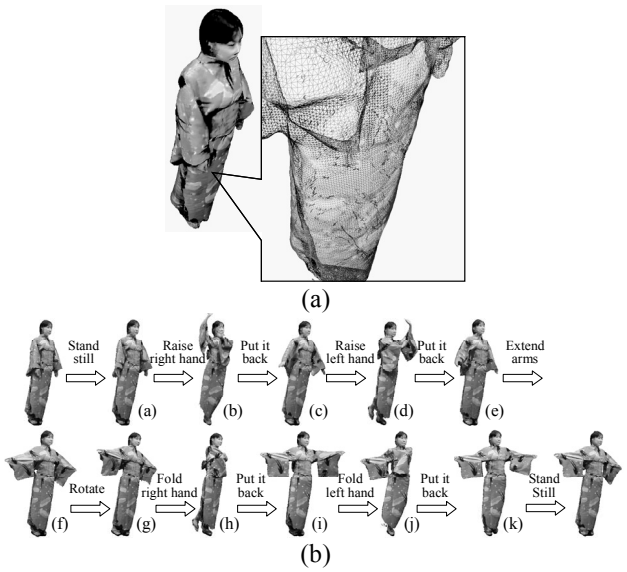

Fig. 1. Example of our TVM data: (a) close-up, (b) sequence. Our TVM consists of three kinds of information: coordinates of vertices, their connection, and color.

this manner, the feature vectors in the same feature domain are extracted from both TVM and motion capture data. The similarity evaluation is accomplished by dynamic programming (DP) matching among the feature vector sequences. Besides, 3DCG animation characters can also be used in the system. Experimental results demonstrated very promising results by retrieving most of the related motion sequences.

2. TIME-VARYING MESH (TVM)

The TVM data in this work were obtained employing the system in [10]. They were generated from multiple-view images taken with 22 synchronous cameras in a dedicated studio of 8 m in diameter and 2.5 m in height.

Different from 2D video, TVM is composed of a consecutive sequence of 3D models. An example of our TVM data seen from a certain view point is shown in Fig. 1. Needless to say, the view point can be changed arbitrary according to the users' taste. Each frame of TVM is represented as a polygon mesh model. Namely, each frame is expressed by three kinds of data: coordinates of vertices, their connection, and color (not shown in the figure).

3. RETRIEVAL ALGORITHM

3.1. Shape Feature Extraction

The essential point in our system is that feature vectors compatible to those from TVM are extracted from the motion capture data. For this purpose, the motion capture data are firstly converted to 3D mesh models using off-the-shelf

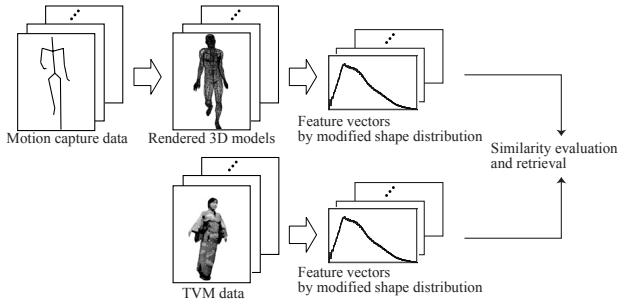

Fig. 2. Flowchart of the proposed retrieval system by the combination of TVM and motion capture data.

3DCG rendering software. Then, feature vectors are generated from both rendered mesh models and TVM using the modified shape distribution algorithm [12][13].

The modified shape distribution is an algorithm to calculate the global shape feature of 3D models. The algorithm derives from the shape distribution developed in [17] and has been modified to extract feature vectors more stably and accurately. In the original shape distribution [17], a number of points are randomly scattered on the 3D model surface and distances among all possible combination of the points are calculated to form a distance histogram, which is used as a feature vector. The shape distribution algorithm is robust to objects' translation, rotation, scale change, mirror, tessellation, simplification, and so on. Therefore, model size inconsistency between TVM and motion capture data can be neglected in the matching. In our modified shape distribution algorithm [12][13], the vertices of the 3D mesh models are clustered in advance to scatter the representative points uniformly, thus achieving stable and accurate feature vector generation.

3.2. Similarity Evaluation

Once the motion capture data and TVM are represented as feature vectors, the similarity evaluation among sequences is simple. In this paper, the similarity is calculated using a DP matching as in [12][13].

Assume that we have a database (Y) of TVM and motion capture data. Then, let us denote the feature vector sequences of the query (Q) and the i-th clip in Y, $Y^{(i)}$, as follows:

$$Q = \{\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_s, \cdots, \mathbf{q}_l\}$$
$$Y^{(i)} = \{\mathbf{y}_1^{(i)}, \mathbf{y}_2^{(i)}, \cdots, \mathbf{y}_t^{(i)}, \cdots, \mathbf{y}_m^{(i)}\} \quad (1)$$

where $\boldsymbol{q}_s$ and $\boldsymbol{y}_t^{(i)}$ 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 $Y^{(i)}$. Important to note here is that it does not matter whether Q and $Y^{(i)}$ are from TVM or motion capture data.

