A 3D Shape Descriptor for Segmentation of Unstructured Meshes into Segment-wise Coherent Mesh Series

Tomoyuki Mukasa Shohei Nobuhara Tony Tung Takashi Matsuyama
Graduate School of Informatics, Kyoto University
Yoshida Honmachi, Sakyo-Ku, Kyoto, Japan, 606–8501
mks@vision.kuee.kyoto-u.ac.jp

Abstract

This paper presents a novel shape descriptor for topology-based segmentation of 3D video sequence. 3D video is a series of 3D meshes without temporal correspondences which benefit for applications including compression, motion analysis, and kinematic editing. In 3D video, both 3D mesh connectivities and the global surface topology can change frame by frame. This characteristic prevents from making accurate temporal correspondences through the entire 3D mesh series. To overcome this difficulty, we propose a two-step strategy which decomposes the entire sequence into a series of topologically coherent segments using our new shape descriptor, and then estimates temporal correspondences on a per-segment basis. We demonstrate the robustness and accuracy of the shape descriptor on real data which consist of large non-rigid motion and reconstruction errors.

1. Introduction

3D surface reconstruction from multi-view videos *i.e.*, 3D video, has become a popular technique in computer vision and graphics communities [13] [6]. 3D video consists of a temporal series of 3D surfaces reconstructed on a frame-by-frame basis from multi-view videos of real-world objects. The captured 3D surface meshes are unstructured, *i.e.*, are not semantically labeled, and can have different numbers of vertices and mesh connectivities. This fact indicates that no vertex-to-vertex correspondences between different meshes are available in general. Moreover 3D video can only capture surfaces that are visible from the cameras, *i.e.*, the envelopes. That is, not only the local mesh structure but the global surface topology also can change through the entire 3D video sequence (Fig.1).

On the other hand, once a time-invariant structure, *e.g.*, kinematic structure describing the object motion is obtained, a wide variety of 3D video applications such as motion analysis of dance or sports activities, kinematic editing of captured data, inter-frame mesh data compression, *etc.* will be realized.

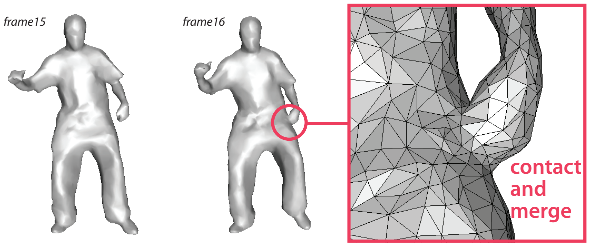

Figure 1. The global topology change of surface. Left is a genus-0 surface while Right is a genus-1 due to the body contact.

To obtain a time-invariant structure, it is desirable that the captured 3D surfaces are converted to a temporally coherent mesh sequence such that all the meshes share a same number of vertices and their connectivities.. Various techniques have been investigated to acquire a temporally coherent structure in mesh sequence [5] [2] [16] .However, [2] does not explicitly account for global topology changes caused by the invisibility in the 3D video production process. Although [5] learns fixed mesh topology from an unstructured mesh sequence, there is a possibility to be disturbed by deficits and artifacts (See Fig.11) which are inevitable in real data reconstruction. [16] acquires temporally coherent segmentations of a mesh sequence. As it was mainly based on ICP algorithm in Euclidean space, its capability was limited to small displacements. Hence applying such techniques to long sequences involving complex motions can fail in theory.

Based on these observations, we propose a two-step scheme that segments the original 3D mesh sequences into time intervals in which the global topology is unchanged, and then applies a remeshing process which converts each segment so as to be represented by a single deforming mesh. To this end, this paper proposes a novel shape descriptor for the 3D surface sequence segmentation that is robust against apparent global surface topology changes falsely caused by observation noise.

The rest of the paper is organized as follows. Section 2 discusses the work related to our technique. Section 3 describes the our topology-based shape descriptor. Section 4 shows experimental results. Final section concludes with a discussion on our contribution.

2. Related Work

In the context of computer vision and computer graphics, various types of 3D shape descriptors has been investigated in the recent years.. They are roughly classified into two groups, one is for static shapes and the other for dynamic shapes. The former is mainly designed for shape retrieval, and is optimized for finding similar 3D shapes that are not necessarily a same object of different postures [10] [4]. In general, these descriptors aim to discriminate different objects. On the other hand, the latter considers the shape matching problem of a same object in motion described as a temporal sequence, *i.e*., 3D video. That is, given a time-series of 3D surfaces, they try to describe a similarity of frame pairs. Our contribution belongs to the latter ones.

