The Visual Computer, 2014 - Draft - manuscript No.
(will be inserted by the editor)

Invariant Shape Descriptor for 3D Video Encoding

Tony Tung · Takashi Matsuyama

Received: date / Accepted: date

Abstract This paper presents a novel approach to represent spatio-temporal visual information. We introduce a surface-based shape model whose structure is invariant to surface variations over time to describe 3D dynamic surfaces (e.g., 3D video obtained from multiview video capture). The descriptor is defined as a graph lying on object surfaces and anchored to invariant local features (e.g., surface point extrema). Geodesic consistency based priors are used as cues within a probabilistic framework to maintain the graph invariant, even though the surfaces undergo non-rigid deformations. Our contribution brings to 3D geometric data a temporally invariant structure that relies only on intrinsic surface properties, and is independent of surface parametrization (i.e., surface mesh connectivity). The proposed descriptor can therefore be used for efficient dynamic surface encoding, through transformation into 2D (geometry) images, as its structure can provide an invariant representation for dynamic 3D mesh models. Various experiments on challenging publicly available datasets are performed to assess invariant property and performance of the descriptor.

Keywords Invariant shape descriptor · Dynamic surface · Geometry image · 3D video · Reeb graph

1 Introduction

It is one of the major goals of natural sciences to find invariant properties. In the 90s, computer vision scientists found several projectively invariant properties (e.g., viewpoint, illumination and curvature invariants) to characterize 3D object shape for recognition tasks [16, 34]. As it is difficult to find invariants on general 3D shapes that are not planar (or simple), local descriptors are used as well to model invariants and represent 3D object surface as a collection of small patches [39]. In this paper, we propose a new invariant surface-based shape descriptor for dynamic geometric objects, that is invariant to surface parameterization (e.g., surface mesh complexity or connectivity) and visual features (e.g., texture) as it relies only on intrinsic surface properties and geodesic paths. The descriptor is defined as a graph lying on object surface and anchored to invariant local features (e.g., extremal points). The graph structure invariance relies on the fact that geodesic paths are theoretically invariant to surface parametrization. Positions of graph edges and nodes are optimized using a Bayesian probabilistic framework driven by two

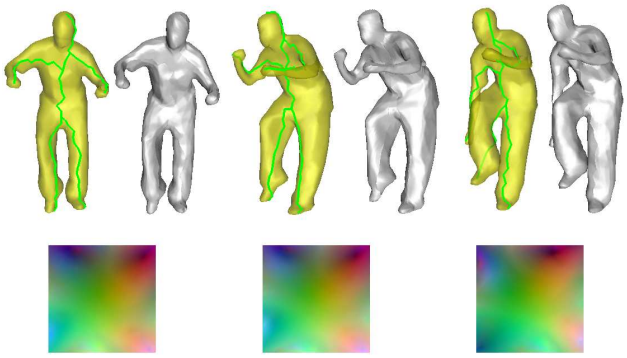


Fig. 1 The proposed invariant surface-based shape descriptor is shown in green, geometry images from planar parametrization are shown at the bottom, and corresponding reconstructed surfaces in gray. Despite deformations, dynamic surface parametrization remains unchanged across time (i.e., geometry images are similar). Hence sequence encoding can be optimal.

Kyoto University, Graduate School of Informatics, Yoshida-Honmachi, Sakyo-ku, Kyoto, Japan
E-mail: tony2ng@gmail.com, E-mail: tm@i.kyoto-u.ac.jp

geodesic consistency cues that handles ambiguities introduced by numerical approximations: when surfaces undergo non-rigid deformations over time, the overall graph structure remains invariant to surface variations (see Fig. 1). We show that the descriptor can be applied for efficient encoding of 3D video data (or free-viewpoint video), which is becoming a popular media developed by several research laboratories from all over the world (e.g., Japan [29], France [3], UK [44], Germany [1], etc.). See Fig. 2 for an example of 3D video captured by Kyoto University [30].

As each 3D video frame is usually reconstructed individually (to avoid error propagation across time), the produced 3D surface models have no geometric consistency between each other: vertex number and mesh connectivity are different. It is then not trivial to find an optimal encoding scheme for the data structure, while it is a crucial problem as (high resolution) 3D video data usually require lots of storage space (i.e., gigabytes for few minutes). Moreover, as no adaptive resolution streaming mechanism exists for 3D video, communication and telepresence applications are still tedious on low-bandwidth networks. Although 3D video data can be post-processed to obtain meshes with consistent topology and connectivity (see Sect. 2), how to cope with geometry variations is still unclear (e.g., when the mesh resolution has to dynamically change). However, the proposed invariant surface-based shape descriptor can be used to define cut graphs that cut open surface meshes for parameterization into a square domain. As the cut graphs are invariant regardless of mesh resolution, 3D video data can be transformed into sequences of 2D images that are suitable for any 2D video encoding technology (e.g., MPEG-4). This strategy is inspired from the geometry image technique proposed by [18].

A preliminary version of the proposed model was presented in [48]. However, we give here more technical details on deformation invariant shape representation (and particularly on surface extremal point extraction using Reeb graphs), discussions on 3D video data encoding (e.g., sensitivity with respect to surface noise, and difference with an approach using topology-based shape descriptor), additional quantitative evaluations and related work references. Related work is discussed in Sect. 2. The invariant surface-based shape descriptor is presented in Sect. 3. Section 4 introduces 3D video data encoding using the proposed model. Section 5 describes various experiments on challenging datasets. Section 6 concludes the paper with discussions.

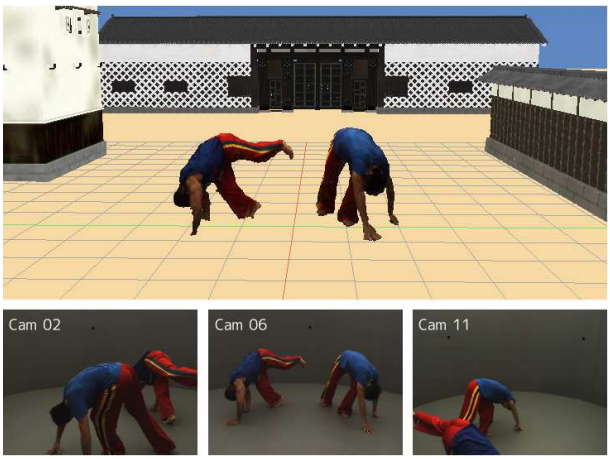


Fig. 2 3D video data captured at Kyoto University [30]. Top 3D video data reconstructed frommulti-view video cameras. Bottom examples of video frames

2 Related work

Multi-view sensing systems have become ubiquitous during the past decade as image sensing technologies are rapidly evolving. Dynamic scenes can be captured using fixed of active vision video cameras [24, 29, 17, 3], broadcast cameras [44], 3D laser scanner [1], or even handheld depth cameras [23]. Full 3D models of dynamic surfaces representing a scene can be reconstructed using multiple view stereo reconstruction techniques applied frame-by-frame (see [42] for a survey). Unlike depth maps (2.5D data) which are unclosed surface, 3D video data represent objects in full 3D as a sequence of reconstructed closed surfaces. This technology has potentially several applications in medicine, culture, communication, entertainment, etc. A review of the whole 3D video reconstruction process and applications is presented in [30].

To encode a sequence of 3D meshes, the state-of-the-art consists mainly of: (1) methods to compress every frame independently (see survey [4]); however since redundant information between frames is not managed, encoding cannot be optimal. (2) techniques designed for 3D animation sequences (e.g., using principal components [2], geometry videos [7], linear prediction coding [25], Frame-based Animated Mesh Compression of MPEG4 [28], etc.); however as they are dedicated to meshes sharing the same connectivity, they cannot be applied directly to 3D video data. A step for surface alignment across time after surface reconstruction is necessary as post-processing (e.g., using spherical matching [43], multidimensional scaling [8], template deformation [52], patch-based surface tracking [9], geodesic mapping [47], etc.), but would not be sufficient for applications requiring adaptive bitrate streaming.

