Timing-Based Local Descriptor for Dynamic Surfaces

Tony Tung Takashi Matsuyama
Graduate School of Informatics, Kyoto University, Japan
tony2ng@gmail.com tm@i.kyoto-u.ac.jp

Abstract

In this paper, we present the first local descriptor designed for dynamic surfaces. A dynamic surface is a surface that can undergo non-rigid deformation (e.g., human body surface). Using state-of-the-art technology, details on dynamic surfaces such as cloth wrinkle or facial expression can be accurately reconstructed. Hence, various results (e.g., surface rigidity, or elasticity) could be derived by microscopic categorization of surface elements. We propose a timing-based descriptor to model local spatiotemporal variations of surface intrinsic properties. The low-level descriptor encodes gaps between local event dynamics of neighboring keypoints using timing structure of linear dynamical systems (LDS). We also introduce the bag-of-timings (BoT) paradigm for surface dynamics characterization. Experiments are performed on synthesized and real-world datasets. We show the proposed descriptor can be used for challenging dynamic surface classification and segmentation with respect to rigidity at surface keypoints.

1. Introduction

Non-rigid surfaces or soft tissues (such as human bodies, faces, organs, cloths, or fluids) are dynamic surfaces that can be represented by sequences of 3D models. The complexity and constant change in geometry and topology of these objects pose challenges for applying traditional vision algorithms to the sequences. Over the past few years, new algorithms have been proposed for vision tasks such as registration, segmentation and categorization of such data. Among these tasks, the *microscopic* categorization of surfaces is of specific interest to us, because it is critical to surveillance applications, such as detecting organ anomalous behavior, skin aging, leaks in tanks at power plant, or assessing fabric quality.

The reconstruction of 3D dynamic surfaces that represent real-world visual and spatial information can nowadays be achieved with high accuracy (i.e., below 0.5 cm) and in reasonable time thanks to recent progresses in sensing technologies. For example, 3D video (i.e., sequence of full 3D models) representing live human performances or daily activities (e.g., dance, or yoga) can be obtained using various techniques (see multiview stereo [28, 18], or depth data fusion [20]).

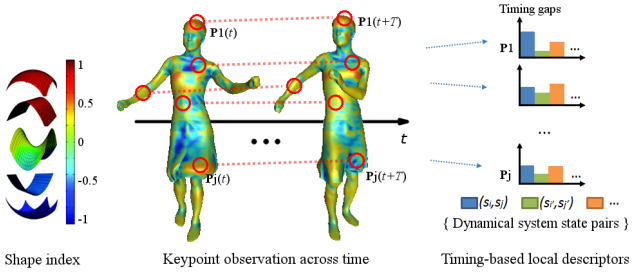

Figure 1. Timing-based local descriptors extracted from dynamic surface keypoints for local surface characterization (e.g., rigidity classification). Center) Dynamic surface with intrinsic information (shape index) observed across time. Right) Local descriptors obtained at keypoints from gaps between local event dynamics using timing structure of dynamical systems.

As we can assume that dynamic surfaces representing real-world objects show temporally continuous variations and have remarkable temporal statistics (e.g., clothing that wrinkles, or speaking face), they can therefore be characterized by modeling surface local deformation dynamics across time. In [43], the authors demonstrated that surface rigidity can be characterized by modeling intrinsic property dynamics using linear dynamical systems (LDS), which have long history in dynamics modeling (see dynamic textures [34, 11, 12, 4, 44, 33]). However their model only accounts for system state duration of independent observations, and is so far limited to binary classification. On the other hand, we propose to: 1) take into account spatial and timing distribution of observations by considering keypoint neighborhood, 2) form low-level local descriptors based on timing structure gaps between LDS states (see Fig. 1 for illustration), and 3) introduce bag-of-timings (BoT) for classification. To the best of our knowledge, this is the first local descriptor designed to represent spatiotemporal event dynamics of dynamic objects (see Sect. 2). In addition, it outperforms prior results for classification tasks on public

datasets, and returns finer granularity as we classify different rigidity levels, which is also a new contribution.

Our experiments on synthesized and real data with ground truth showcase classification of 3D dynamic surface keypoints with respect to deformation types. In particular, we use public datasets of 3D video of human performances which are challenging, of good quality, and popular in CV and CG publications (i.e., data from MIT [9], Univ. Surrey [37], and INRIA [3]). The next section deals with related work. Section 3 presents the proposed dynamic surface spatiotemporal local descriptor. Section 4 gives details on the BoT paradigm. Section 5 shows experimental results on synthetic and real-world datasets. Section 6 concludes with a discussion on our contributions.

2. Related Work

Local descriptors have been widely used as core methods in computer vision research and applications for the past two decades (e.g., for 3D reconstruction, or object recognition in videos). Performances and limitations usually rely on their ability to be invariant to certain classes of transformation. The literature contains numerous work on texture-based local description of 2D images [27, 29, 41]. As well, the literature has provided several descriptors for 3D shape models [39, 40]. However, most of existing 3D descriptors are either designed for static and synthesized objects, require surface texture, or are too sparse for object characterization, while in our framework we deal with real-world dynamic surfaces which can be textureless and noisy.