Shape descriptors for 3D video are also broadly classified by means of how it handles the global topology change of surface meshes. Huang *et al*. proposed Shape Histogram [1] [3] which aims to retrieve similar poses in the sequence for smooth motion transitions of 3D video. Shape Histogram is designed to be robust to global topology change by utilizing a volumetric occupancy of the shape.

In contrast, Mukasa *et al*. have proposed a shape descriptor designed to detect global topology changes for kinematic structure estimation while ignoring shape deformations[9]. Their descriptor is a histogram of distribution of μ values defined as the sum of geodesic distances from a vertex to all the others, and their similarity is defined as the correlations of the histograms.

This method shares the same motivation with this paper, but it is not discriminative enough to handle global topology changes in real data with small holes and loops due to observation noise and reconstruction errors. This is because it integrates the geodesic distances over the surface, and smoothes out changes of vertex-wise geodesic distances caused by a global topology change.

Based on this observation, we introduce a part-wise geodesic histogram description of 3D mesh. The key idea is to identify body parts having a prominent geometric characteristic first, and then define the geodesic histogram on a part-wise basis in order to avoid mixing them up. The evaluations in Section 4 demonstrates that our part-wise geodesic histogram successfully account for noisy inputs while the conventional method cannot.

3. Part-wise Geodesic Histogram Shape Descriptor

As described in Section 2, our descriptor requires identifying body regions having prominent geometric properties first. The key point here is how we can identify a position on a surface in another surface of a different frame. Our algorithm utilizes extreme points of the 3D surfaces. That is, our strategy is first extracts extreme points from a pair of 3D surfaces in question, and then tries to establish correspondences between them. Intuitively this part returns a set of sparse corresponding points between two surfaces.

Once obtained such sparse corresponding points, then the next step is to decompose the surface into disjoint subsurfaces each of which includes a corresponding point as an anchor point.

Finally we compute a histogram of geodesic distances for each subsurface and integrate them to be a single descriptor named part-wise geodesic histogram shape descriptor (PGH-SD).

In what follows, we introduce these three steps, and then provide the definition of the similarity between PGH-SDs.

3.1. Sparse Corresponding Points Estimation

As described above, our descriptor utilizes surface extreme points, *i.e*. tips, to obtain 3D points which appear stably on surfaces under deformation. In particular, we utilize the integral of geodesic distances μ and Morse theory.

3.1.1 μ Computation

Let us denote a 3D surface mesh sequence by $M(t) = \{V(t), E(t)\}$ where $V(t)$ and $E(t)$ denote the set of vertices and edges at frame t respectively. We define a continuous function $\mu : M(t) \to \mathbb{R}$ as the sum of geodesic distances from a vertex to all the others :

$$\mu(v) = \sum_{u \in V(t)} g(v, u), \tag{1}$$

where $v, u \in V(t)$, and $g(v, u)$ is the geodesic distance between v and u. If the global surface topology changes, the distribution of μ changes drastically because topological changes introduce or remove shortest paths on $M(t)$ (Fig.2). On the other hand, since we define μ as an integral over the surface, its distribution is robust to local surface deformations caused by object motions or per-vertex reconstruction errors.

3.1.2 Tip Detection and Correspondence Estimation

According to the Morse theory[8], a continuous function μ defined on a surface can characterize the surface topology using its critical points. As we choose the integral of

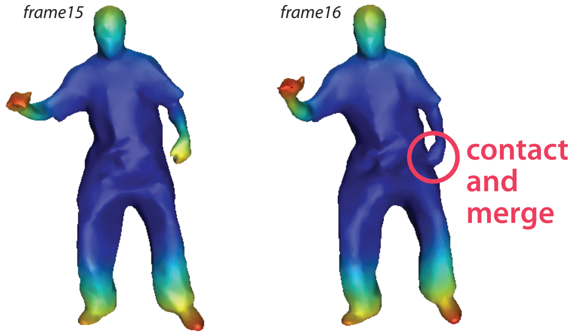

Figure 2. μ distribution on surface mesh. Warm color indicates a higher μ value.

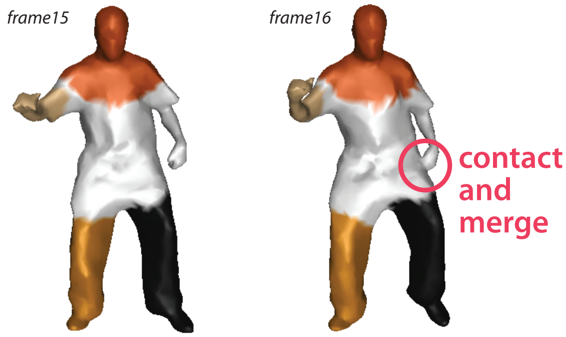

Figure 3. Segmentation results. Note that left hand is merged to stem area as a tip of left hand was not found in frame 16 because of body contact.

geodesic distance as μ, critical points coincide to highly concave or convex regions [7] [15].

