On Mean Pose and Variability of 3D Deformable Models

Benjamin Allain, Jean-Sébastien Franco, Edmond Boyer, Tony Tung

On Mean Pose and Variability of 3D Deformable Models

Benjamin Allain[1], Jean-Sébastien Franco[1], Edmond Boyer[1], and Tony Tung[2]

[1] LJK, INRIA Grenoble Rhône-Alpes, France
[2] Graduate School of Informatics, Kyoto University, Japan
firstname.lastname@inria.fr, tony2ng@gmail.com

Abstract. We present a novel methodology for the analysis of complex object shapes in motion observed by multiple video cameras. In particular, we propose to learn local surface rigidity probabilities (i.e., deformations), and to estimate a mean pose over a temporal sequence. Local deformations can be used for rigidity-based dynamic surface segmentation, while a mean pose can be used as a sequence keyframe or a cluster prototype and has therefore numerous applications, such as motion synthesis or sequential alignment for compression or morphing. We take advantage of recent advances in surface tracking techniques to formulate a generative model of 3D temporal sequences using a probabilistic framework, which conditions shape fitting over all frames to a simple set of intrinsic surface rigidity properties. Surface tracking and rigidity variable estimation can then be formulated as an Expectation-Maximization inference problem and solved by alternatively minimizing two nested fixed point iterations. We show that this framework provides a new fundamental building block for various applications of shape analysis, and achieves comparable tracking performance to state of the art surface tracking techniques on real datasets, even compared to approaches using strong kinematic priors such as rigid skeletons.

Keywords: Shape dynamics, Motion analysis, Shape spaces

1 Introduction

Recent years have seen the emergence of many solutions for the capture of dynamic scenes, where a scene observed by several calibrated cameras is fully reconstructed from acquired videos using multiview stereo algorithms [24,12,1,20]. These techniques have many applications for media content production, interactive systems [2] and scene analysis [28] since they allow to recover both geometric and photometric information of objects' surface, and also their shape and evolution over time. Since these temporal evolutions were initially reconstructed as a sequence of topologically inconsistent 3D models, significant research work has been done for full 4D modeling and analysis of geometrically time-consistent 3D sequences.

In particular, several techniques propose to deform and match a template to either image data, or to intermediate 3D representations of the surface [25,17,9,26].

These methods allow the recovery of both shape and motion information. However they usually do not consider the intrinsic dynamic properties of a surface. These are either assumed, for instance through a kinematic structure (rigging) or through the surface tension parameters, or are simply ignored. Hence, there is a large interest in better understanding rigidity and motion properties of shapes, with the prospect of improving dynamic models, extracting more useful information, and better automation. In this work, we take the estimation a step further and investigate how to infer dynamics or statistical properties of shapes given temporal sequences.

Recovering this information is yet a largely open research topic with only few exploratory representations proposed for dynamics characteristics of surfaces, e.g. [11,29]. We propose a novel inference framework for the analysis of complex object shapes in motion that learns local surface rigidity probabilities (i.e., deformations), and estimates a mean pose over a temporal sequence. Based on recent advances in surface tracking techniques, we formulate a generative model of 3D temporal sequences using a probabilistic framework, which conditions shape fitting over all frames to a simple set of intrinsic surface rigidity properties. Surface tracking and rigidity variables can then be obtained iteratively using Expectation-Maximization inference by alternatively minimizing two nested fixed point iterations. Thus, our main contribution is a framework that allows the simultaneous tracking and inference of dynamic properties of object surfaces given temporal observations. We show how these properties contribute to a better understanding of surface motion and how they can be used for the dynamic analysis of 3D surface shapes through mean pose estimation and rigidity-based segmentation, while achieving competitive surface tracking.

The remainder of the paper is organized as follows. The next section discusses related work. Details on the mean pose inference model are given in Sect. 3. Section 4 presents various applications and experimental results. Section 5 concludes with a discussion on our contributions.

2 Related Work

The analysis of deformable surfaces captured by multi-video systems has gained lot of interest during the last decade due to the rapid progression of computer and image sensing technologies. We focus here on works that relate to dynamic properties of shapes.

Kinematic structures. Many popular tracking methods propose to rigidly constrain a model using an articulated structure, for instance a skeleton or a cage, which must be scaled and rigged to a 3D template, and optimally positioned through a sequence of models representing the observed subjects [4,30,17,19]. The template is usually deformed using a skinning technique, according to the optimized structure across the sequence [5]. Such kinematic structures provide intrinsic information on the associated shapes through their parameter evolutions (e.g. their averages can define a mean pose). These approaches require a

priori knowledge on the observed shapes, such as the topology and the rigid parts, and cannot be applied to arbitrary object shapes. Moreover global template deformation across time is subject to loss of local details such as cloth wrinkles and folds.