Nowadays, full 3D capture systems have become popular and accurate 3D dynamic surface can be obtained using various kinds of sensing devices (e.g., multiview video cameras, RGB-D sensors, or 3D laser scanners) [22, 28, 1, 9, 19, 20]. Collected information is usually represented as a stream of surface mesh models undergoing free-form deformation across time, which geometrical structure (e.g., surface mesh connectivity) can be kept consistent using 3D scene flow estimation, or surface-point matching and tracking [10, 26, 3, 42, 17]. Hence, low-frequency surface details (e.g., wrinkles on solid color clothing) can be tracked accurately, and deformation dynamics can be characterized for various classification tasks [43].

Dynamic event modeling has received lots of interest from the scientific community. Particularly, linear dynamical systems (LDS), which are a generalization of Hidden Markov Models (HMM) [32] where the underlying state-space is continuous (instead of discrete), have been successful at modeling complex time series. Note that the Kalman filter is actually a popular method to estimate internal states from a sequence of observations assuming the underlying system is a LDS. Dynamical models have been widely used for dynamic texture modeling [34, 11], segmentation [12, 4, 44], recognition [34, 33], and for facial movement synchronization [24], or action recognition [6].

Let us mention that despite the large amount of work on static and dynamic 3D facial expression recognition (FER), temporal modeling is still unexplored for 3D dynamic facial expression recognition (see survey [36]). FER methods usually involve tracking of landmarks (obtained from photometric-based local feature extraction), and rely on feature displacements (e.g., FACS). However to date, no *free* dataset with *annotated landmarks* is available despite recent progress [8]. Similarly, spatiotemporal statistical analysis of medical data employs learning of specific motion patterns (e.g., based on growth [14] or velocity [13]) that often result from data simulations or involve additional physiological features. Finally, work on non-rigid surface detection or modeling also refer to surface registration techniques or (local) stress/fold region localization, which are out of scope of this paper. For example, in [35] a 3D template (planar mesh) is fit to a 2D textured surface using linear local models assuming a single class of object.

In this paper, we use complex datasets representing real human performances as introduced in Sect. 1. Deformation pattern models are necessary to perform classification tasks as opposed to fitting rigid transformations, just as fitting pixel trajectory is insufficient to perform classification of dynamic texture [11, 4]. Note that the task is also difficult for humans because viewpoint and visual aspect can be ambiguous.

3. Timing-Based Local Descriptor

We propose a low-level local descriptor to model surface deformation dynamics. The descriptor captures spatiotemporal event statistics between neighboring keypoints using timing structure of linear dynamical systems (LDS).

3.1. Background

Complex surface deformations can be modeled using sets of LDS where observations across time are given by surface intrinsic property variations such as shape indices [43].

Shape index. The shape index σ describes local surface topology at each surface point in terms of the principal curvatures as a continuous parameter, while being invariant to surface orientation [25]. σ encodes a curvature type (cup, rut, saddle rut, saddle ridge, ridge, dome, cap) following its value (see Fig. 1): $\sigma = \frac{2}{\pi} \arctan \frac{\kappa_2+\kappa_1}{\kappa_2-\kappa_1} \in [-1, 1]$, where κ_1 and κ_2 are the principal curvatures ($\kappa_1 \geq \kappa_2$). Natural scene classification is known to be more stable with shape index than with Gaussian and mean curvatures. For curvature tensor estimation, we implemented [7] based on normal cycle. The tensor is averaged with Laplacian over geodesic region. It can adapt to resolution sampling and filter noisy objects.

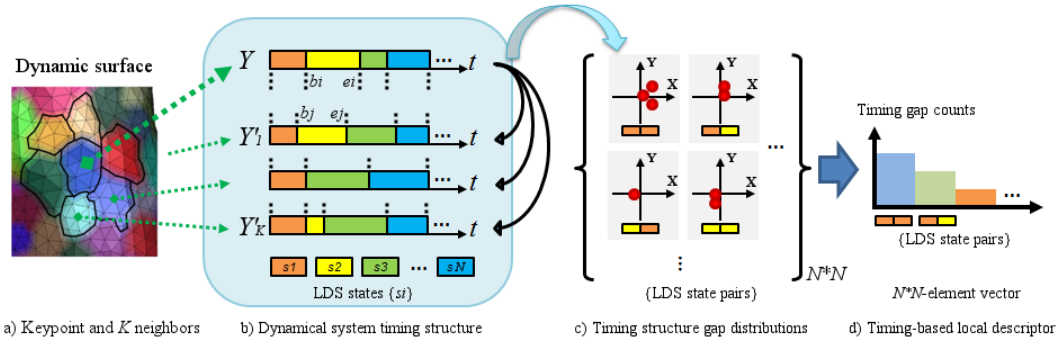

Figure 2. Overview of timing-based local descriptor creation. a) Multivariate observations of intrinsic surface property at neighboring keypoints. b) Representation of dynamical system timing structures using intervals $\{(s_k, \tau_k)\}$. c) Distributions of timing structure gaps between system state pairs (each dot is an overlapping occurrence). The X-Y axis represent differences between starting ($\Delta b = b_i - b_j$) and ending ($\Delta e = e_i - e_j$) times of two states. d) Vector-based representation of the descriptor as a histogram.