Tips of body extremities, *e.g.*, fingertips, appear in $\mu(v)$ distribution on the surface as its local maxima. By assuming the frame rate of 3D video is high enough for capturing the object's motion, these points can be tracked based on Euclidean distance over time in each pair of successive 3D video frames.

We find tips as local surface feature points $v_n^{tip}(t) \in L(t)(n = 1, \ldots, N(L(t)))$ at local maxima of $\mu(v)$ at each time frame t, and take $L(t)$ as the tips. Here we introduce a function $E(v, L)$ which returns a tip $v_n^{tip} \in L$ nearest to v in Euclidean space. We then find mapping $F_{t,t+1}$ between $L(t)$ and $L(t+1)$ as follows:

$$\begin{aligned} F_{t,t+1} = \{ \quad < \quad & v_n^{tip} \in L(t), v_m^{tip} \in L(t+1) > | \\ & v_n^{tip} = E(v_m^{tip}, L(t)), \\ & v_m^{tip} = E(v_n^{tip}, L(t+1)), \\ & \mathbf{n}(v_n^{tip}(t)) \cdot \mathbf{n}(v_m^{tip}(t+1)) > 0\}, \quad (2) \end{aligned}$$

where $\mathbf{n}(v)$ stands for the normal vector of v. As the result, we find a number of tips L_i in each pair of 3D video frames, and establish their correspondences.

3.2. Surface Segmentation

Up to this point, we have obtained a set of corresponding tips for a given pair of 3D surfaces. The goal of this section is to decompose each surface into a set of disjoint subsurfaces each of which includes a corresponding point as an anchor. Here, the point is to make the subsurfaces having corresponding tips to be identical to each other even under deformations. We realize this by assuming that (1) subsurfaces including tips can be modeled as a generalized cone, and (2) the entire surface can be modeled as a union of subsurfaces including tips and the other regions called stems.

For the preparation of geodesic histogram computation, the entire surface mesh is segmented into sub-surfaces which correspond to body extremities and body stem.

We first apply Voronoi segmentation to entire surface using tips as seed points. As the result, we have a set of sub-surfaces $\{S_n(t)\}$ corresponding to tips $\{v_n^{tip}(t)\}$. Secondly, in each $S_n(t)$, we find point $p_n(t)$ which is geodesically nearest to $v_n^{tip}(t)$ on the border lines between sub-surfaces. Then we define geodesic range g_n^{max} as follows:

$$g_n^{max} = \min\{g_n^{tip}(p_n(t)), g_n^{tip}(p_n(t+1))\} \qquad (3)$$

$$\text{where } g_n^{tip}(p_n(t)) = g(v_n^{tip}(t), p_n(t)). \qquad (4)$$

We remove vertices from each sub-surface if the vertex $v(t) \in S_n(t)$ meet the following condition:

$$g_n^{tip}(v(t)) > g_n^{max}. \qquad (5)$$

We merge these removed vertices into a sub-surface $S_{n+1}(t)$ which is corresponding to body stem. Fig.3 shows the example of our surface segmentation.

3.3. Definition of PGH-SD

To define our PGH-SD using the subsurfaces, this section introduces a geodesic coordinates based on the tips, and then combines them as histograms.

As we have made correspondences between tips v_n^{tip} between frames in the last section, we can define a geodesic coordinate $\boldsymbol{x}_i^{geo}(v(t))$ whose element is the geodesic distances to the tips as follows:

$$\boldsymbol{x}_i^{geo}(v(t)) = (g_1^{tip}(v(t)), ..., g_{N(L)}^{tip}(v(t))). \qquad (6)$$

When the global topology does not change between frames, the distance between each vertex $v(t) \in V(t)$ and each of tips $v_n^{tip}(t)$ does not change even if the object has deformed. Therefore geodesic coordinate is invariant against mesh deformations between frames if there are no global topology changes [14].