Locally rigid structures. The literature also contains several methods that relax the constraint on the shape structure using looser rigidity priors. A body of works consider deformations that preserve local intrinsic surface properties, e.g. isometric deformations [21,8,22,23]. Such properties relate to local rigidities, for instance in [31,32] local surface distorsions are constrained, however they are usually known priors. While efficient to register or match surfaces, intrinsic surface properties are not necessarily sufficient to track complex shapes such as human bodies. In that case, several approaches introduce local deformation models to drive surface evolutions. For instance, in [9], the observed surface is treated as a piece-wise body with locally rigid motions. We consider a similar model to represent surface deformations which is used to learn local rigidities as well as mean poses along with the tracking. Interestingly, recent approaches also in this category were proposed to characterize local surface deformations. In [11], the authors propose a probabilistic framework for rigid tracking and segmentation of dynamic surfaces where the rigid kinematic structure is learned along time sequences. Our framework does not assume such structure but learns instead local rigidities and mean poses. In [29], the authors model complex local deformation dynamics using linear dynamical systems by observing local curvature variations, using a shape index, and perform rigidity-based surface patch classification. The latter approach assumes surface alignment is given, in contrast to our proposed generative model that simultaneously performs surface tracking and local rigidity estimation.

Shape Spaces. Following the work of Kendall [18], a number of works consider shape spaces that characterize the configurations of a given set of points, the vertices of a mesh for instance. This has been used in medical imaging to estimate mean shapes through Procrustes analysis, e.g. [16]. In this case, the shape of the object is the geometrical information that remains when the pose (i.e., similarities) is filtered out. Thus Procrustes distances can be used to measure shape similarities and to estimate shape averages with Fréchet means. We follow here a different strategy where a shape space represents the poses of a single shape and where we estimate a mean pose instead of a mean shape. This relates to other works in this category that also consider shapes spaces to model shape poses with mesh representations. They can either be learned, e.g. [3,15] or defined a priori, e.g. [27] and are used to constrain mesh deformations when creating realistic animations [3,27] or estimating shape and poses from images[15]. While sharing similarities in the deformation model we consider, our objective is not only to recover meaningful shape poses but also to measure pose similarities and intrinsic shape properties. Unlike [3,15], we do not need a pose or shape database and the associated hypothesis of its representativeness. Moreover, our methodology specifically addresses robust temporal window integration.

3 Mean Pose Inference Model

Fig. 1. Example of patch template used.

We assume given a temporal sequence of 3D reconstructions, incoherent meshes or point clouds, obtained using a multi-view reconstruction approach, e.g. [12,1,20]. We also assume that a template mesh model of the scene is available, e.g. a particular instance within the reconstructed sequence under consideration. The problem of local surface rigidity and mean pose analysis is then tackled through the simultaneous tracking and intrinsic parameter estimation of the template model. We embed intrinsic motion parameters (e.g. rigidities) in the model, which control the motion behavior of the object surface. This implies that the estimation algorithm is necessarily performed over a sub-sequence of frames, as opposed to most existing surface tracking methods which in effect implement tracking through iterated single-frame pose estimation. We first describe in details the geometric model (§3.1) illustrated with Fig. 1, and its associated average deformation parameterization for the observed surface (§3.2). Second, we describe how this surface generates noisy measurements with an appropriate Bayesian generative model (§3.3). We then show how to perform estimation over the sequence through Expectation-Maximization (§3.4).

3.1 Shape Space Parameterization

To express non-rigid deformability of shapes, while de-correlating the resolution of deformation parameters from mesh resolution, we opt for a patch-based parameterization of the surface similar to [9]. The reference mesh is partitioned in an overlapping set of patches, pre-computed by geodesic clustering of vertices. Each patch P_k is associated to a rigid transformation $\mathbf{T}_k^t \in SE(3)$ at every time t. Each position $\mathbf{x}_{k,v}$ of a mesh vertex v as predicted by the transform of P_k can then be computed from its template position $\mathbf{x}_v^0$ as follows:

$$\mathbf{x}_{k,v} = \mathbf{T}_k(\mathbf{x}_v^0). \tag{1}$$

We thus define a *pose* of the shape space as the set of patch transforms $\mathbf{T} = \{\mathbf{T}_k\}_{k\in\mathcal{K}}$ that express a given mesh deformation. Note here that a pose in the shape space does not necessarily correspond to a proper geometric realization of the reference mesh and, in practice, patch deformations are merged on the template to preserve the mesh consistency.