Linear dynamical systems. Let us assume a temporal sequence of multivariate observations $Y = \{y(t)\}_{t\geq 0}$, $y(t) \in \mathbf{R}^c$, and their hidden states $X = \{x(t)\}_{t\geq 0}$, $x(t) \in \mathbf{R}^a$ belonging to a continuous state space. A linear dynamical system D_i can then be defined as:

$$\begin{cases} x(t+1) &= A_i x(t) + g_i + v_i(t) \\ y(t) &= C x(t) + w(t), \end{cases} \quad (1)$$

where $A_i \in \mathbf{R}^{a\times a}$ is the state transition matrix which models the dynamics of D_i, g_i is a bias vector and $C \in \mathbf{R}^{c\times a}$ is the observation matrix which maps the hidden states to the output of the system by linear projection. $v_i(t) \sim \mathcal{N}(0, Q_i)$ and $w(t) \sim \mathcal{N}(0, R)$ are process and measurement noises modeled as Gaussian distributions with null averages and Q_i and R covariances respectively. Eq. 1 has been widely utilized to model complex spatiotemporal variations (e.g., for dynamic textures [11, 33], human actions [6]). For heterogeneous scenes or patterns, a mixture of N LDS are used and all parameters are estimated by Expectation-Maximization (e.g., see dynamic texture segmentation [4], and facial movement recognition [24]).

3.2. System Dynamics Timing Structure

Timing structure of LDS represents the temporal relationship between multiple LDS. Particularly, we model timing gaps between event dynamics at neighboring keypoints of dynamic surface.

Timing structure. We use a hybrid linear dynamical system model (HDS) to represent both dynamical and discrete-event systems [4, 24]. Dynamical systems are usually described by differential equations and are suitable for modeling smooth and continuous physical phenomena (see Eq. 1), while discrete-event systems describe discontinuous changes in physical phenomena and in subjective or intellectual activities. The HDS consists of: (1) a set of N LDS $\mathcal{D} = \{D_1, ..., D_N\}$, and (2) a finite state machine (FSM) that serves as an abstraction for LDS states and transitions. The FSM models the system state transitions (i.e., switching) between the discrete set of states $\mathcal{S} = \{s_i\}_{i=1...N}$, where each FSM state s_i corresponds to an LDS D_i (see [2, 31]). Particularly, state transitions usually occur when (sudden) changes are observable. Thus, a sequence (of length T) of observed dynamic events Y can be represented using N_t intervals (or segment models [30]): $\{I_k\} = \{(s_k, \tau_k)\}$, where $k = 1, ..., N_t$, $s_k \in \mathcal{S}$ identifies a state, $\tau_k = e_k - b_k$ is the duration of s_k, b_k and e_k are starting and ending times of s_k respectively, and $\sum_{k=1}^{N_t} \tau_k = T$. We exploit this representation to characterize repetitive events. See Fig. 2b) for illustration.

System dynamics timing gaps. Let us assume two observation sequences $Y \equiv \{I_k\}$ and $Y' \equiv \{I_{k'}\}$ (e.g., observations from two neighboring surface keypoints), and the set of overlapping interval pairs $\mathcal{I} = \{(I_k, I_{k'}) \in Y \times Y' : [b_k, e_k] \cap [b_{k'}, e_{k'}] \neq \emptyset\}$. The distribution of timing structure gaps between two states $s_m \in \mathcal{S}$ and $s_n \in \mathcal{S}'$ (of Y and Y' respectively) can be defined as follows:

$$\begin{aligned} &P(b_k - b'_k = \Delta b, e_k - e_{k'} = \Delta e| \\ &s_k = s_m, s_{k'} = s_n, (I_k, I_{k'}) \in \mathcal{I}), \end{aligned} \quad (2)$$

where Δb and Δe represent differences between starting and ending of two states s_k and $s_{k'}$ respectively. Note that if $\forall (I_k, I_{k'})$, $|b_k - b_{k'}| \to 0$ and $|e_k - e_{k'}| \to 0$, then all pairs of overlapping intervals are synchronized. Eq. 2 tells how much two states are synchronized statistically (Fig. 2c). In practice, we introduce a temporal threshold $d_{\max}$ to discard unrelated events:
$\|(b_k - b_{k'}, e_k - e_{k'})\|_2 > d_{\max} \Rightarrow [b_k, e_k] \cap [b_{k'}, e_{k'}] = \emptyset$.

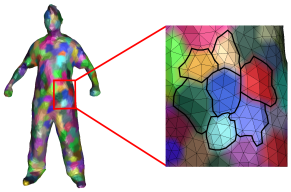

Figure 3. Local descriptor spatial layout on Bouncing model (450 keypoints) [3]. Here, a keypoint consists of 6 connected vertices, and we consider the $K = 6$ nearest keypoints as neighborhood.

3.3. Descriptor Spatial Layout

State-of-the-art dynamic event models usually consider sets of independent features sparsely distributed in space [4, 38] or in time [24, 43]. However, natural scenes (e.g., in 3D video or dynamic texture) often exhibit both time and space dependent observations: a point $\mathbf{P}(t)$ at time t is correlated to observations of $\mathbf{P}(t+1)$ and $\mathbf{P}'(t)$, where $\mathbf{P}'(t)$ belongs to the (spatial) neighborhood $\mathcal{V}(\mathbf{P}(t))$ of $\mathbf{P}(t)$. Hence, we propose to define our local descriptor by considering multivariate observations at neighboring keypoints.