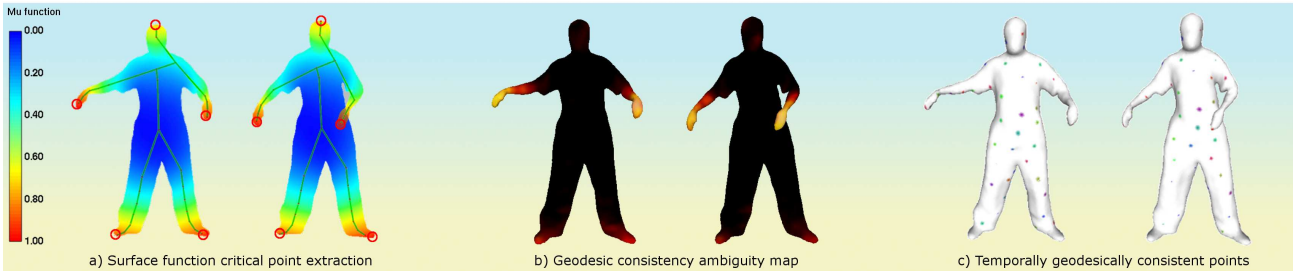


Fig. 3 Temporal geodesic consistency. a) Critical points extracted on surface mesh using Reeb graphs [50]. b) Geodesic consistency ambiguity map (darker means less position ambiguity). c) 50 temporally consistent points chosen randomly.

On the other hand, the literature has provided numerous 3D shape models based on volume, surface, global or local properties (e.g., medial axis [6], skeleton-curve [12], Reeb graphs [21]). Although most of descriptors can capture intrinsic shape property, they are not suited for dynamic representation as their structure is usually too noisy. Similarly, skeleton representation can capture intrinsic information of shape based on surface or volume (e.g., using a priori human model [11], thinning [36], rigging [5], etc.) but are either not invariant in time or often need prior knowledge on the shape to be described (e.g., a human skeleton). Also, once the intrinsic structure is found (i.e., shape topology), local surface details are usually excluded from the representation [51]. Hence, we propose here a new surface-based shape descriptor that can be used to define cut graphs on 3D model surfaces and encode 3D mesh sequences using a transformation into 2D video by cutting and parameterizing 3D surface meshes on image planes (e.g, using [40]) as in geometry video [7] and skin-off [20] methods. However, unlike existing methods, our model has invariant property to surface deformations and is thus well suited to sequences of inconsistent geometric data such as 3D video data. To our knowledge, no similar model has been designed in the literature (e.g., see surveys on invariant descriptors [16, 34], invariant skeleton [33], etc.).

3 Deformation invariant shape representation

3.1 Local feature point tracking

3.1.1 Critical point extraction

Let us assume that dynamic surfaces representing real-world objects in motion can be approximated by compact 2-manifold meshes. We consider geodesic distances to characterize surface intrinsic properties, as geodesic distances are invariant to isometric shape transformations when normalized (and can also be used to measure distortion between shapes [8]). Let $\mu : \mathcal{S} \rightarrow \mathbb{R}$ denote the continuous function defined on the object surface $\mathcal{S}$:

$$\mu(v) = \int_{\mathcal{S}} g(v, s) dS, \tag{1}$$

where $g : \mathcal{S}^2 \rightarrow \mathbb{R}$ is the geodesic distance between two points on $\mathcal{S}$. Eq. 1 is the geodesic integral function whose critical points can be used to characterize shape (see Morse theory [32]). The function μ is normalized with respect to its minimal and maximal values $\mu_{\min}$ and $\mu_{\max}$ as $\mu_N : S \rightarrow [0, 1]$, where $\mu_N(v) = \frac{\mu(v)-\mu_{min}}{\mu_{\max}-\mu_{\min}}$. Maximal values of μ_N usually correspond to limb extremities (e.g., of human or animal models) while global minimum corresponds to body center.

3.1.2 Reeb graph construction

As illustrated in Fig. 3a, we can use μ_N to build Reeb graphs in order to identify and match critical points over time using geometry and topology information. Reeb graphs are high level shape descriptors that can be used for shape matching and retrieval in large datasets (e.g., see [38] for Reeb graph theory, [37] for efficient Reeb graph construction, [21, 50] for shape retrieval in datasets of 3D models and [22] for 3D videos). Let us assume that 3D surface models are defined as compact 2-manifold surfaces approximated by 3D meshes, and let S denote a surface mesh. According to the Morse theory, a continuous function $\mu : S \rightarrow \mathbb{R}$ defined on S characterizes the topology of the surface on its critical points. The surface connectivity between critical points can then be modeled by the Reeb graph of μ, which is the quotient space defined by the equivalence relation $\sim$. Assuming the points $x \in S$ and $y \in S$, then $x \sim y$ if and only if:

$$\begin{cases} \mathbf{y} \in \text{same connected component of } \mu^{-1}(\mu(\mathbf{x})), \\ \mu(\mathbf{x}) = \mu(\mathbf{y}), \end{cases} \tag{2}$$

where the Morse function μ is defined as above. Hence, the Reeb graph of μ on S describes the connectivity of

the level sets of μ. Note that the inverse function μ^{-1} is defined on $\mathbb{R}$ and returns regions on S corresponding to level sets of μ at some isovalues.

3.1.3 Critical point matching

As defined above, the Reeb graph structure relies on surface critical points and captures surface topology. Consequently, graph nodes can be used to embed various local, global, geometry or topology information. Hence, critical point matching (i.e., Reeb graph leaf nodes) can be achieved by defining a similarity function accounting for embedded information in Reeb graph nodes. In practice, each node embeds the range of μ_N values it belongs to, the corresponding relative surface area, and topology information (i.e., graph node valence) as in [50, 49]. This method is efficient for discriminating and matching nodes in real-world object datasets such as the ones used in this paper, as natural shapes and poses are usually asymmetric (i.e., nodes contain different weights). However, in the particular case of synthesized data, when node matching can become ambiguous because of symmetry (e.g., if two legs are exactly similar), it may be necessary to leverage embedded information with geometrical or prior information (e.g., using node coordinates). Methods involving time-varying Reeb graphs [13] or scale-space [27] can be used for complex scenario of critical point tracking.

3.2 Temporal geodesic consistency

Definition 1. Assuming a set of N points $\mathcal{B} = \{b_1, ..., b_N\}$ defined on a 2-manifold $\mathcal{S}$, the points v_1 and v_2 on $\mathcal{S}$ are said geodesically consistent with respect to $\mathcal{B}$ if and only if:

$$\forall i \in [1, N], \quad |g(v_1, b_i) - g(v_2, b_i)| \leq \epsilon, \tag{3}$$

where $\epsilon \to 0$. If the points in $\mathcal{B}$ do not have any particular configuration of alignment or symmetry, the geodesic consistency property can be used to uniquely locate points on $\mathcal{S}$ when $N > 2$. In practice, the uniqueness is verified by checking the number of intersections of isovalue lines from $\mathcal{B}$, and ambiguities are solved by increasing N or adding geometric constraints (e.g., Euclidean distance). The framework is similar to a generalized barycentric coordinate system defined in a Riemannian manifold equipped with geodesic distance.

Definition 2. Assuming a set of N points $\mathcal{B}^t = \{b_1^t, ..., b_N^t\}$ defined on a deformable 2-manifold $\mathcal{S}^t$ at time $t \in [t_b, t_e]$, the points v_1^t and v_2^t on $\mathcal{S}^t$ are said temporally geodesically consistent with respect to $\mathcal{B}^t$ in $[t_b, t_e]$ if and only if:

$$\forall t \in [t_b, t_e], \forall i \in [1, N], \quad |g(v_1^t, b_i^t) - g(v_2^{t+\delta}, b_i^{t+\delta})| \leq \epsilon, \tag{4}$$

where $t_b < t_e$, $t + \delta \in [t_b, t_e]$ and $\epsilon \to 0$. g is normalized using the maximum geodesic distance over all pairs of points on $\mathcal{S}^t$ to preserve geodesic consistency when surfaces undergo non-rigid deformations (e.g., scale changes). Figure 3 illustrates temporal geodesic consistency with respect to critical points (top: 8, bottom: 5) extracted automatically using local geometry and topology properties (see [50]). Ambiguity maps are obtained by counting the number of candidate pairs $(v_1^t, v_2^{t+\delta})$ when $\epsilon > 0$. We observe that the regions located around object centers have very low ambiguity (i.e., numerical approximation is not an issue). In practice, we can search for $v_2^{t+\delta} = \arg\min_{v \in \mathcal{S}^t} \sum_{i=1}^{N} |g(v_1^t, b_i^t) - g(v, b_i^{t+\delta})|$.

