3D VIDEO PERFORMANCE SEGMENTATION

Tony Tung *Takashi Matsuyama*

Graduate School of Informatics, Kyoto University, Japan

ABSTRACT

We present a novel approach that achieves segmentation of subject body parts in 3D videos. 3D video consists in a free-viewpoint video of real-world subjects in motion immersed in a virtual world. Each 3D video frame is composed of one or several 3D models. A topology dictionary is used to cluster 3D video sequences with respect to the model topology and shape. The topology is characterized using Reeb graph-based descriptors and no prior explicit model on the subject shape is necessary to perform the clustering process. In this framework, the dictionary consists in a set of training input poses with a priori segmentation and labels. As a consequence, all identified frames of 3D video sequences can be automatically segmented. Finally, motion flows computed between consecutive frames are used to transfer segmented region labels to unidentified frames. Our method allows us to perform robust body part segmentation and tracking in 3D cinema sequences.

***Index Terms*—** 3D video, topology dictionary, shape matching, body segmentation

1. INTRODUCTION

Performance captures have become popular since the recent advance of digital technologies. Solutions produce free-viewpoint videos of real-world subjects in motion that can be immersed in virtual worlds. In particular the 3D video technology enables us to capture subjects without using any special suit or markers as opposed to motion capture methods (and thus the subjects can wear loose clothing). In order to perform the acquisitions, several calibrated and synchronized video cameras are set around the scene (e.g. a studio or a stadium). 3D video sequences are reconstructed using multiple-view stereo techniques. Each 3D video frame is therefore composed of one or several 3D models. The technique suits to a very wide range of applications such as 3D cinema, video game, medicine, sports, surveillance, etc.

We present a novel approach that achieves segmentation of subject body parts (such as head, body, limbs) in 3D videos using a topology-based shape descriptor dictionary. The dictionary is used to cluster 3D video sequences with respect to the model topology and shape. As Reeb graph-based descriptors are used as topology descriptors, no prior explicit model on the subject shape (such as a skeleton) is necessary to perform pose recognition. In our framework, the revisited dictionary consists in a set of training input sequences with a priori segmentation and labels. As a consequence, using topology and shape matching, all identified frames of 3D video sequences are automatically segmented. Finally motion flows computed between consecutive frames are used to validate region segmentation and transfer labels to unidentified regions. Our method allows us to perform robust body part segmentation and tracking in 3D cinema sequences. This can be useful, for example, to edit a body part in one frame and then automatically transfer the modification to the whole sequence. The rest of the paper is organized as follows. The next section discusses work related to the techniques presented in this paper. Section 3 makes a brief recap on the topology dictionary concept. Section 4 describes the 3D video performance segmentation process. Section 5 shows experimental results. Section 6 concludes with a discussion on our contributions.

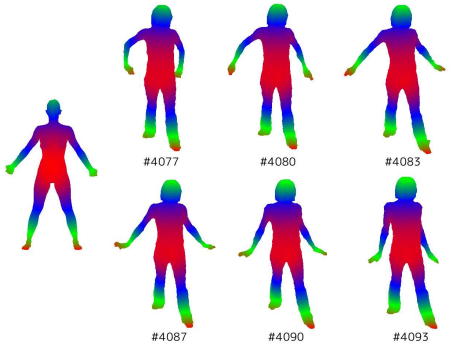

Fig. 1. **3D video performance segmentation**. 3D video sequences are modeled by a topology dictionary whose *words* contain segmentation informations. Here, a model (left) from the dictionary is used to segment a set of similar models (right) retrieved from a 3D video sequence.

2. RELATED WORK

Since a decade an increasing number of research groups have been working on 3D performance capture using multiple view camera settings [1, 2, 3, 4, 5, 6]. The systems usually require a dedicated studio or area, where video cameras are set to surround a scene (cf. Fig. 2). The video cameras are synchro-

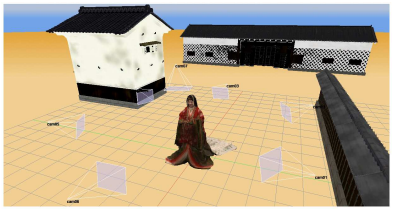

Fig. 2. **3D video performance.** The subject model is represented in a virtual environment with video cameras.