We use a star-shaped spatial layout where observation sequences are collected at the central keypoint and from its neighbors [27, 45]. In the case of 3D video data, dynamic surfaces are aligned and consistently segmented into regular regions (e.g., sets of 6 connected vertices) which represent keypoints (Fig. 3). This strategy allows us to alleviate 3D video reconstruction surface noise and numerical approximation inherent to mesh-based representations [3].

3.4. Histogram of Timing Gaps

Following the description above, we form a low-level local descriptor as a sparse histogram by assigning length counts of timing structure gaps between the LDS states computed at a keypoint and its K neighbors:

1. For each keypoint $\mathbf{P}_i, i = 1, ..., V$, and its nearest neighbors $\{\mathbf{P}_j\}, j = 1, ..., K$, we compute the LDS timing structures from $Y_i \equiv \{I_k\}$ and $\{Y_j\} \equiv \{\{I_{k'}\}\}$ respectively (see Sect. 3.2).

2. For each interval I_k, we compute the timing structure gaps between I_k and all intervals $I_{k'}$ (see Fig. 2b)).

3. Assuming Y_i and $\{Y_j\}$ have N_i and N_j states respectively, we initialize an empty histogram with $N_i * N_j$ bins for each $\mathbf{P}_i$; and then the pairs of overlapping intervals $(I_k, I_{k'}) \in \mathcal{I}$ (see Sect. 3.2) contribute to a descriptor bin $b_{m,n}$, where $m = 1, ..., N_i$ and $n = 1, ..., N_j$, as follows:

$$b_{m,n} = \sum_{(I_k, I_{k'}) \in \mathcal{I}} \Big(P(I_k | s_k = s_m, Y_i) \\ * P(I_{k'} | s_{k'} = s_n, Y_j) * w(I_k, I_{k'}) \Big), \tag{3}$$

where $P(I_k | s_k = s_m, Y_i) = \tau_k / T$ and $P(I_{k'} | s_{k'} = s_n, Y_j) = \tau_{k'} / T$ are probabilities that measure the relevance of intervals I_k and $I_{k'}$ respectively (e.g., relative duration of intervals), and $w(I_k, I_{k'})$ quantifies the interval pair synchronization:

$$w(I_k, I_{k'}) = 1 - \frac{\|(b_k - b_{k'}, e_k - e_{k'})\|_2}{d_{\max}}. \tag{4}$$

Hence, contributions to $b_{m,n}$ are higher with pairs of long and well synchronized intervals. The overall scheme for descriptor creation is given in Fig. 2. As observations are obtained from orientation invariant intrinsic features (see shape index in Sect. 3.1) , and contributions are collected from unordered circular neighborhood, the proposed descriptor is also orientation invariant.

In practice, a local descriptor is computed for each surface keypoint, whose neighborhood consists of the $K = 6$ nearest keypoints. The actual implementation uses a set of $N = 6$ LDS to model all intrinsic surface variation dynamics, which leads to a $6 * 6 = 36$-element vector for each local descriptor. We set $d_{\max} = 0.1s$ by heuristics.

4. Bag-of-Timings

We introduce the bag-of-timings (BoT) paradigm for dynamic surface classification, by treating timings of local surface element (i.e., keypoint) dynamics as words to be represented in a codebook, while in previous work LDS state parameters were used as words (i.e., not timing structure gaps) [33, 43]. Bag-of-words (BoW) are also used for image classification where image features serve as words [15].

Definition. A bag-of-timings (BoT) is a sparse vector of occurrence counts of a vocabulary of dynamic surface local descriptors. It is represented as a sparse histogram of dynamic state timings as presented in the previous section. To form a codebook (i.e., find the codewords), we cluster all local descriptors using the K-medoids algorithm [23], which is known to improve the robustness to noise and outliers of the clustering, and is computationally more efficient compared to K-means. Here, a medoid is the descriptor of a cluster, whose average distance to all other descriptors in the cluster is minimal. Distances between descriptors can be computed using standard histogram kernel d_H (see Sect. 5). Assuming the total number V of local descriptors on a dynamic surface, a pairwise distance matrix $D \in \mathcal{R}^{V \times V}$ is computed only once to obtain the set of G clusters, whose centers $\{F_1, ..., F_G\}$ stand for the codewords.

Soft-weighting. Let us consider the set of V descriptors $\{d_i\}_{i=1...V}$ extracted from all points (i.e., the keypoints) of

a dynamic surface. Within the BoT framework, each descriptor d_i contributes to a set of weights $\{w_{i1}, ..., w_{iG}\}$ associated to the codewords $\{F_1, ..., F_G\}$ that characterize the object (e.g., for classification). We use *soft-weighting* as it is less sensitive to noise compared to other weighting schemes (e.g.,*term frequency* and *inverse document frequency* [21, 33]):

$$w_{il} = \sum_{j=1}^{M} \sum_{i=1}^{N_j} \frac{1}{2^{j-1}} \text{sim}(d_i, F_l), l = 1, ..., G, \quad (5)$$

where $M = 4$ is the number of nearest codewords, N_j is the number of descriptors whose j^{th} closest codeword is F_l, $\text{sim} = 1 - \frac{d_H}{\max(D)}$ is a similarity measure between descriptors, and $\max(D)$ is the biggest element of the matrix D (d_H and D are as defined above). Finally, the set of weight characterizing the surface is normalized with L_1 norm.