3.3 Invariant surface-based graph construction

Definition 3. Let $\mathcal{C}^t = \{c_1^t, ..., c_N^t\}$ denote a set of invariant local features (e.g., local extrema) on $\mathcal{S}^t$ that are tracked over time in $[t_b, t_e]$. The surface-based shape descriptor $\mathcal{T}(\mathcal{V}^t, \mathcal{P}^t)$ is a graph on $\mathcal{S}^t$ whose nodes $\mathcal{V}^t$ are temporally geodesically consistent with respect to $\mathcal{C}^t$ in $[t_b, t_e]$. Every edge in $\mathcal{P}^t$ of $\mathcal{T}$ is linked to a feature in $\mathcal{C}^t$, and nodes of $\mathcal{T}$ represent edge junctions. Here, an edge consists of a path[1] on $\mathcal{S}^t$. To maintain the graph structure invariant over time independently from the parameterization of $\mathcal{S}^t$, we develop a probabilistic framework where edge positions are optimized using two geodesic consistency cues (see Fig. 4), while being located in regions of low ambiguity (see Def. 2).

Construction. First, we define an initial graph structure ρ_0 on $\mathcal{S}^{t_b}$ at t_b as either the global minimum (i.e., one point) given by Eq. 1 if $\mathcal{S}^{t_b}$ is genus-0, or as a graph cutting handles if the genus is higher (see [46], [14] and Sect. 4). Second, we initialize the graph: $\rho \leftarrow \rho_0$. The graph $\mathcal{T}$ is then built by iteratively adding the shortest edge linking a local feature (e.g., local maxima) in $\mathcal{C}^{t_b}$ to the current graph structure ρ until all elements in $\mathcal{C}^{t_b}$ are linked. At each step, the path ρ_j given by the pair of points $(c_j^{t_b}, v_j^{t_b}) \in \mathcal{C}^{t_b} \times \rho$ verifies:

$$(c_j^{t_b}, v_j^{t_b}) = \underset{(c,v) \in \mathcal{C}^{t_b} \times \rho}{\arg\min} \; g(c, v). \tag{5}$$

ρ_j is linked to ρ: $\rho \leftarrow \rho \cup \rho_j$, and $v_j^{t_b}$ is inserted into the set $\mathcal{V}^{t_b}$ (initially empty). When every feature in $\mathcal{C}^{t_b}$

[1] A path on a surface is a set of points linked two-by-two by a line.

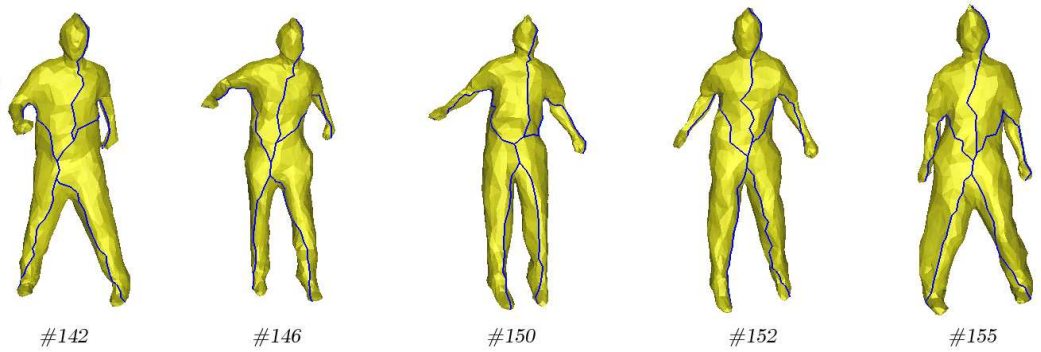


Fig. 4 Deformation invariant descriptor. The descriptor is a graph (in blue) defined on the object surfaces. Graph nodes are maintained geodesically consistent over time, while edges vary adaptively to surface deformations. (Bouncing sequence.)

is linked to ρ, we obtain $\rho = (\bigcup \rho_j) \cup \rho_0$ and we finally set: $\mathcal{T} \leftarrow \rho$ at t_b.

For all $t > t_b$, the invariant model is obtained by building a graph whose nodes have temporal geodesic consistency with the prior graph nodes and are located at local maxima or in non-ambiguous regions.. The problem is formulated as an MRF to find the optimal paths linking the graph nodes using intrinsic surface properties, so that graph constructions across time are independent from surface parameterization. The algorithm to construct a graph at t is the following:

1. Extract local features $\mathcal{C}^t = \{c_1^t, ..., c_N^t\}$ on $\mathcal{S}^t$ using Eq. 1 and match them to prior ones in $\mathcal{C}^{t-1}$ (e.g., using geometry and topology information [50]).
2. Derive an initial structure ρ_0^t on $\mathcal{S}^t$ geodesically consistent to the prior one. Note that for genus-0 surface, ρ_0^t is usually a point located around the object center. Set the graph at t: $\rho^t \leftarrow \rho_0^t$.
3. Edges that link the features $\mathcal{C}^t$ to the current graph structure ρ^t are added iteratively, and in the same order as the prior steps. Let $\mathcal{P}^t = \{p_i^t\}$ denote the set of points forming a path (a graph edge) linking a feature c^t to a node v^t at t, and $\mathcal{D}^t = \{d_i^t\}$ denote the set of points forming the shortest path linking c^t to v^t (e.g., using Dijkstra's algorithm). To obtain the optimal path $\mathcal{P}^t = \{p_i^t\}$, the problem is expressed as a MAP-MRF where the surface mesh vertices at t serve as sites. Probabilities of p_i^t to be at some positions at t are computed given known priors $\mathcal{P}^{t-1}$ and $\mathcal{D}^t$. The posterior probability to maximize is:

$$\Pr(\mathcal{P}^t|\mathcal{D}^t, \mathcal{P}^{t-1}) \propto \prod_i E_d(p_i^t, d_i^t) E_p(p_i^t, p_i^{t-1}) \prod_i \prod_{j \in \mathcal{N}(i)} V(p_i^t, p_j^t), \quad (6)$$

where E_d and E_p are the local evidence terms for a point p_i^t to be at positions inferred from d_i^t and p_i^{t-1} respectively, $\mathcal{N}(i)$ is the neighborhood of i, and V is a pair-wise smoothness assumption (so that $\mathcal{P}^t$ forms a path on $\mathcal{S}^t$). E_d and E_p are defined as what follows:

$$E_d(p_i^t, d_i^t) = f_d(\sum_{k \in [1,N]} ||g(p_i^t, c_k^t) - g(d_i^t, c_k^t)||), \quad (7)$$

$$E_p(p_i^t, p_i^{t-1}) = f_p(\sum_{k \in [1,N]} ||g(p_i^t, c_k^t) - g(p_i^{t-1}, c_k^{t-1})||), \quad (8)$$

where f_d and f_p are Gaussian distributions centered on d_i^t and p_i^{t-1} respectively, g is the normalized geodesic distance, $c_k^t \in \mathcal{C}^t$ and $c_k^{t-1} \in \mathcal{C}^{t-1}$. Note that indices were simplified for clarity: $\mathcal{P}^{t-1}$, $\mathcal{P}^t$ and $\mathcal{D}^t$ may not have the same number of elements, and d_i^t and p_i^{t-1} are the closest point to p_i^t on $\mathcal{D}^t$ and $\mathcal{P}^{t-1}$. Hence, Eq. 6 estimates the probability of $\mathcal{P}^t$ to be geodesically consistent to the previous edge $\mathcal{P}^{t-1}$, while being influenced by the shortest path $\mathcal{D}^t$. Let $\mathcal{P}^\star$ denote the optimal path linking the feature c^t to the node v^t. Thus, we have to estimate:

$$\mathcal{P}^\star = \arg\max_{\{\mathcal{P}^t\}} \Pr(\mathcal{P}^t|\mathcal{D}^t, \mathcal{P}^{t-1}), \quad (9)$$

where $\{\mathcal{P}^t\}$ denotes all the possible paths linking c^t to v^t. Shortest paths are added one-by-one to avoid edge overlapping when linking local features. E_d acts as a force that attracts the path to a state where the stress is lower (see Fig. 4) when an elastic deformation occurs or in case of surface noise (e.g., 3D reconstruction artifact). As well, E_d prevents the model to be subject to error accumulation over time, causing drift effects. On the other hand, E_p maintains the graph structure consistent over time, which can be crucial for some applications (see Sect. 4).