Let us define $d(s,t)$ as the Euclidean distance between $\boldsymbol{q}_s$ and $\boldsymbol{y}_t^{(i)}$ as in (2):

Table 1. List of motion sequences.

ID	Motion	Source
a	stretch	TVM
b	stretch	TVM
c	stretch	motion capture
d	pitching	TVM
e	pitching	motion capture
f	pitching	motion capture
g	pitching	motion capture
h	walk	TVM
i	walk	TVM
j	walk	TVM
k	walk	TVM
l	walk	TVM
m	walk	motion capture
n	walk	motion capture
o	skip	motion capture
p	skip	motion capture
q	skip	motion capture
r	skip	motion capture
s	skip	motion capture

$$d(s,t) = \left\| \mathbf{q}_s - \mathbf{y}_t^{(i)} \right\| \quad (2)$$

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

$$D\left(Q, Y^{(i)}\right) = \mathrm{c}(l,m) \Big/ \sqrt{l^2 + m^2} \quad (3)$$

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

$$\mathrm{c}(s,t) = \begin{cases} d(1,1), & \text{for } l = m = 1 \\ d(s,t) + \min\left\{\mathrm{c}(s,t-1), \mathrm{c}(s-1,t), \mathrm{c}(s-1,t-1)\right\}, & \text{otherwise} \end{cases} \quad (4)$$

In this manner, similar motion search is achieved. The flowchart of the overall process is shown in Fig. 2.

3.3. Discussion

The counterpart of our approach would be extracting skeleton models from TVM that are compatible to motion capture data and apply retrieval algorithms developed for motion capture data. Once the structural information is obtained from TVM, it would be possible to use well-established and sophisticated retrieval algorithms for motion capture data [14]-[16].

Examples of extracting structural information from TVM or 3D volume data can be found in [19]-[21]. And, some preliminary results of the retrieval based on the idea described above is demonstrated in [21]. However, it is generally difficult to extract structural information from TVM stably. Since TVM is constructed only from multiple images, the generated 3D models are noisy by nature. When noise such as holes, truncation, inconsistent connection, and so on exist in TVM data, the topology of the extracted skeleton models change drastically. In addition, when some of the body parts touch to each other, the topology of the original 3D model itself changes, making it harder to extract structural information consistent to neighboring frames. Therefore, extracting feature vectors from motion capture data that are compatible to those of TVM is more practical approach.

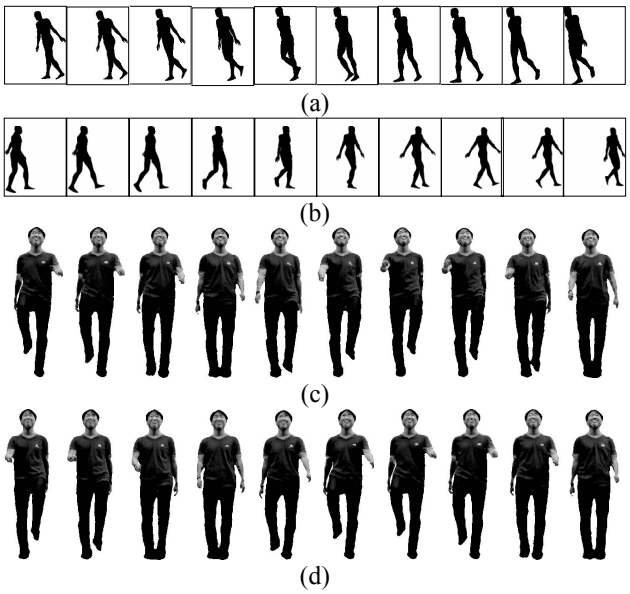

Fig. 3. Results of content-based retrieval: (a) query (sequence *m*), (b) the most similar sequence (sequence *n*), (c) the second (sequence *k*), (d) the third (sequence *j*). Only the first 10 frames are shown. The scores for sequences *n*, *k*, and *j* were 805, 1245, and 1280, respectively.

4. EXPERIMENTAL RESULTS

In this paper, we used motion capture data available on [5]. We used stretching, pitching, walking, and skipping sequences. Table 1 shows a list of the motion data used in the experiments. The motion capture data in the Biovision Hierarchical (bvh) format were converted to naked 3DCG mesh models using a rendering software called Poser [18]. The number of sampled vertices and the number of bins (i.e., dimension of the feature vectors) were both set at 1,024.

Fig. 3 demonstrates one of the retrieval results using a walking sequence in motion capture data as a query. In the figure, only the first 10 frames are shown due to the space limitation. We can see that walking sequences in both motion capture data and TVM are correctly retrieved.