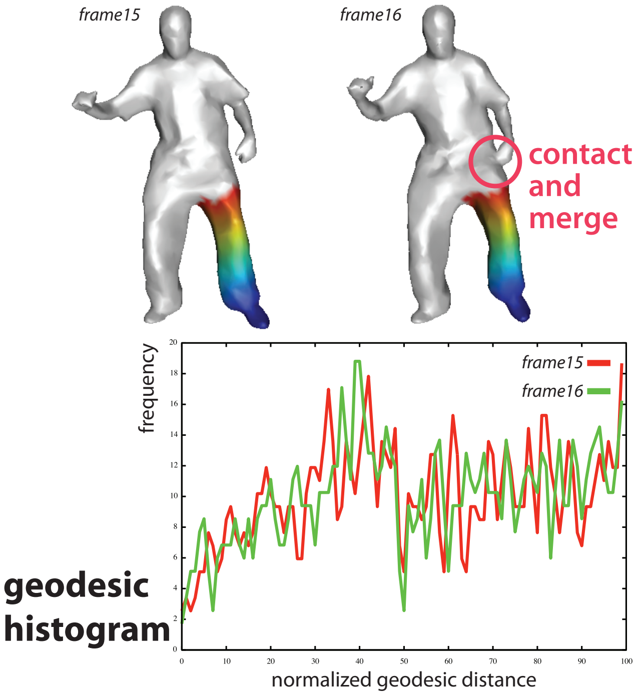

Figure 4. Distributions of geodesic distance to left leg tip on left leg sub-surfaces and their histograms. The vertical axis indicates the number of vertices whose geodesic distance is in a range correlated to a bin.

Our shape descriptor is a set of geodesic histograms $\{H^{geo}_{n,m}(t)\}$ where $m = \{1, ..., n\}$ denotes geodesic coordinate corresponding to a tip. We call the shape descriptor as Part-wise Geodesic Histogram Shape Descriptor (PGH-SD). For each sub-surface $S_n(t)$ and tip m, we compute histograms $\{H^{geo}_{n,m}(t)\}$ of distribution of m-th geodesic coordinate value of vertices in $S_n(t)$ (Fig. 4, 5).

3.4. Similarity Computation

We employ Earth Mover's Distance (EMD) [17] for computing distance between 3D video frames. EMD is defined as the minimum amount of work to transform one histogram into the other, and robust to outliers and small shifts of values among histogram bins. Given two histograms $H^1 = (h^1_1, ..., h^1_l)$ and $H^2 = (h^2_1, ..., h^2_l)$, each having l bins, a flow matrix F, where $f_{i,j}$ indicates flow to move from h^1_i to h^2_j, and a cost matrix C, where $c_{i,j}$ models cost of moving flow from the i-th bin to the j-th bin, EMD can be defined as follows:

$$EMD(H^1, H^2) = \min_F \sum_{i=1}^{l} \sum_{j=1}^{l} f_{i,j} c_{i,j}. \quad (7)$$

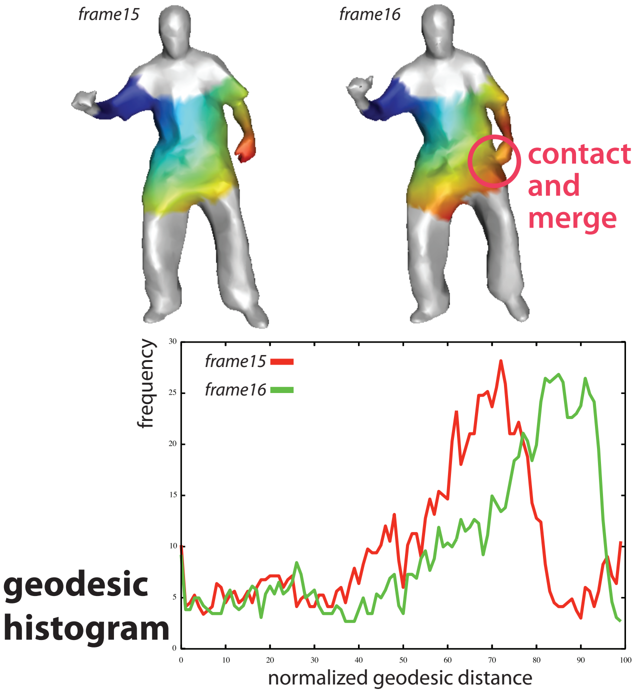

Figure 5. Distributions of geodesic distance to right hand tip on body stem sub-surfaces and their histograms.

We define the distance $D(t, t+1)$ between 3D video frames as a weighted sum of EMD of corresponding geodesic histograms as follows:

$$D(t, t+1) = \sum_n w_n \sum_m EMD(H^{geo}_{n,m}(t), H^{geo}_{n,m}(t+1)), \quad (8)$$

where w_n is a weight factor for sub-surface $S_n(t)$ which is size ratio of $S_n(t)$ to the entire surface area. In Fig.4, the histograms are similar and EMD is small. In contrast, EMD is large as the distribution of geodesic distance is significantly affected by the contact between the left hand and body stem in Fig.5.

We segment entire 3D video sequence at each pair of frames if their distance $D(t, t+1)$ is larger than a threshold τ.