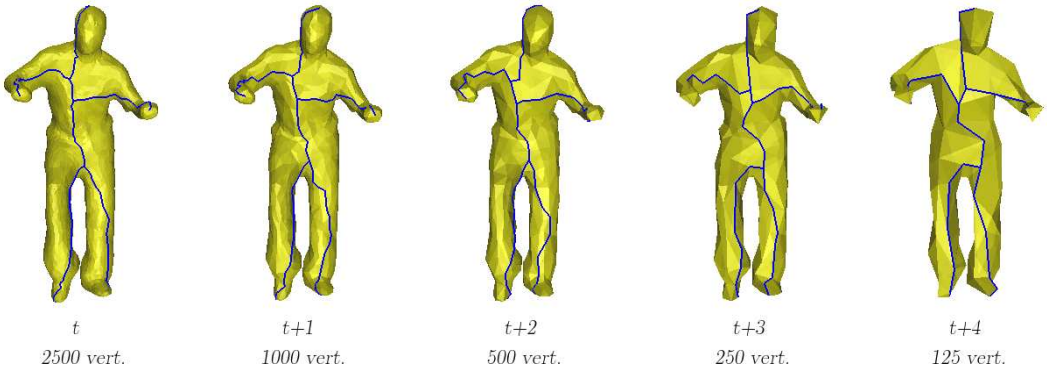


Fig. 5 Invariant property against surface parameterization. The graph structure is maintained invariant even though the surface mesh complexity and connectivity change. Here, the number of vertices varies from 2500 to 125 vertices. (Lock sequence.)

ρ^t is obtained by iteratively adding paths $\mathcal{P}_j^\star$ linking $c_j^t \in \mathcal{C}^t$ to the current graph at node v_j^t. v_j^t is the closest point on the current graph to:

$$\bar{v}_j^t = \arg\min_{v \in \rho_t} [\lambda . g(\hat{v}_j^t, v) + (1-\lambda) . g(\check{v}_j^t, v)], \qquad (10)$$

where $\hat{v}_j^t$ is the point in $\mathcal{S}^t$ geodesically consistent to v_j^{t-1} in $\mathcal{S}^{t-1}$ with respect to $\mathcal{C}^t$, $\check{v}_j^t$ is the intersection point given by the shortest path from c_j^t to ρ^t, and $\lambda = 0.5$ is a weight. (Dependence to temporal priors are canceled if $\lambda = 0$.) In addition, v_j^t is constrained to belong to the edge derived from the edge containing v_j^{t-1}. The structure of $\mathcal{T}$ is therefore maintained invariant over time. Note that priors can be extended to $\{\mathcal{P}^{t-k}\}_{t_b<k<t}$.

4. Repeat Step 3. until every feature in $\mathcal{C}^t$ is linked to ρ^t. Finally, the graph $\mathcal{T}$ at t is given by $\rho^t \leftarrow (\bigcup \mathcal{P}_j^{\star,t}) \cup \rho_0^t$.
5. Set $t \leftarrow t+1$ and repeat Step 1. to 4. for all $t < t_e$.

Note that the optimization problem in Step 3. can be effectively solved by dynamic programming. Finally, as illustrated in Fig. 5 we obtain a graph that is invariant over time regardless of the surface parameterization (i.e., mesh complexity and connectivity).

4 3D video data encoding

We propose to apply the descriptor to 3D video data for encoding purpose using a strategy inspired by geometry images [18]. The overall geometry image transformation scheme is illustrated in Fig. 6. The strategy consists in transforming 3D video data stream into 2D video. Any mature 2D encoding algorithm (such as Windows Media, MPEG-4, Quicktime, etc.) can then be used for (lossless) compression. Particularly, the surface-based shape descriptor $\mathcal{T}$ introduced in the previous section provides an invariant structure to 3D data obtained independently, such as a sequence of 3D meshes obtained from multiple view stereo (e.g., from Kyoto University [29] or University of Surrey [44]). The invariant description can then be exploited to obtain optimal encoding, as successive geometric data representations present small variations.

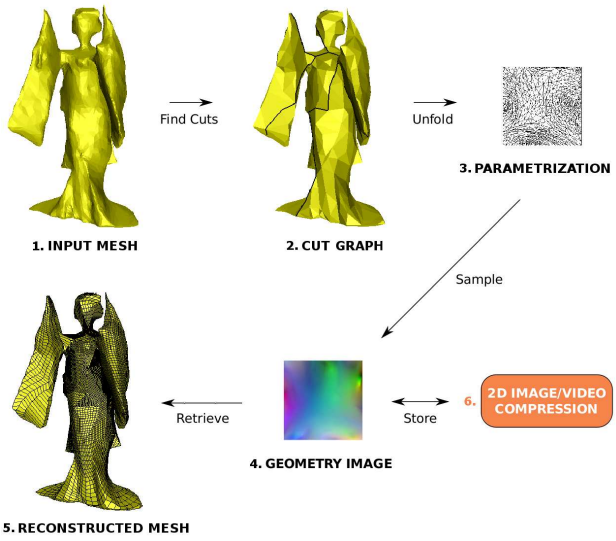


Fig. 6 Invariant property against surface parameterization.

For each frame, the graph structure of $\mathcal{T}$ is used as a cut graph ρ that cuts and opens the 3D surface mesh M into a disk (a genus-0 chart). M is then mapped onto a flat parameter domain, which will be used as an image plane. Finally, M is resampled on a regular grid, where the 3D coordinates XYZ are scaled and stored as RGB pixel components to form a 2D (geometry) image $\mathcal{I}$. To retrieve M from $\mathcal{I}$, RGB values are simply reconverted to 3D coordinates. When applying the transformation on a sequence of 3D meshes, the process returns

a sequence of images (i.e., a video). As the graph structure is invariant over the sequence, consecutive frames vary smoothly and can therefore be efficiently encoded using any popular codec for 2D video. Note that if a lossy compression method is used for encoding and alters the border of $\mathcal{I}$, cracks may be observed on the reconstructed surface around the cut ρ. In that case, a post-processing step (e.g., mesh joining or hole filling) may be necessary to preserve the topology of the initial mesh. The advantage of the proposed invariant surface-based shape descriptor for 3D video encoding is at least twofold:

1. The shape descriptor can be used as cut graphs to produce smoothly varying geometry images from real-world 3D video data independently from the surface parameterization, i.e., even though the mesh resolution or connectivity is inconsistent between consecutive frames. Hence the model allows for adaptive bitrate streaming application, whereas state-of-the-art methods cannot be applied (e.g., geometry video [7], independent planar parametrization [40]).
2. In standard parameterization approaches (see [40] for state-of-the-art implementation of [15, 18]), the computation of the cut graph ρ is obtained iteratively and requires several parameterization steps to detect all the local extrema one-by-one (e.g., using triangle geometric stretch). On the other hand, the proposed strategy is one-shot, and still guarantees that the generated cut path passes through all local extrema of M (i.e., surface protrusions), which is a crucial condition to preserve the geometry accuracy after transformation. When the cut graphs are well defined, the transformation can be used for lossless compression of 3D meshes.

Sensitivity to topological perturbations. As the cut graph passes through all extrema, critical points usually lie at the boundaries of the parameter domain. When surface topology changes, the number of critical points may vary, and the graph structure can locally change. This results in a discontinuity between consecutive geometry images that cannot be avoided. On the other hand, it guarantees that the original surface topology is preserved and can be reconstructed from a single chart. Otherwise a surface alignment method should be applied as preprocessing (see Sect. 2), but large resolution variations as shown in Fig. 5 would not be handled and original topology would be lost. Methods that estimate global geodesic distortions for shape matching are usually robust to local surface deformation (e.g., using isometry invariant framework [31], generalized multidimensional scaling [8], etc.). However, the global measures can be strongly affected by surface topology changes, as opposed to the proposed descriptor which is only locally affected. Figure 7 shows geometry image discontinuities when altering geodesic consistency of nodes and adding an arbitrary critical point. As critical points are matched across time, image regions with no perturbation remain aligned (see left part of images).