For classification tasks, we use Support Vector Machines (SVM) with Radial Basis Function (RBF) kernel to discriminate the codewords: $\mathcal{K}(x, y) = \exp^{-\gamma d(x,y)}$, where γ is a free parameter learned by cross-validation, and d is a distance in the histogram space. We use RBF kernels with distances $d(x_i, y_i) = \sum_i |x_i - y_i|^b$, where $b < 2$, that are Laplacian and sub-linear, popular in image retrieval, and satisfy the Mercer's condition [16]. In our experiments, SVM show more stability than Nearest Neighbor (NN).

5. Experiments

Experimental results for dynamic surface classification and segmentation are obtained with synthesized and real-world public datasets of 3D video data. For baseline comparison, we use state-of-the-art methods employing dynamical system models [34, 5, 33, 43].

5.1. Classification

Baseline for comparison. In [5, 34], the authors use a single LDS to model a video sequence (of dynamic textures), the Martin distance is used to calculate distances between LDS, and NN and SVM are used for classification. In [33], the authors use one LDS per video feature, and a bag-of-system (BoS) model with SVM for classification. In [43], the authors use a set of N LDS to characterize surface dynamics, and a BoS that accounts for the statistical distribution of LDS in time. We abbreviate these approaches D+NN, D+SVM, BS+SVM_1, and BS+NN_N and BS+SVM_N respectively. In what follows, we show that the proposed local descriptor returns best performance, while state-of-the-art approaches are ineffective for certain challenging scenario.

Synthesized datasets. We created a synthesized dynamic surface dataset to evaluate the timing-based local descriptor performance.The dataset represents a rectangular surface divided into 8 equal regions (ROI) undergoing various deformations across time. The sequence consists of 176 frames and the mesh contains 4000 vertices ($40 * 100$). Each region undergoes a sinusoidal deformation: $y = \lambda\alpha \sin(\beta\pi x)$, where λ is a constant scale factor, and $\alpha = 6, 4, 2, 8$ and $\beta = 10, 2, 7, 5$ respectively, that occurs in one out of two orthogonal directions (in order to evaluate the orientation invariance property of the local descriptor). Furthermore, a random noise ($< 5\%$) is added to all vertex positions to simulate real 3D video reconstruction artifacts. As designed, state-of-the-art methods [34, 5, 33] (without timing information) cannot be used to discriminate the different regions as all dynamical system models are identical all over the surface. Hence, they are all assigned to the same class.

Figure 4a) shows a frame of the sequence. As deformations are subtle, all 8 regions appear flat. A shape index (see Sect. 3) is computed at each surface point and allows us to represent local topology (e.g., cups in blue, caps in yellow as shown in Fig. 4b)). In Fig. 4c), the surface is sampled into 250 regular regions (i.e., keypoints) where observation sequences are obtained across time. We use $N = 6$ LDS to model surface deformation dynamics. Figure 4d) shows timing structures with interval representation at some keypoints (with $K = 4$ neighbors). Evaluations of our method were performed with different parameters for the sake of optimization ($K = 1, 4, 6, 8$, $d_{\max} = 0.1, 0.2$, etc.). We also tested several histogram distances d_H and found that Manhattan distance (reported) and histogram intersection return best performances. Classification tasks were performed using NN, and SVM where 50% of the keypoints served for training, and 50% for testing (as described in Sect. 4). We reported in Table 1 the baseline comparison with $K = 4$ and $d_{\max} = 0.1$.

Table 1. Synthesized data ROI classification.

	ROI1+5	ROI2+6	ROI3+7	ROI4+8
[34, 5, 33]	cannot discriminate			
BS+NN_N[43]	30.0%	100%	52.0%	78.3%
BS+SVM_N[43]	62.4%	94.8%	78.4%	74.8%
BT+NN[ours]	16.7%	97.2%	82%	86.6%
BT+SVM[ours]	**90.0%**	**98.3%**	**83.3%**	**86.6%**

Thus, the experiments on synthesized data show that our method outperforms state-of-the-art techniques (which are not designed to cope with such scenario) for classification of dynamical surfaces undergoing subtle deformations (see Fig. 4). The results also highlight orientation invariance, resistance to noise, and high classification performance.