4. Experiments

This section evaluates the performance of the proposed method using a real dataset called *free* [12] in which a subject wearing a loose cloth performs a break dance. The sequence contains 291 frames of unstructured meshes each of which consists of around 3,000 vertices.

4.1. Comparison to Existing Methods

We implemented following methods for comparison:

1. μ-histogram shape descriptor (μ-SD) [9]
2. Geodesic histogram shape descriptor without surface segmentation (GH-SD).

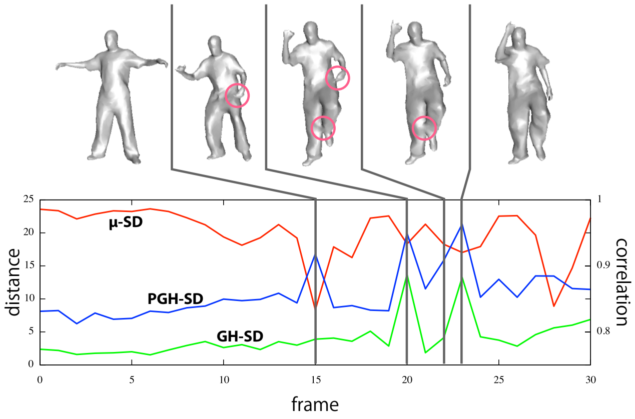

Figure 6. Comparison of three methods. Circles on the top figure indicate body contacts.

μ-SD is given by the following correlation coefficient index defined by covariance Cov and standard deviation σ to measure the surface-to-surface topological difference:

$$C(H^{\mu}(t), H^{\mu}(t+1)) = \frac{\mathrm{Cov}(H^{\mu}(t), H^{\mu}(t+1))}{\sigma(H^{\mu}(t))\sigma(H^{\mu}(t+1))}, \quad (9)$$

where $H^{\mu}(t)$ denotes the histogram of μ distribution on surface mesh $M(t)$. Then, distance $D^{\mu}(t, t+1)$ between 3D video frames can be defined as follows:

$$D^{\mu}(t, t+1) = 1 - C(H^{\mu}(t), H^{\mu}(t+1)). \quad (10)$$

Another method for comparison, GH-SD is a simplified version of the proposed method to evaluate the importance of the surface segmentation in our method. GH-SD is the set of geodesic histograms $\{H^{geo}_{m}(t)\}$ defined on the entire surface mesh. Therefore, based on GH-SD, we can define the distance $D^{GH}(t, t+1)$ between 3D video frames as follows:

$$D^{GH}(t, t+1) = \sum_{m} EMD(H^{geo}_{m}(t), H^{geo}_{m}(t+1)). \quad (11)$$

All algorithms were implemented in C++ using an Intel Core-i7 2.3GHz computer.

Figure 6 reports the frame-by-frame distances returned by the three methods from frame 0 to 30 for illustration. The vertical lines indicate the segmentation boundaries where global topology changes occur. These boundaries are given by hand as the ground truth. The rendered 3D shapes on top of the plots illustrate a typical pose of each segments, and the red circles show where the body parts contacted and made global topology changes. the *free* sequence is segmented based on a thresholding for $D^{\mu}(t, t+1)$, $D^{GH}(t, t+1)$ and $D(t, t+1)$. The detection results of segment points are shown in Fig.7 and summarized in Table 1 and Fig.8. Evaluation of Receiver Operator Characteristics (Fig.8) for segmentation against ground truth, shows that the proposed method most correctly reflects the global topology change in the sequence. In Table 1, the proposed method returns no false negatives which is crucial requirement for further mesh alignment relying on the global topology consistency.

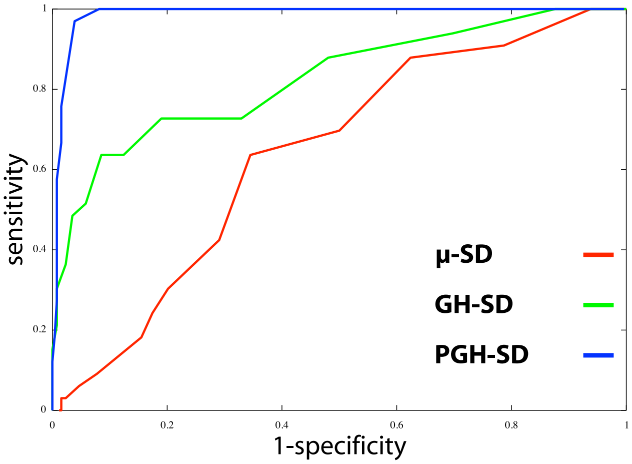

Figure 8. Evaluation of ROC performance of three methods.