Surface-based shape descriptor versus topology-based shape descriptor. In [51]. the authors introduced a method for 3D video encoding using a topology-based shape descriptor, namely the augmented Multiresolution Reeb graph (aMRG) [50], which is similar to the graph depicted in Sect. 3. The approach consists of: (1) automatically extracting intrinsic information of surface shape by computing an aMRG for each 3D video frame, (2) tracking and recording each aMRG node relative displacement across time, and (3) compactly encode 3D video sequences using the obtained representation. To reconstruct a 3D video sequence, a reference mesh (e.g., frame #0) is deformed by deriving the positions of all aMRG nodes across time. The dynamic surface is recovered by skinning (e.g., using dual quaternions [26]), where the aMRG serves as skeleton and aMRG edges serve as bones. However, like deformation mesh transfer methods (e.g., [5]), local surface details are not encoded and therefore cannot be recovered. Although these methods are well adapted to synthesized 3D mesh sequences or low resolution 3D video sequences, they are not sufficient to handle high resolution 3D video sequences reconstructed from real-world objects that contain local details and subtle variations (e.g., cloth wrinkles). In Sect. 5, we show the performance of the proposed surface-based shape descriptor for 3D video encoding and lossless compression (i.e., reconstruction).

5 Experimental results

Datasets. For experimental validations, we have tested the algorithm on publicly available datasets of 3D video reconstructed from multi-view images (from the University of Surrey [44], and the MIT [52]). The 3D video datasets consist of real subjects wearing loose clothing and performing various actions, such as dancing or jumping. Surfaces can therefore vary a lot between two consecutive frames when the motion is fast. Results on a synthesized dataset representing a galloping elephant (from [45]) are also given. The model presents several protusions which numbers are different from human models (see Fig. 8). Our experiments aim to assess the invariant property of the proposed descriptor regardless of surface parameterization, and its performance (e.g.,

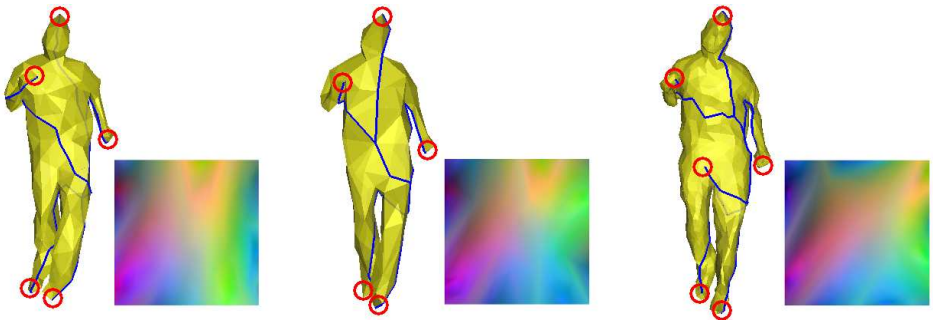


Fig. 7 (Left and Center) Geometry images show discontinuities when graph nodes are not geodesically consistent. (Right) Adding an arbitrary critical point alters locally the image boundaries. (Bouncing sequence.)

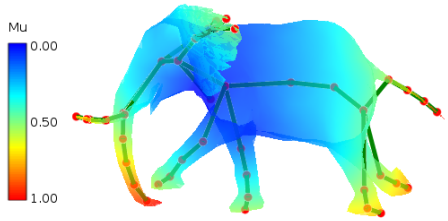


Fig. 8 Extremal point extraction on a mesh model from the sequence Elephant [45]. Reeb graph [50] computation returns 10 extrema (i.e., limbs, tail, ears, tusks and trunk).

reconstruction accuracy) when applied for 3D video adaptive bitrate streaming. For quantitative evaluation, we use 3D mesh sequences processed by [9] as the surface genus is theoretically consistent over the sequences. (In practice, 3D video data can be post-processed using any surface alignment method to prevent surface topology changes as described in Sect. 2.) In addition, we remeshed the sequences to cancel all mesh connectivity consistency, and produced mixed resolution 3D video data containing alternatively 3D meshes of 1000, 500, 250 and 125 vertices.

We perform comparisons to a state-of-the-art parameterization technique [40], where cut graphs are obtained by iterative parameterizations. The approach, here named Geometry Image Sequence (GIS), is known to optimally encode closed 3D surface meshes. Results obtained with our proposed technique on mesh sequences having same resolution are denoted 'fixed', whereas results obtained on sequences with meshes having various resolutions are denoted 'mixed'.

Computation time. All computations were performed on a dual-core PC (Intel Core2 Duo CPU @3.00 GHz, 4GB RAM). The proposed approach requires: (1) extremal point extraction and matching, (2) surface-based descriptor construction, and (3) one-shot parametrization. Step (1) depends on μ function computation (see Eq. 1). Running time is longer when using the geodesic integral function, while using the height function as in [37] is faster by several orders of magnitude (e.g., 1min to 1min30s for a 4000 vertex mesh against few seconds). The latter is used when it is safe to assume that all meshes in the sequence are oriented. The geodesic integral function is implemented using Dijkstra shortest path algorithm with binary heap, whose complexity is $O((E+V)logV)$, where E is the number of edges and V is the number of vertices in a mesh. The overall running time using Eq. 1 is about 2min if $V = 4000$, 50s if $V = 2000$, 15s if $V = 1000$, and below 6s if $V < 500$.

Invariant property evaluation. To assess the invariant property of the descriptor to surface variations and its ability to produce consistent geometry images that varies smoothly, the mean square error of pixel values (MSE) between consecutive geometry images is computed (smaller MSE is better). It allows us to estimate how much the geometry images vary over a sequence. In our experiments, the size of geometry images has been fixed to 128×128 pixels (encoded in RGB with 8bit per pixel component) for the sake of consistent comparison. (To achieve optimal streaming, the geometry images should indeed be resized with respect to the mesh resolution.) Table. 1 shows average MSE obtained on various sequences. The proposed descriptor shows remarkable invariant property between consecutive frames: average MSE(fixed) values are very low. Moreover, the resolution changes do not affect the performance: MSE(mixed) are low as well. Note that as GIS does not contain any stabilization mechanism: average MSE(GIS) values are high and are given for comparison.

Fig. 9 shows MSE graphs for several sequences. Figure 10 illustrates invariant graphs obtained with our approach with fixed and mixed mesh resolution.

Fig. 9 Mean Square Errors (MSE) of pixels between consecutive geometry images. Results obtained with [40] are denoted 'gis', results obtained with our method on sequences with fixed resolution are denoted 'fixed', and results obtained with our method on sequences with mixed resolutions are denoted 'mixed'.

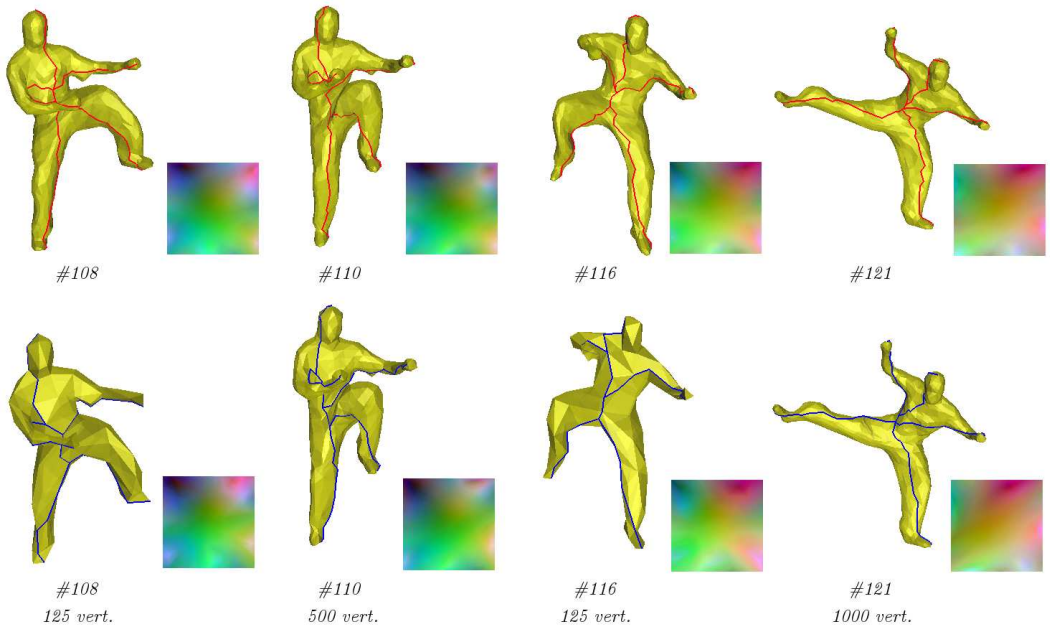


Fig. 10 Graph invariant property regardless of surface mesh complexity and connectivity. (top) shows a mesh sequence with 1000 vertices. (bottom) shows the same sequence with meshes at different resolutions. Although surface parameterizations are different, the proposed surface-based shape descriptor computed on the Lock sequence shows invariant property and adaptivity: graphs and geometry images remain similar.