Real-world datasets. For further evaluations, we use real-

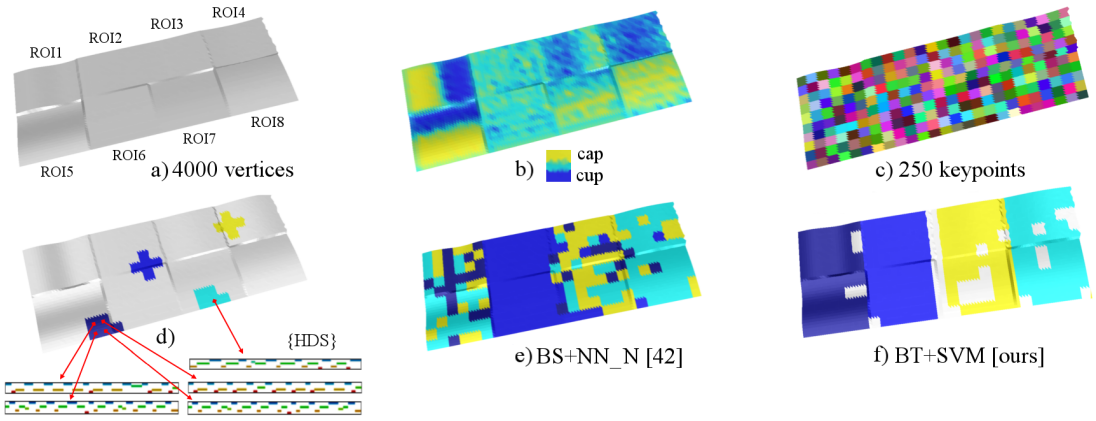

Figure 4. Synthesized data classification using timing-based local descriptor. a) Synthesized dynamic surface containing various deformations (4000 vertices with noise). b) Local curvature estimated using shape index. c) Surface sampling into 250 keypoints. d) HDS computed for each keypoint and $K = 4$ neighbors. e) Classification using BS+NN_N [43]. f) BT+SVM [ours] outperforms state-of-the-art.

world public datasets of 3D video sequences reconstructed from real human performances: Free and Lock sequences from the University of Surrey [37], Samba, Bouncing, Handstand and Crane from MIT CSAIL [9]. The sequences represent real humans performing various actions, such as turning, dancing, and jumping. Most subjects wear loose clothing (e.g., T-shirt) whose details were accurately reconstructed. The sequences were aligned using [3] in order to track surface points and extract observations (i.e., shape index) across time. Nevertheless, we can observe that reconstructed surfaces from [37] (such as Free) have local surface noise due to drawbacks from multiview stereo reconstruction, despite being visually compelling. On the other hand, surfaces from [9] contain drawbacks for spatiotemporal reconstruction, and therefore reconstructed surfaces seem more rigid and less prone to wrinkle. However, temporal statistics can still be observed in different regions of the object.

We evaluate our method against [34, 33, 43] for quantitative evaluations of rigid/non-rigid surface classification. Each surface point is manually labeled as rigid or non-rigid (see Fig. 5). Note that it can be confusing to determine whether a region is rigid or non-rigid without referring to videos. For example, in the Samba dataset, the subject wears a dress that moves during the dance. However, the top of the dress is tight and has little variations, while the bottom of the dress is looser. For classification with SVM, we selected 50% of keypoints for training, and evaluated the classification with the remaining keypoints. Results are reported in Table 2 (with $K = 6$). Confusion matrices between sequences also return consistent keypoint rigid/non-rigid classification ($> 90\%$ true positives). Our approach performs much better than state-of-the-art methods.

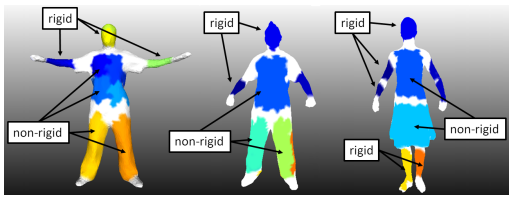

Figure 5. Annotations of rigid/non-rigid surface regions.

Table 2. Rigid/non-rigid surface classification (success ratio %).

	Free	Bouncing	Samba
#vertices	4284	3848	5530
#keypoints	514	450	361
#frames	175	174	174
D+NN [34]	56%	42%	50%
D+SVM [34]	52%	50%	45%
BS+SVM_1 [33]	66%	70%	67%
BS+SVM_N [43]	85%	89%	82%
BT+SVM_N [ours]	**89%**	**92%**	**87%**

5.2. Segmentation

As discussed above, some regions could have be misclassified due to wrong annotation (e.g., around the chest for Samba), hence we perform timing-based descriptor segmentation to identify those regions. Note that in that case we have to cope with surface noise, such as 3D reconstruction artifacts or drawbacks from spatiotemporal constraints (i.e., unique templates are deformed over time) [3]. Here, we use $K = 6$ neighbors for each descriptor, and achieve clustering using K-medoids. Segmentation is evaluated qualitatively. Surfaces are clustered into 4 clusters, where each cluster stands for a rigidity class (i.e., from less