4.2. Segment-wise Coherent Mesh Series

Based on the criteria $D(t, t+1)$ described in above sections, the entire 3D video sequence is segmented into a set of intervals $\{I_i\}$. Here, we apply tracking-by-matching approach to acquire segment-wise coherent mesh series from each mesh interval. In particular, we used the geodesic mapping[14] since the geodesic coordinate required in [14] has already been computed for our shape descriptor.

Once obtained vertex-wise correspondences between frames, we deformed the mesh of a reference frame so as to fit to the others. We choose a 3D video frame in the middle of each interval as the reference mesh M_i^{ref} because it is likely to represent an average posture of the object in the interval and hence is likely to minimize the deformation artifacts.

Then we establish per-vertex correspondences between successive frames $f^{geo}_{t\rightarrow t+1} : v(t) \in V(t) \rightarrow v(t+1) \in V(t+1)$ by finding a point $v(t+1)$ such that:

$$v(t+1) = \underset{v' \in V(t+1)}{\arg\min}\,(d(v(t), v')), \quad (12)$$

where $d(v(t), v')$ denotes the geodesic distance between vertices in different frames in the same interval as follows:

$$d(v \in V(t1 \in I_i), v' \in V(t2 \in I_i)) = ||\boldsymbol{x}^{geo}_{i}(v) - \boldsymbol{x}^{geo}_{i}(v')||. \quad (13)$$

If we move each vertex of $M_{t\in I_i}$ according to $f^{geo}_{t\rightarrow t+1}$, it can introduce self-intersections because **Eq.**(12) returns the