Table 1 Average MSE of pixel values between consecutive geometry images.

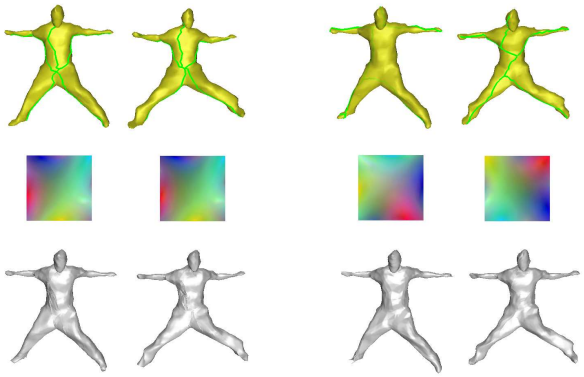
	MSE(GIS)	MSE(fixed)	MSE(mixed)
Bouncing	35302	2886	3224
Crane	28485	1670	2025
Handstand	30671	1261	3125
Kickup	27700	1938	3753
Lock	22466	1700	3037
Samba	35302	2886	3282
Elephant	23232	4931	8151

Table 2 Average Hausdorff distances Δ to ground truth.

	Δ(GIS)	Δ(ours)	Δ(ref)
Bouncing	0.0122	0.0126	0.0885
Crane	0.0126	0.0132	0.0729
Handstand	0.0119	0.0122	0.0920
Kickup	0.0118	0.0120	0.0862
Lock	0.0079	0.088	0.0783
Samba	0.0223	0.0237	0.0913
Elephant	0.0315	0.0466	0.0985

Reconstruction accuracy. To assess the reconstruction accuracy of geometry images obtained from the invariant surface-based shape descriptor used as cut graphs, Hausdorff distances are computed between original meshes and reconstructed meshes [35]. Average Hausdorff distances Δ between ground truth sequences and reconstructed surfaces by GIS and our proposed method (with fixed resolution) are reported in Table. 2. We can observe similar performances between the proposed approach and GIS as Δ is very low for both methods. Results between original data and simplified mesh with arbitrary resolution (e.g., 125 vertices) are given for comparison (see Δ(ref)).

Figure 11 illustrates cut graphs obtained with the proposed method against [40]. Note the strong geometry image variations obtained with unstabilized cut graphs. Additional examples given in Fig. 12 show the



a) proposed approach b) geometry image sequence (GIS)

Fig. 11 Encoding and reconstruction of sequence Bouncing. a) Our approach produces stable cut graphs that are used to produce smoothly varying geometry images. b) Results with unstabilized cut graphs [40]).

descriptor invariant property to surface undergoing large deformations. Furthermore, as shown in Fig. 13, our method can achieve accurate reconstruction (comparable to GIS which is optimal according to [18, 40]) while using an original one-shot processing, as opposed to standard iterative parameterizations employed by GIS.

Encoding performance. Table. 3 shows 3D video encoding performance with respect to different strategies (zipped OFF, GIS, ours). For each format, the size of each sequence is given in KB. Standard H.264/MPEG-4 is used for compression of geometry images (128×128p). As can be observed, our method clearly performs better. Compressed geometry image sequences obtained with stable cut graphs are 100 times smaller than zipped sequences of objects in OFF format, and around 40% smaller than sequences obtained with unstable cut graphs.

Table 3 3D video encoding. For each format, the size of each sequence is given in KB. Standard H.264/MPEG-4 is used for compression of geometry images (128×128p).

	#fr.	OFF(zip)	GIS	ours
Bouncing	174	16,300	304.4	169.9
Crane	173	14,100	283.7	162.7
Handstand	173	24,700	283.4	154.7
Kickup	219	29,900	365.1	197.0
Lock	249	32,400	388.2	204.0
Samba	174	22,200	304.0	173.2
Elephant	48	2,197	84.8	78.7

6 Conclusion

We present a novel invariant shape descriptor to represent spatio-temporal visual information that varies over time, such as 3D dynamic surfaces. The proposed descriptor consists in a surface-based graph that lies on object surfaces, and is anchored to local features. The overall graph structure is made invariant to surface variations using surface intrinsic geometric properties while surfaces undergo non-rigid deformation. In particular, the graph is defined within a probabilistic framework using temporal geodesic consistency cues as priors, and is independent to surface parameterization. Hence, the descriptor can be used to bring an invariant structure to 3D geometric data that are produced independently, such as 3D video obtained from multiple view stereo.

We show that the proposed shape descriptor can be employed as surface cut graphs, which enables 3D surface models to be transformed into 2D (geometry) images using a one-shot strategy while geometry is accurately preserved. Moreover, the invariant property of the representation allows the production of smoothly varying images, regardless of the 3D surface mesh complexity and connectivity. Therefore, the approach is suitable for adaptive bitrate streaming of 3D video data, which was a challenging issue as state-of-the-art techniques are only designed to optimally encode 3D animated mesh sequences sharing a same mesh connectivity.

For further research, it would be interesting to tackle large scale 3D video encoding, e.g., where dynamic surfaces represent outdoor scenes composed of multiple objects. In this case, each individual foreground or background object could be described independently by an invariant surface-based descriptor. A reasonable strategy would consist of transforming a scene into multiple (planar) geometry images, and eventually combine the geometry images into one or several atlases. Similarly, large objects or objects composed of several pieces (i.e., non-manifold surfaces) could be efficiently represented using multi-chart geometry images (see [41, 10]). To ensure the invariance of each surface-based descriptor in each surface region, graph nodes would have to be placed on the boundaries of the regions (i.e., charts) and maintained geodesically consistent across the sequence using the approach proposed in Sect. 3.3. Furthermore, additional surface features such as color (when available) may be exploited. Also, it would be interesting to extend the model to texture mapping such as in [19] where geodesic paths are also exploited.

Acknowledgements This work was supported in part by the JST-CREST project "Creation of Human-Harmonized Information Technology for Convivial Society". The authors thank Dr. Lyndon Hill for his preliminary work on this project.

References

1. de Aguiar E, Stoll C, Theobalt C, Ahmed N, Seidel HP, Thrun S (2008) Performance capture from sparse multi-view video. ACM Trans Graphics 27(3)
2. Alexa M, Müllen W (2000) Representing animations by principal components. Computer Graphics Forum 19(3)
3. Allard J, Ménier C, Raffin B, Boyer E, Faure F (2007) Grimage: Markerless 3d interactions. ACM SIGGRAPH - Emerging Technologies
4. Alliez P, Gotsman C (2005) Recent advances in compression of 3d meshes. Advances in Multiresolution for Geometric Modelling Springer-Verlag