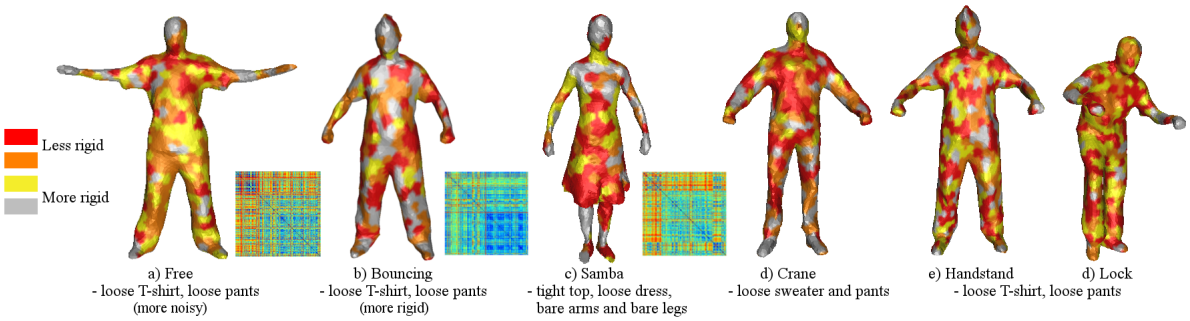

Figure 6. Real-world dynamic surface segmentation (using $K = 6$ neighbors and K-medoids for clustering). Rigid regions are mostly located on faces and bare limbs, while less rigid regions (i.e., that deform the most) follow clothing folds and tips, and body joints. Unaccuracies are due to 3D reconstruction artifacts (e.g., surface noise). In a), b) and c), pairwise descriptor distance matrices highlight good clustering. (Blocs correspond to different body regions.)

to more rigid). Results are shown in Fig. 6. For the Samba sequence in a), most rigid regions concern the face, bare limbs, and the top of the body which is covered by tight clothing. For other sequences, the face region and forearms are mostly always rigid, while neck and joints are usually less rigid. Regions corresponding to loose clothing (and particularly along folds and at tips) belong to non-rigid clusters, which is correct. We also provide pairwise descriptor distance matrices. Here, rows and columns were consistently reordered with respect to hand-made classes to highlight the clustering efficiency (see matrix blocs).

Finally, we perform further segmentation validation using dynamic face datasets, as the face structure is readable and does not require a tedious annotation process. Sequences representing human faces are obtained by fitting a face model (112 vertices) to real-world RGB-D data captured using Kinect sensor. Face models are then segmented into 3 clusters using our approach (with $K = 6$ neigbors). The sequence shown in Fig. 7 contains 300 frames and represents a singing human face. As observed, timing-based descriptor segmentation highlights the different face regions that are stressed during the performance (i.e., eyes, nose, mouth, jaw, and forehead). To our knowledge, timing-based surface segmentation with respect to rigidity level is a new result that can potentially have numerous applications. Particularly segmentation cannot efficiently be achieved using LDS state-based methods [34, 33, 43].

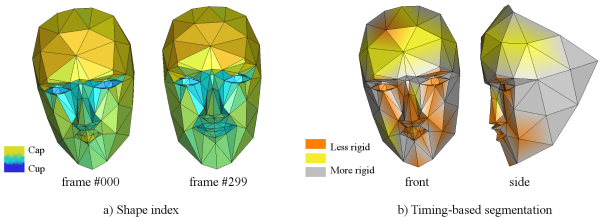

Figure 7. Dynamic face sequence. a) Face model shape indices. b) Timing-based descriptor segmentation highlights active regions.

6. Conclusion

In this paper, we present the first local descriptor designed for surface dynamics modeling. The descriptor captures local event dynamics timing structure between surface keypoints. The approach is novel compared to the state-of-the-art that usually relies on dynamical system state parameters to characterize dynamic events (of dynamic texture). Timing-based local descriptors are computed from surface intrinsic property variation dynamics, which are modeled using hybrid linear dynamical systems (HDS). Particularly, the descriptor accounts for local event dynamics timing gaps. We also introduce the bag-of-timings (BoT) paradigm to locally characterize dynamic surfaces at keypoints. The proposed descriptor is orientation invariant by design, and can be used with (noisy) real-world data. Evaluations are performed on challenging synthesized and real-world datasets. We show that the local descriptor can be used for dynamic surface classification and segmentation with respect to rigidity level, which is a new result. We believe our model is promising for future research and applications involving dynamic geometrical data obtained from accurate 3D vision techniques or depth sensors (e.g., Kinect), and also data such as soft tissue organs (e.g., heart, liver), cloths or fluids for diagnosis and anomaly detection. Real-timeness should be achievable using observation windows with online Viterbi algorithm (as with online HMM).

References

[1] J. Allard, C. Ménier, B. Raffin, E. Boyer, and F. Faure. Grimage: Markerless 3d interactions. *SIGGRAPH - Emerging Technologies*, 2007. 2

[2] C. Bregler. Learning and recognizing human dynamics in video sequences. *CVPR*, 1997. 3

[3] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. *ECCV*, 2010. 2, 4, 6

[4] A. B. Chan and N. Vasconcelos. Mixtures of dynamic textures. *ICCV*, 2005. 1, 2, 3, 4

[5] A. B. Chan and N. Vasconcelos. Classifying video with kernel dynamic textures. *CVPR*, 2007. 5

[6] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. *CVPR*, 2009. 2, 3

[7] D. Cohen-Steiner and J.-M. Morvan. Restricted delaunay triangulations and normal cycle. *Symposium on Computational Geometry*, 2003. 2

[8] D. Cosker, E. Krumhuber, and A. Hilton. A facs valid 3d dynamic action unit database with applications to 3d dynamic morphable facial modelling. *ICCV*, 2011. 2

[9] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Trans. Graphics*, 27(3), 2008. 2, 6

[10] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Markerless deformable mesh tracking for human shape and motion capture. *CVPR*, 2007. 2

[11] G. Doretto, A. Chiuso, Y. Wu, and S. Soatto. Dynamic textures. *IJCV*, 51(2):91–109, 2003. 1, 2, 3