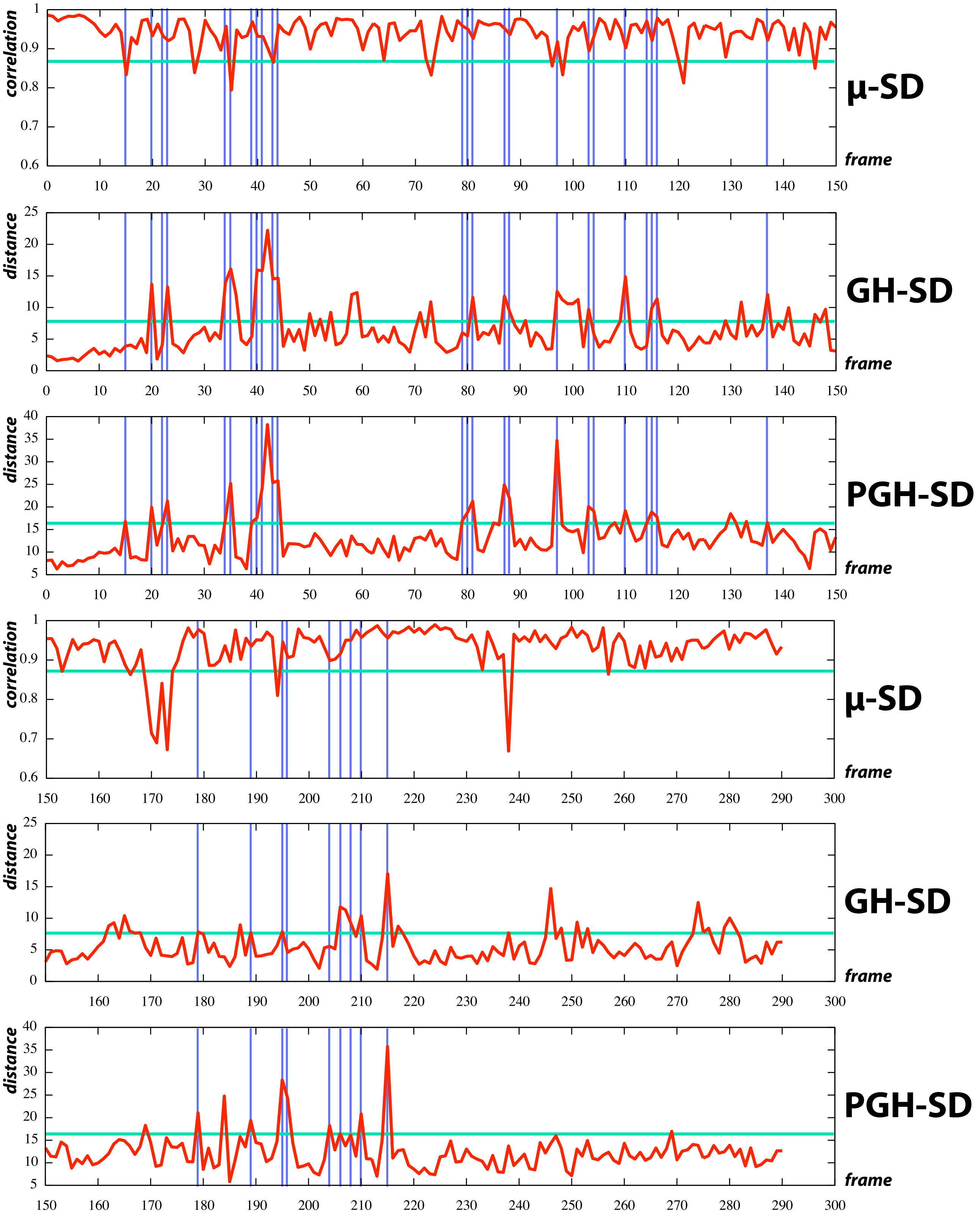

Figure 7. Segmentation results using a typical threshold for each method. The blue vertical lines indicates ground truth segmentation boundaries . The green horizontal ones shows thresholds for each method.

method	threshold	True Positive	True Negative	False Positive	False Negative
μ-SD	0.87	3	243	14	31
GH-SD	7.50	26	221	37	7
PGH-SD	16.0	36	247	9	0

Table 1. Detection result of temporal segments.

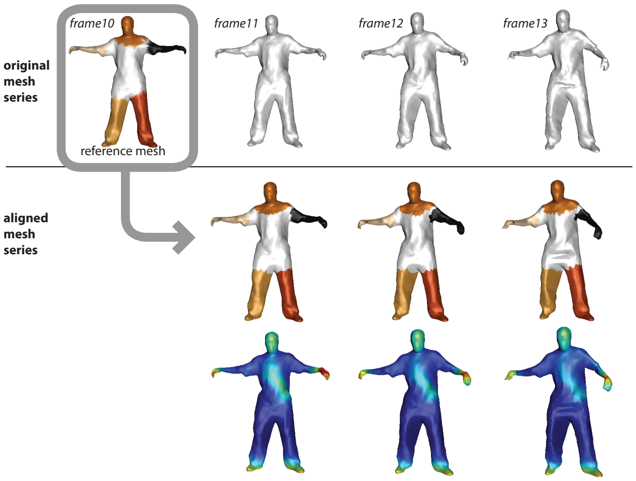

Figure 9. Segment-wise coherent mesh series acquired by mesh alignment based on geodesic mapping. The middle row shows how the surface segmentation on the reference mesh is propagated over frames. In the bottom figures, warm color indicates high residual error to original mesh.

geodesically nearest vertex without considering connectivities among vertices. To enforce a smooth deformation, we employ as-rigid-as-possible (ARAP) deformation method [11] to M_i^{ref} using $f_{i\to j}^{geo}$ as the soft constraint. ARAP deformation preserves surface details by keeping rigidities of each local area around vertex while the whole mesh is deformed so as to satisfy the soft constraint. With the combination of geodesic mapping and ARAP deformation, we replace 3D mesh sequence $M_{t\in I_i}$ by $M'_{t\in I_i}$ in which mesh topology is guaranteed to be equal to the reference mesh M_i^{ref}. Fig. 9 shows structured mesh sequence acquired by above explained deformation. By restricting the deformation only in each interval, we can avoid potential danger for erroneous deformations (see Fig.10) caused by global topology changes.

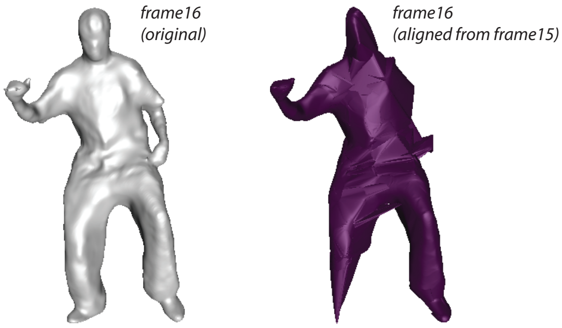

Figure 10. Example of erroneous deformation caused by the alignment over the segment point.