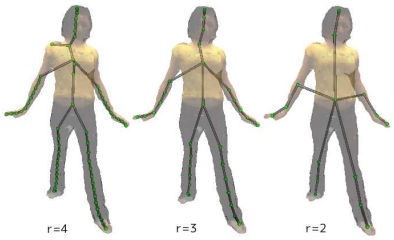

Fig. 3. **Multiresolution Reeb graphs at resolution r=4, 3 and 2**.

nized and geometrically (and eventually color) calibrated. It is then possible to reconstruct 3D videos (as a streaming of 3D models) of subjects in motion using multi-view stereo techniques applied to every set of frames simultaneously acquired. According to the literature the best performing techniques to achieve 3D shape reconstruction from multi-view images (without using additional HR device such as 3D scanner) combine silhouettes and stereo (see [7] for a survey). In addition textures are mapped onto the reconstructed 3D object surfaces to deliver high quality visual effects (e.g. for cloth rendering).

Many methods have focused on reconstruction quality [8, 9, 10, 11]. Nevertheless few have addressed the management of the overall data produced by 3D video systems whereas it is crucial as 3D video sequences are very tedious to handle due to the huge size of datasets. Recently [12, 13] have proposed methods to produce new synthesized 3D video sequences from original 3D video sequences by: (1) clustering similar frames using shape matching methods, and (2) producing comprehensive motion graphs. In particular, [13] proposed a content-based encoding strategy to achieve 3D video compression, skimming and description. 3D video sequences are modeled by a topology dictionary with Markov network: a weighted directed graph where nodes represent clusters of topology descriptors, and edges represent transitions between different poses. The topology dictionary is used to learn and index 3D video patterns. In addition, semantic annotations provide automatic video description and action recognition [14, 15].

In our framework, we extend the application of the topology dictionary to 3D video performance segmentation. To date, human body segmentation has been mainly performed in monocular videos using explicit human models (e.g. skeletal graph) [16, 17]. The system recognition ability is therefore bounded to predesigned descriptions learned from training datasets and cannot cope with arbitrary articulated models. Furthermore to our knowledge no strategy to segment 3D video performance has been proposed so far. Hence we propose to use the topology dictionary with Reeb graphs as shape descriptors to identify subject body parts in 3D videos. The Reeb graph allows us indeed to automatically extract shape and topology information without fitting a predesigned model on the model shape. The training dataset consists in extracted poses or sequences with segmented regions (and labels). As in [18, 13], the recognition step relies on a fast and efficient multiresolution topology matching process where 3D video frames are segmented as frames are identified by the system. Finally a post-processing step transfers labels from identified frames to unidentified frames using a surface matching approach [20, 19].

3. THE TOPOLOGY DICTIONARY

The topology dictionary is generated from a set of extracted patterns or learned from training input sequences. Pattern extraction is obtained by unsupervised clustering of the dataset using enhanced Reeb graphs. The Reeb graphs are canonical representations of the topology of surface models. They have been designed for shape matching in large database and can perform efficient shape retrieval queries [18]. The dictionary features a weighted directed graph structure $\mathbf{G} = (\mathbf{C}, \mathbf{E})$ where nodes (or states) $\mathbf{C} = \{c_i\}$ represent patterns and edges $\mathbf{E} = \{e_{ij}\}$ characterize topological changes (state transitions). The Markov network structure allows to draw statistical information on the video content such as duration and occurrence probability of frame sets: each edge $e_{ij} \neq \emptyset$ carries a weight w_{ij} that models the transition probability between the two states c_i and c_j. Those informations are combined with a Reeb graph multiresolution matching scheme in order to accelerate queries [13].

3.1. Multiresolution Reeb graph

The Reeb graph is a high level 3D shape descriptor. It is an elegant solution to analyze 3D mesh topology and shape as it gives a graphical representation of surface properties. As designed, Reeb graphs at coarser resolution levels $r < R$ can be derived from finer resolution representations (cf. Fig. 3). This is a great advantage in practice as it enables us to perform graph matchings at lower resolution levels, thus avoiding NP-completeness complexity. Furthermore the augmented Multiresolution Reeb Graph (aMRG) is an enhanced version of a multiresolution Reeb graph. It embeds topologi-

cal and geometrical informations in order to perform accurate shape matching in large database (cf. [18] for a detailed description).

3.2. Motion graph structure

We propose to analyze the content of training 3D data to identify poses, and encode sequences using pattern references. The search is operated by Reeb graph matchings. The dataset is clustered into topology classes **C** and a weighted directed graph **G** is built upon the timing of the sequence as a Markov network (cf. Fig. 4). The overall structure stands for the topology-based shape descriptor dictionary.