[12] G. Doretto, D. Cremers, P. Favaro, and S. Soatto. Dynamic texture segmentation. *ICCV*, 2003. 1, 2

[13] N. Duchateau, M. D. Craene, G. Piella, E. Silva, A. Doltra, M. Sitges, B. H. Bijnens, and A. F. Frang. A spatiotemporal statistical atlas of motion for the quantification of abnormal myocardial tissue velocities. *Medical Image Analysis*, 15(3):316–328, 2011. 2

[14] S. Durrleman, X. Pennec, A. Trouvé, J. Braga, G. Gerig, and N. Ayache. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. *IJCV*, 103(1):22–59, 2013. 2

[15] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. *CVPR*, 2005. 4

[16] C. Fowlkes, S. Belongie, F. Chung, and J. Malik. Spectral grouping using the nystrom method. *PAMI*, 2009. 5

[17] J. Franco and E. Boyer. Learning temporally consistent rigidities. *CVPR*, 2011. 2

[18] J. Franco, C. Menier, E. Boyer, and B. Raffin. A distributed approach for real-time 3d modeling. *CVPR Workshop*, 2004. 1

[19] P. Huang, A. Hilton, and J. Starck. Shape similarity for 3d video sequences of people. *IJCV*, 89(2-3):362–381, 2010. 2

[20] H. Jiang, H. Liu, P. Tan, G. Zhang, and H. Bao. 3d reconstruction of dynamic scenes with multiple handheld cameras. *ECCV*, 2012. 1, 2

[21] Y.-G. Jiang, C.-H. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. *CIVR*, 2007. 5

[22] 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

[23] L. Kaufman and P. Rousseeuw. Clustering by means of medoids. *Statistical Data Analysis Based on the L1-Norm and Related Methods, Y. Dodge Ed., North-Holland*, 1987. 4

[24] H. Kawashima and T. Matsuyama. Interval-based modeling of human communication dynamics via hybrid dynamical systems. *NIPS Workshop on Modeling Human Communication Dynamics*, 2010. 2, 3, 4

[25] J. Koenderink and A. van Doorn. Surface shape and curvature scales. *Image and Vision Computing*, 1992. 2

[26] H. Li, R. W. Sumner, and M. Pauly. Global correspondance optimization for non-rigid registration of depth scans. *Computer Graphics Forum, Proc. SGP*, 27(5), 08. 2

[27] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 2, 4

[28] 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*, 96(3):393–434, 2004. 1, 2

[29] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *PAMI*, 10(27):1615–1630, 2005. 2

[30] M. Ostendorf, V. Digalakis, and O. Kimball. From hmm's to segment models: a unified view of stochastic modeling for speech recognition. *IEEE Trans. Speech and Audio Processing*, 4(5):360–378, 1996. 3

[31] V. Pavlovic and J. M. Rehg. Impact of dynamic model learning on classification of human motion. *CVPR*, 2000. 3

[32] L. R. Rabiner. A tutorial on hidden markow models and selected applications in speech recognition. *IEEE*, 77(2):257–286, 1989. 2

[33] A. Ravichandran, R. Chaudhry, and R. Vidal. View-invariant dynamic texture recognition using a bag of dynamical systems. *CVPR*, 2009. 1, 2, 3, 4, 5, 6, 7

[34] P. Saisan, G. Doretto, Y. Wu, and S. Soatto. Dynamic texture recognition. *CVPR*, 2001. 1, 2, 5, 6, 7

[35] M. Salzmann and P. Fua. Linear local models for monocular reconstruction of deformable surfaces. *PAMI*, 33(5):931–944, 2011. 2

[36] G. Sandbach, S. Zafeiriou, M. Pantic, and L. Yin. Static and dynamic 3d facial expression recognition: A comprehensive survey. *Image Vision Computing*, 30(10):683–697, 2012. 2

[37] J. Starck and A. Hilton. Surface capture for performance-based animation. *IEEE CGA*, 2007. 2, 6

[38] Y. Sun and L. Yin. Facial expression recognition based on 3d dynamic range model sequences. *ECCV*, 2008. 4

[39] J. W. H. Tangelder and R. C. Veltkamp. A survey of content based 3d shape retrieval methods. *Multimedia Tools Appl.*, 39(3):441–471, 2008. 2

[40] A. Tevs, A. Berner, M. Wand, I. Ihrke, and H.-P. Seidel. Intrinsic shape matching by planned landmark sampling. *Computer Graphics Forum*, 30:543–552, 2011. 2

[41] E. Tola, V. Lepetit, and P. Fua. A fast local descriptor for dense matching. *CVPR*, 2008. 2

[42] T. Tung and T.Matsuyama. Dynamic surface matching by geodesic mapping for 3d animation transfer. *CVPR*, 2010. 2

[43] T. Tung and T.Matsuyama. Intrinsic characterization of dynamic surfaces. *CVPR*, 2013. 1, 2, 4, 5, 6, 7

[44] R. Vidal and A. Ravichandran. Optical flow estimation and segmentation of multiple moving dynamical textures. *CVPR*, 2005. 1, 2

[45] S. Winder, G. Hua, and M. Brown. Picking the best daisy. *CVPR*, 2009. 4