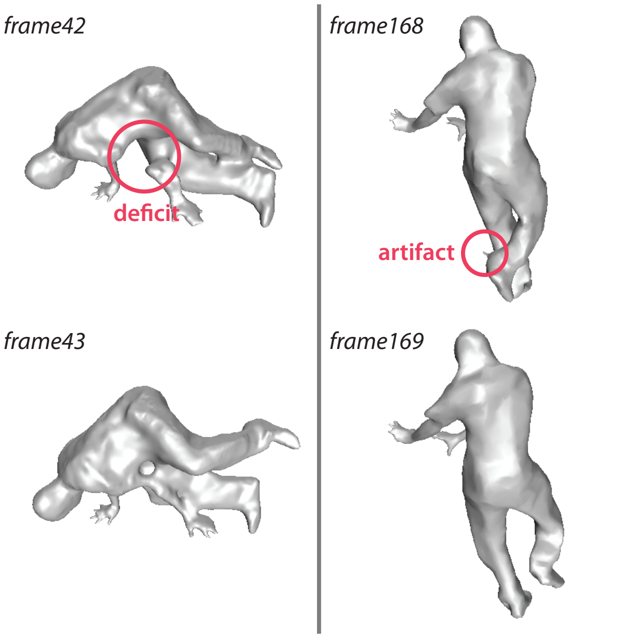

Figure 11. Reconstruction errors. In the top left figure, we can see a deficit of left upper hand. On the other hand, an artifact is found in right leg in the top right figure.

5. Discussion

As reported in Section 4.1, proposed method shows a high enough sensitivity to global topology change. However, the detection results contains a few false positives (See Table.1). As we can observe in Fig.11, most of false positives occurred between pairs of 3D video frames in which reconstruction error can be found. In practice, these erroneous 3D video frames are undesirable to be included in an interval as it may cause erroneous mapping in tracking-by-deformation step. Thanks to this sensitivity to erroneous 3D video frames, deformation result contains low residual errors as we have seen in the last Section 4.2.

6. Conclusion

In this paper, we proposed a new shape descriptor for incoherent 3D video frame sequence. The contribution of this paper is to introduce a practical shape descriptor which explicitly accounts for global topology changes and works on real data containing large non-rigid motion and reconstruction errors. The proposed shape descriptor showed the ability to segment the sequence into intervals in which global topology is coherent. The evaluation demonstrated that we can obtain segment-wise coherent mesh series with low residual errors from each interval by tracking-by-deformation based on geodesic mapping.

We will estimate a time invariant structure in each interval, and integrate them to a unique kinematic structure in future work.

Acknowledgment

This research is partially supported by JSPS Bilateral Joint Research Project "Visual Gesture Perception".

References

[1] M. Ankerst, G. Kastrnmuller, H. P. Kriegel, and T. Seidl. 3d shape histograms for similarity search and classification in spatial databases. In *Advances in Spatial Databases Lecture Notes in Computer Science*, volume 1651, pages 207–226, 1999. 2

[2] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, 2010. 1

[3] P. Huang, J. Starck, and A. Hilton. Temporal 3d shape matching. In *The 4th European Conference on Visual Media Production*, pages 1–10, 2007. 2

[4] J.W.H.Tangelder and R.C.Veltkamp. A survey of content based 3d shape retrieval methods. In *Shape Modeling International*, pages 145–156, 2004. 2

[5] A. Letouzey and E. Boyer. Progressive shape models. In *CVPR*, 2012. 1

[6] T. Matsuyama, S. Nobuhara, T. Takai, and T. Tung. *3D Video and Its Applications*. Springer, 2012. 1

[7] M.Hilaga, Y.Shinagawa, T.Kohmura, and T. Kunii. Topology matching for fully automatic similarity estimation of 3d shapes. In *SIGGRAPH*, pages 203–212, 2001. 3

[8] M. Morse. The calculus of variations in the large. *American mathematical Society, Colloquium Publication*, 18, 1934. 2

[9] T. Mukasa, S. Nobuhara, T. Tung, and T. Matsuyama. Tree-structured mesoscopic surface characterization for kinematic structure estimation from 3d video. *IPSJ Transactions on Computer Vision and Applications*, 6, 2014. 2, 4

[10] N.Iyer, S.Jayanti, K.Lou, Y.Kalyanaraman, and K.Ramani. Three-dimensional shape searching: state-of-art review and future trends. *Computer-Aided Design*, 37(5):509–530, 2005. 2

[11] O. Sorkine and M. Alexa. As-rigid-as-possible surface modeling. In *Eurographics Symposium on Geometry Processing*, 2007. 7

[12] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE Comput. Graph. Appl.*, 27(3):21–31, May 2007. 4

[13] J. Starck, A. Maki, S. Nobuhara, A. Hilton, and T. Matsuyam. The multiple-camera 3-d production studio. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(6):856–869, 2009. 1

[14] T. Tung and T. Matsuyama. Dynamic surface matching by geodesic mapping for 3d animation transfer. In *CVPR*, 2010. 3, 5

[15] T. Tung and F. Schmitt. The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes. *International Journal of Shape Modeling*, 11(1):91–120, 2005. 3

[16] K. Varanasi and E. Boyer. Temporally coherent segmentation of 3d reconstructions. In *The 5th International Symposium on 3D Data Processing, Visualization and Transmission*, 2010. 1

[17] C. T. Y. Rubner and L. J. Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. 4