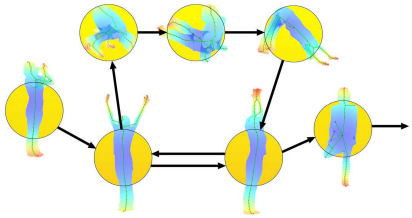

Fig. 4. **Motion graph**. 3D video sequences are clustered in topology classes.

4. PERFORMANCE SEGMENTATION

The topology dictionary introduced in [13] is revisited to serve for 3D video performance segmentation. Training datasets consist in segmented and labeled sequences. Topology and shape matchings (as in [18]) are performed to identify and automatically segment 3D video sequences. Finally motion flows computed between consecutive frames are used to transfer labels to unclassified frames. Our method allows us to perform robust body part segmentation and tracking in 3D sequences.

4.1. Body part learning

The system is trained with extracted patterns from test sequences. The patterns $\mathbf{C} = \{c_1, ..., c_N\}$ can be chosen by clustering the sequence $\mathcal{S} = \{s_1, ..., s_T\}$ using a topology-based descriptor as the Reeb graph, and selecting a pattern from each cluster c_i (cf. Fig. 4); therefore the training dataset contains segmented models with labels. In practice, models are partitioned according to Reeb graph nodes computed at a fine resolution level $r = 4$ and a region label is associated to each node (cf. Fig. 3).

4.2. Body part recognition

Similar poses of a model can be retrieved in 3D videos by queries with respect to topology and shape. The matching process involves a multiresolution strategy were similarity between Reeb graphs are evaluated using coarse-to-fine representations [18, 13]. The motion graph structure **G** (cf. 3.2) allows us to quickly select the best candidates in **C** to be evaluated based on occurrence statistics w_{ij}. Hence as a frame s is classified $s \in c_i$, all the frames belonging to the same topology cluster c_i are then segmented and labeled accordingly. Furthermore, we assume that unclassified frames $\{s \in \mathcal{S} \cap s \notin \mathbf{C}\}$ can be correctly segmented using the closest identified frame in **C**. Labels are transfered using dense surface matching [20, 19].

5. EXPERIMENTAL RESULTS

The algorithms were developed in C++ using a PC Core2Duo 3.0GHz 4GB RAM. The performances of our approach have been evaluated on real-world 3D videos of human performances such as yoga, martial arts, aerobic. Every 3D video frame contains a 3D mesh of approximatively 15000-30000 triangles with texture. The current unoptimized implementation takes 15s to generate an augmented multiresolution Reeb graph at resolution level $R = 4$. The similarity computation between two models takes 10ms. The sequences were captured in a dedicated studio using a set of 15 video cameras synchronized at 25fps. Figures 1 and 5 illustrate the segmentation performances of our approach on challenging datasets. In Fig. 1, the segmentations are performed with respect to a model from a different sequence. In Fig. 5, similar frames are efficiently retrieved and similarly segmented disregarding timeframe.

6. CONCLUSION

In this paper we present a new scheme to achieve 3D video performance segmentation. 3D video sequences of subjects in motion are modeled by a revisited topology dictionary. The dictionary consists in topology classes obtained by sequence clustering. The extracted sequences contain user-defined segmentations and labels. The dictionary features as well a comprehensive motion graph that allows us to easily navigate through the sequences. The Reeb graph is used as topology-based shape descriptor as it enables us to represent arbitrary subjects without using predesign explicit models such as human skeleton. Identified frames that are similar to a topology dictionary *word* are classified and then segmented accordingly. Finally, motion flows computed between consecutive frames are used to transfer labels to eventual unidentified frames. Our method allows us to perform robust body part segmentation and tracking in 3D cinema sequences.

7. ACKNOWLEDGMENTS

This work was supported in part by the JST-CREST project “Creation of Human-Harmonized Information Technology for Convivial