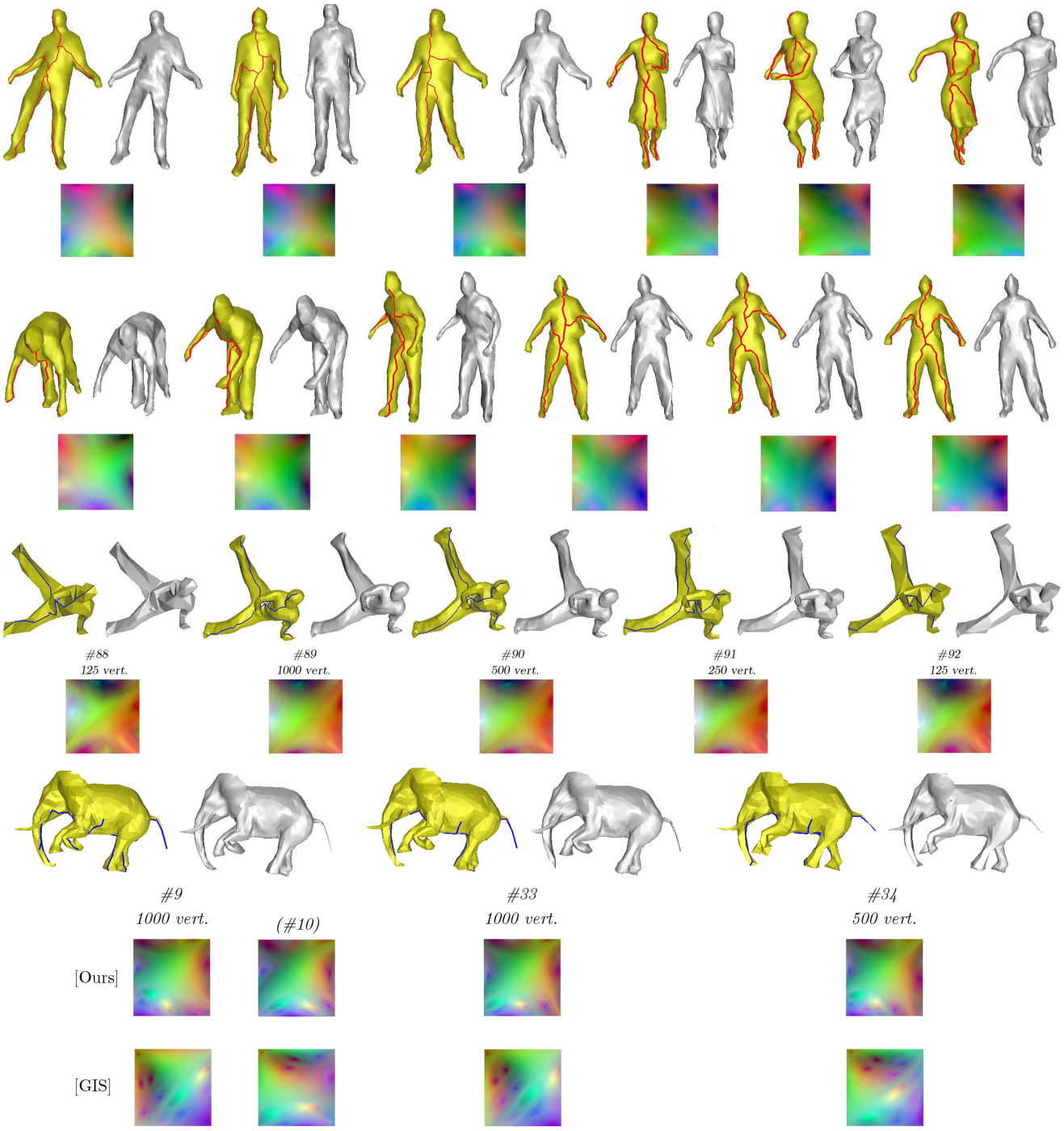


Fig. 12 Encoding and reconstruction. Surface-based graphs are shown in red and blue (for mixed resolutions), reconstructions are shown in gray. Although surfaces undergo strong variations, the invariant surface-based shape descriptor produces smoothly varying geometry images and accurate surface reconstruction. Sequences are: Crane and Samba (1^{st} row), Handstand (2^{nd} row), Kickup (3^{rd} row), and Elephant (4^{th} row).

5. Baran I, Popovic J (2007) Automatic rigging and animation of 3d characters. ACM Trans Graphics 26(3):27
6. Blum H (1967) A transformation for extracting new descriptors of shape. Models for the perception of speech and visual form, MIT Press pp 362–380
7. Briceno H, Sandler P, McMillian L, Gortler S, Hoppe H (2003) Geometry videos: A new representation for 3d animations. Eurographics/SIGGRAPH Symp Computer Animation pp

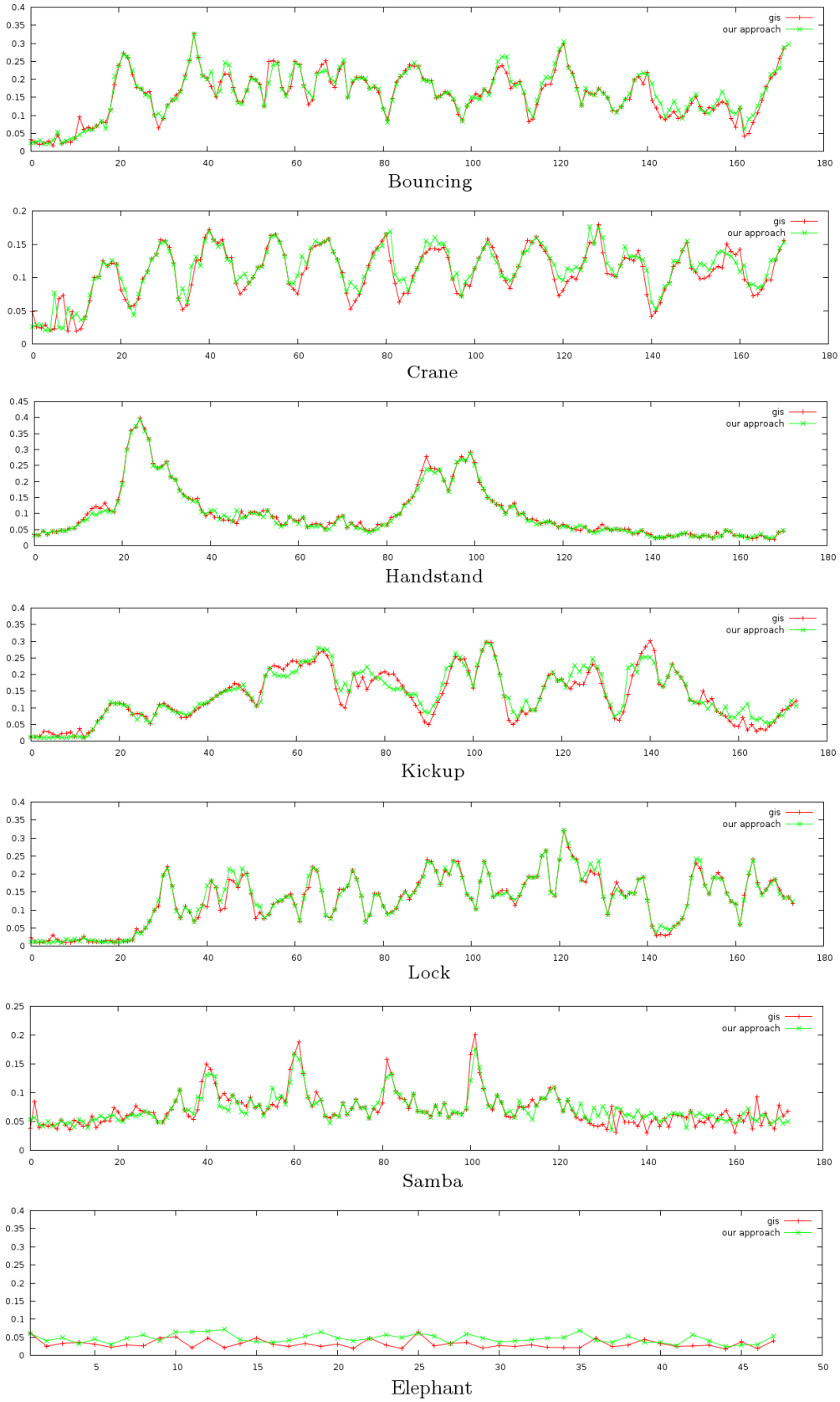


Fig. 13 Hausdorff distances Δ for various sequences.. Our approach allows accurate reconstruction of mesh sequences comparable to state-of-the-art implementation [40].