Fig. 4 shows the similarity scores among the sequences. The darker the color is, the more similar the sequences are. The maximum value (the largest dissimilarity D mapped to the grayscale value of 255) is about 3,000. When the se-

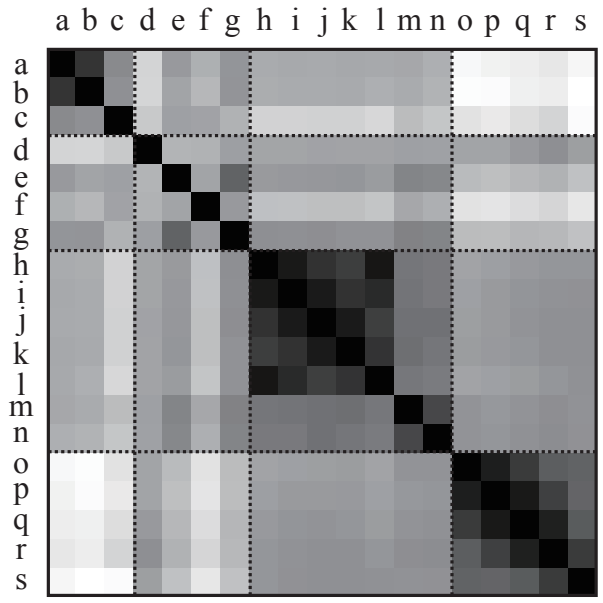

Fig. 4. Similarity scores among sequences. The indices from *a* to *s* corresponds to those in Table 1. It shows that he darker the color is, the similar the sequences are.

quences are similar, the distance is generally below 2,000. It is demonstrated that the similar motions yields higher similarity score properly regardless of its data type. In addition, the second most similar motion group to "walk" sequences is "skip" motion, which coincide with our perception of similarity. Therefore, we can conclude that the similarity evaluation in the feature vector space of the modified shape distribution is a reasonable approach for the retrieval using TVM and motion capture data. For the retrieval performance itself using modified shape distribution, please refer to [13].

5. CONCLUSIONS

In this paper, we have developed an efficient content-based retrieval system for TVM and motion capture data. In conventional retrieval systems for TVM, low cost query generation has been a problem. In the present system, motion capture data were rendered as 3DCG mesh models and the shape feature vectors were extracted from both TVM and the 3DCG mesh models using the modified shape distribution algorithm. The similarity evaluation criteria has been employed from our previous work. As a result, efficient TVM retrieval using motion capture data as queries and vice versa has been made possible. Experimental results demonstrated the validity of our approach by retrieving proper sequences.

6. REFERENCES

[1] T. W. Calvert, J. Chapman and A. Patla, "Aspects of the kinematic simulation of human movement," IEEE Computer Graphics and Applications, vol. 2, no. 9, pp. 41-50, 1982.

[2] Carol M. Ginsberg and Delle Maxwell, "Graphical marionette," Proc. ACM SIGGRAPH/SIGART Workshop on Motion, pp. 172-179, 1983.

[3] A. Menache, Understanding Motion Capture for Computer Animation and Video Games, Morgan Kaufmann, 1999.

[4] http://mocap.cs.cmu.edu/

[5] http://www.mocapdata.com/

[6] http://www.bvhfiles.com/

[7] 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.

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

[9] 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.

[10] 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.

[11] J. Starck, and A. Hilton, "Virtual view synthesis of people from multiple view video sequences," Graphical Models, vol. 67, no. 6, pp. 600-620, 2005.

[12] T. Yamasaki and K. Aizawa, "Similar motion retrieval of dynamic 3D mesh based on modified shape distribution," Proc. Eurographics2006, pp. 9-12, 2006.

[13] T. Yamasaki and K. Aizawa, "Motion segmentation and retrieval for 3D video based on modified shape distribution," EURASIP Journal on Applied Signal Processing, vol. 2007, Article ID 59535, 11 pages, 2007.

[14] Y. Sakamoto, S. Kuriyama, and T. Kaneko, "Motion map: image-based retrieval and segmentation of motion data," Proc. 2004 ACM SIGGRAPH/Eurographics symposium on Computer animation, pp. 259-266, 2004.

[15] C.Y. Chiu, S.P. Chao, M.Y. Wu, S.N. Yang, and H.C. Lin, "Content-based retrieval for human motion data," Journal of Visual Communication and Image Representation, vol. 15, no. 3, pp. 446-466, 2004.

[16] M. Muller, T. Roder, and M. Clausen, "Efficient content-based retrieval of motion capture data" Proc. SIGGRAPH2005, pp. 677-685, 2005.

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

[18] http://www.e-frontier.com/go/poser

[19] T. Mukasa, S. Nobuhara, A. Maki, and T. Matsuyama, "Finding articulated body in time-series volume data," The 4th International Conference on Articulated Motion and Deformable Objects, LNCS 4069, pp. 395-404, 2006.

[20] M. Iiyama, Y. Kameda, and M. Minoh, "Estimation of the location of joint points of human body from successive volume data," Proc. ICPR2000, pp. 699-702, 2000.

[21] R. Tadano, T. Yamasaki, and K. Aizawa, "Fast and robust motion tracking for time-varying mesh featuring Reeb-graph-based skeleton fitting and its application to motion retrieval," in Proc. ICME2007, 2007. (accepted)