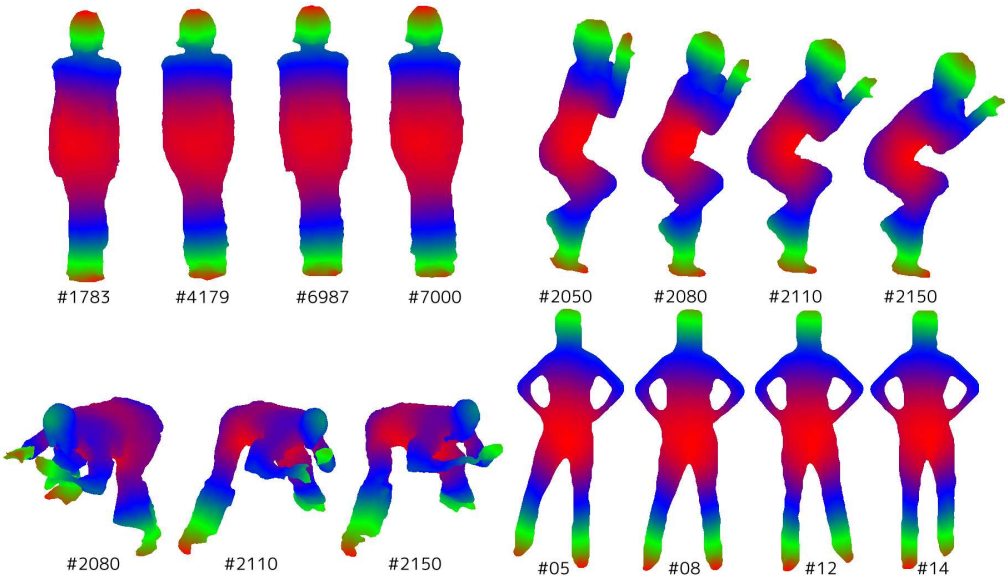

Fig. 5. **3D video performance segmentation.** The subjects are segmented according to the topology class they belong to. (Top) Yoga sequences. (Bottom left) Capoeira sequence. (Bottom right) Aerobic sequence.

Society", and the International Information Science Foundation (Grant No. 2010.1.2.066).

8. REFERENCES

[1] T. Kanade, A. Yoshida, K. Oda, H. Kano, and M. Tanaka, "A stereo machine for video-rate dense depth mapping and its new applications," *CVPR*, 1996.

[2] J. Starck and A. Hilton, "Model-based multiple view reconstruction of people," *ICCV*, 2003.

[3] T. Matsuyama, X. Wu, T. Takai, and S. Nobuhara, "Real-time 3d shape reconstruction, dynamic 3d mesh deformation, and high fidelity visualization for 3d video," *CVIU*, vol. 96, no. 3, pp. 393–434, 2004.

[4] J.S. Franco, C. Menier, E. Boyer, and B. Raffin, "A distributed approach for real-time 3d modeling," *CVPR Workshop on Real-Time 3D Sensors and their Applications*, p. 31, 2004.

[5] K. M. Cheung, S. Baker, and T. Kanade, "Shape-from-silhouette across time: Part ii: Applications to human modeling and markerless motion tracking," *IJCV*, vol. 63, no. 3, pp. 225–245, 2005.

[6] J. Allard, C. Ménier, B. Raffin, E. Boyer, and F. Faure, "Grimage: Markerless 3d interactions," *SIGGRAPH - Emerging Technologies*, 2007.

[7] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," *CVPR*, 2006.

[8] J. Starck and A. Hilton, "Surface capture for performance-based animation," *IEEE Computer Graphics and Applications*, 2007.

[9] E. De Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun, "Performance capture from sparse multi-view video," *SIGGRAPH*, 2008.

[10] T. Tung, S. Nobuhara, and T.Matsuyama, "Simultaneaous super-resolution and 3d video using graph-cuts," *CVPR*, 2008.

[11] T. Tung, S. Nobuhara, and T.Matsuyama, "Complete multi-view reconstruction of dynamic scenes from probabilistic fusion of narrow and wide baseline stereo," *ICCV*, 2009.

[12] P. Huang, A. Hilton, and J. Starck, "Human motion synthesis from 3d video," *CVPR*, 2009.

[13] T. Tung and T.Matsuyama, "Topology dictionary with markov model for 3d video content-based skimming and description," *CVPR*, 2009.

[14] J. Sullivan and S. Carlsson, "Recognizing and tracking human action," *ECCV*, 2002.

[15] D. Weinland, E. Boyer, and R. Ronflard, "Action recognition from arbitrary views using 3d exemplars," *ICCV*, 2007.

[16] G. Mori, X. Ren, A. A. Efros, and J. Malik, "Recovering human body configurations: Combining segmentation and recognition," *CVPR*, 2004.

[17] X. Xu and B. Li, "Learning motion correlation for tracking articulated human body with a rao-blackwellised particle filter," *ICCV*, 2007.

[18] T. Tung and F. Schmitt, "The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes," *Int. Jour. of Shape Modeling*, vol. 11, no. 1, pp. 91–120, 2005.

[19] T. Tung and T.Matsuyama, "Dynamic surface matching by geodesic mapping for 3d animation transfer," *CVPR*, 2010.

[20] J. Starck and A. Hilton, "Correspondence labelling for wide-timeframe free-form surface matching," *ICCV*, 2007.