136–146
8. Bronstein AM, Bronstein MM, Kimmel R (2007) Calculus of non-rigid surfaces for geometry and texture manipulation. IEEE Trans Visualization and Computer Graphics pp 902–913
9. Cagniart C, Boyer E, Ilic S (2010) Probabilistic deformable surface tracking from multiple videos. Proc European Conf Computer Vision
10. Carr N, Hoberock J, Crane K, Hart J (2006) Rectangular multi-chart geometry images. Proc Eurographics Symp Geometry Processing pp 181–190
11. Carranza J, Theobalt C, Magnor M, Seidel HP (2003) Free-viewpoint video of human actors. ACM Trans Graphics 22(3):569–577
12. Cornea N, Silver D, Yuan X, Balasubramanian R (2005) Computing hierarchical curve skeletons of 3d objects. The Visual Computer Journal 21(11):945–955
13. Edelsbrunner H, Harer J, Mascarenhas A, Pascucci V (2004) Time-varying reeb graphs for continuous space-time data. Proc Symp Computational Geometry
14. Erickson J, Har-Peled S (2004) Optimally cutting a surface into a disk. Discrete & Computational Geometry 31(1):37–59
15. Floater M (1997) Parametrization and smooth approximation of surface triangulations. Computer Aided Geometric Design 14(3):231–250
16. Forsyth DA, Mundy JL, Zisserman A, Coelho C, Heller A, Rothwell C (1991) Invariant descriptors for 3d object recognition and pose. IEEE Trans Pattern Analysis Machine Intelligence 13(10)
17. Franco J, Menier C, Boyer E, Raffin B (2004) A distributed approach for real-time 3d modeling. Proc IEEE Conf Computer Vision Pattern Recognition Workshop on Real-Time 3D Sensors and their Applications
18. Gu X, Gortler S, Hoppe H (2002) Geometry images. ACM SIGGRAPH
19. Guo YW, Wang J, Cui XF, Peng QS (2005) A new constrained texture mapping method. Entertainment Computing-ICEC Springer Berlin Heidelberg, Springer LNCS
20. Habe H, Katsura Y, Matsuyama T (2004) Skin-off: Representation and compression scheme for 3d video. Proc Picture Coding Symposium
21. Hilaga M, Shinagawa Y, Kohmura T, Kunii TL (2001) Topology matching for fully automatic similarity estimation of 3d shapes. ACM SIGGRAPH pp 203–212
22. Huang P, Tung T, Nobuhara S, Hilton A, Matsuyama T (2010) Comparison of skeleton and non-skeleton shape descriptors for 3d video. Proc 3DPVT
23. Jiang H, Liu H, Tan P, Zhang G, Bao H (2012) 3d reconstruction of dynamic scenes with multiple handheld cameras. Proc European Conf Computer Vision
24. Kanade T, Yoshida A, Oda K, Kano H, Tanaka M (1996) A stereo machine for video-rate dense depth mapping and its new applications. Proc IEEE Conf Computer Vision Pattern Recognition
25. Karni Z, Gotsman C (2004) Compression of soft-body animation sequence. Computers & Graphics 28:25–34
26. Kavan L, Collins S, Žára J, O'Sullivan C (2007) Skinning with dual quaternions. Proc Symposium on Interactive 3D Graphics and Games pp 39–46
27. Klein T, Ertl T (2005) Scale-space tracking of critical points in 3d vector fields. Proc Topology-Based Methods in Visualization
28. Mamou K, Zaharia T, Preteux F, Stefanoski N, Ostermann J (2008) Frame-based compression of animated meshes in mpeg-4. Proc IEEE Int'l Conf Multimedia and Expo
29. Matsuyama T, Wu X, Takai T, Nobuhara S (2004) Real-time 3d shape reconstruction, dynamic 3d mesh deformation, and high fidelity visualization for 3d video. Computer Vision Image Understanding 96(3):393–434
30. Matsuyama T, Nobuhara S, Takai T, Tung T (2012) 3d video and its applications. Springer
31. Mémoli F, Sapiro G (2005) A theoretical and computational framework for isometry invariant recognition of point cloud data. Found Comput Math 5(3):313–347
32. Morse M (1934) The calculus of variations in the large. American Mathematical Society, Colloquium Publication 18, New York
33. Mortara M, Patanè G (2002) Affine-invariant skeleton of 3d shapes. Proc Shape Modeling International
34. Mundy J, Zisserman A (1992) Geometric invariance in computer vision. MIT Press
35. P Cignoni CR, Scopigno R (1998) Metro: measuring error on simplified surfaces. Computer Graphics Forum 17(2):167–174
36. Palagyi K, Kuba A (1999) A parallel 3d 12-subiteration thinning algorithm. Graph Models and Image Proc 61(4):199–221
37. Pascucci V, Scorzelli G, Bremer PT, Mascarenhas A (2007) Robust on-line computation of reeb graphs: Simplicity and speed. ACM Trans Graphics 26
38. Reeb G (1946) On the singular points of a completely integrable pfaff form or of a numerical

function. Comptes Rendus Acad Sciences Paris 222:847–849
39. Rothganger F, Lazebnik S, Schmid C, Ponce J (2006) 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. Int'l Journ Computer Vision 66(3):231–259
40. Saboret L, Alliez P, Lévy B (2012) Planar parameterization of triangulated surface meshes. In CGAL Reference Manual CGAL Editorial Board, 40 edition
41. Sander P, Wood Z, Gortler S, Snyder J, Hoppe H (2003) Multi-chart geometry images. Proc Eurographics Symp Geometry Processing pp 146–155
42. Seitz S, Curless B, Diebel J, Scharstein D, Szeliski R (2006) A comparison and evaluation of multi-view stereo reconstruction algorithms. Proc IEEE Conf Computer Vision Pattern Recognition
43. Starck J, Hilton A (2005) Spherical matching for temporal correspondence of non-rigid surfaces. Proc IEEE Int'l Conf Computer Vision
44. Starck J, Hilton A (2007) Surface capture for performance-based animation. IEEE Computer Graphics and Applications
45. Sumner RW, Popovic J (2004) Deformation transfer for triangle meshes. ACM Trans Graphics 23(3)
46. Taubin G, Rossignac J (1998) Geometric compression through topological surgery. ACM Trans Graphics 17(2):84–115
47. Tung T, Matsuyama T (2010) Dynamic surface matching by geodesic mapping for 3d animation transfer. Proc IEEE Conf Computer Vision Pattern Recognition
48. Tung T, Matsuyama T (2012) Invariant surface-based shape descriptor for dynamic surface encoding. Proc Asian Conf Computer Vision
49. Tung T, Matsuyama T (2012) Topology dictionary for 3d video understanding. IEEE Trans Pattern Analysis Machine Intelligence 34(8):1645–1657
50. Tung T, Schmitt F (2005) The augmented multiresolution reeb graph approach for content-based retrieval of 3d shapes (code on webpage). Int'l Journ Shape Modeling 11(1):91–120
51. Tung T, Schmitt F, Matsuyama T (2007) Topology matching for 3d video compression. Proc IEEE Conf Computer Vision Pattern Recognition
52. Vlasic D, Baran I, Matusik W, Popovic J (2008) Articulated mesh animation from multi-view silhouettes. ACM Trans Graphics 27(3)

Tony Tung received the M.Sc. degree in Physics and Computer Science from Télécom Physique, France, with a double degree in Photonics and Image Processing in 2000, and the Ph.D. degree in Signal and Image processing from Télécom ParisTech, France, in 2005. He worked as IT consultant (2000-2002) and senior R&D engineer (2005-2008) for companies, and as postdoctoral research fellow at Kyoto University (2005, 2008-2009). Since 2010, he has been an Assistant Professor at Kyoto University, working jointly at the Graduate School of Informatics, and at the Academic Center for Computing and Media Studies. His research interests include computer vision, pattern recognition, shape modeling, and human-machine interaction. He was awarded Fellowships from the Japan Society for the Promotion of Science in 2005 and 2008, and Grant-in-Aid for Young Scientists in 2011.

Takashi Matsuyama received B. Eng., M. Eng., and D. Eng. degrees in electrical engineering from Kyoto University, Japan, in 1974, 1976, and 1980, respectively. He is currently a professor in the Department of Intelligence Science and Technology, Graduate School of Informatics, Kyoto University. His research interests include knowledge-based image understanding, computer vision, 3D video, human-computer interaction, and smart energy management. He wrote more than 100 papers and books including two research monographs, A Structural Analysis of Complex Aerial Photographs, PLENUM, 1980 and SIGMA: A Knowledge-Based Aerial Image Understanding System, PLENUM, 1990. He won ten best paper awards from Japanese and international academic societies including the Marr Prize at ICCV'95. He is on the editorial board of the Pattern Recognition Journal. He was awarded Fellowships from the International Association for Pattern Recognition, the Information Processing Society of Japan, and the Institute of Electronics, Information, and Communication Engineers Japan.