3.2 Mean Pose

To retrieve the mean pose of a given sequence, we provide a definition suitable for the analysis of complex temporal mesh sequences. Following Fréchet's definition of a mean [13], we introduce the *mean pose* $\bar{\mathbf{T}}$ of a given set of poses $\{\mathbf{T}^t\}_{t\in\mathcal{T}}$ over the time sequence $\mathcal{T}$ as the pose minimizing the sum of squared distances to all poses in the set:

$$\bar{\mathbf{T}} = \arg\min_{\mathbf{T}} \sum_{t\in\mathcal{T}} d^2(\mathbf{T}, \mathbf{T}^t), \tag{2}$$

where $d()$ is a distance that measures the similarity of two poses. This distance should evaluate the non-rigidity of the transformation between two poses of a shape and hence should be independent of any global pose. Such a distance is not easily defined in the non-Euclidean shape space spanned by the rigid motion parameters of the patches. However using the Euclidean embedding provided by the mesh representation, we can define a proper metric based on the vertex positions. Inspired by the deformation energy proposed by Botsch *et al.* [7] our distance is expressed as an internal deformation energy between two poses. Let $\mathbf{T}^i$ and $\mathbf{T}^j$ be two poses of the model, the distance can be written as a sum of per patch pair squared distances:

$$d^2(\mathbf{T}^i, \mathbf{T}^j) = \sum_{(P_k,P_l)\in\mathcal{N}} d^2_{kl}(\mathbf{T}^i, \mathbf{T}^j), \tag{3}$$

$$\text{with } d^2_{kl}(\mathbf{T}^i, \mathbf{T}^j) = \sum_{v\in P_k\cup P_l} \|\mathbf{T}^i_{k-l}(\mathbf{x}^0_v) - \mathbf{T}^j_{k-l}(\mathbf{x}^0_v)\|^2, \tag{4}$$

where $\mathbf{T}^i_{k-l} = {\mathbf{T}^i_l}^{-1} \circ \mathbf{T}^i_k$ is the relative transformation between patches P_k and P_l for pose i, and $\mathcal{N}$ is the set of neighboring patch pairs on the surface. The distance sums, for every pair of patches of the deformable model, its rigid deviation from pose i to j. This deviation is given by the sum over each vertex v belonging to the patch pair, of the discrepancy of relative positions of the vertex as displaced by P_k and P_l. It can be verified that d^2 defines a distance as it inherits this property from the L^2 norm used between vertices.

3.3 Generative Model

The expression (2) is useful to characterize the mean over a set of poses *already known*. Our goal however is to estimate this mean in the context where such

poses are indirectly observed through a set of noisy and sparse 3D point clouds of the surface. Thus we cast the problem as the joint estimation of mean pose and fitting of the model to each set of observations. For our purposes, we assume the set of poses $\{\mathbf{T}^t\}_{t\in\mathcal{T}}$ are defined for a set $\mathcal{T}$ corresponding to observations in a temporal sequence. The observed point clouds are noted $\mathbf{Y} = \{\mathbf{Y}^t\}_{t\in\mathcal{T}}$, where $\mathbf{Y}^t = \{\mathbf{y}_o^t\}_{o\in\mathcal{O}_t}$ is the set of point coordinates $\mathbf{y}_o^t$ for an observation o among the set of observations $\mathcal{O}_t$ at time t. Note that this set $\mathcal{O}_t$ is different than $\mathcal{V}$ in general as it is obtained from a 3D reconstruction or depth camera, without any direct correspondence to the deformable shape surface model earlier defined.

To express the noisy predictions of observations, we follow the principle of EM-ICP [14] by introducing a set of assignment variables k_o^t indicating, for each observation o, which patch this observation is assigned to. We are also interested in retrieving information about the variations of the rigid deformation with respect to the mean shape. To keep this information in its simplest form, we express in the generative model that each pair of patches $(k,l) \in \mathcal{N}$ is assigned a *binary rigidity variable* $\mathrm{c}_{kl} \in \{0,1\}$, which will condition the patch pair to accordingly be rigid or flexible. This variable is an intrinsic parameter attached to the original deformable model and is thus time-independent. We note the full set of rigidity variables $\mathrm{C} = \{\mathrm{c}_{kl}\}_{(k,l)\in\mathcal{N}}$. This in turn will allow during inference the estimation of a rigid coupling probability for each patch pair (k,l). We express the generative model through the following joint probability distribution:

$$p(\bar{\mathbf{T}}, \mathbf{T}, \mathbf{Y}, \mathrm{C}, \mathrm{K}, \sigma) = p(\bar{\mathbf{T}}) \prod_{t\in\mathcal{T}} \left(p(\mathbf{T}^t \mid \bar{\mathbf{T}}, \mathrm{C}) \prod_{o\in\mathcal{O}_t} p(\mathbf{y}_o^t \mid \mathrm{k}_o^t,\ \mathbf{T}^t, \sigma^t) \right), \quad (5)$$

with $\sigma = \{\sigma^t\}_{t\in\mathcal{T}}$ the set of noise parameters of the observation prediction model, and $\mathrm{K} = \{\mathrm{k}_o^t\}$ the set of all patch selection variables.

Observation prediction model. Each observation's point measurement is predicted from the closest vertex v within patch $P_{k=\mathrm{k}_o^t}$. Because the prediction is noisy, this prediction is perturbed by Gaussian noise of variance ${\sigma^t}^2$:

$$p(\mathbf{y}_o^t \mid \mathrm{k}_o^t\ \mathbf{T}^t \sigma^t) \ = \ \mathcal{N}(\mathbf{y}_o^t \mid \mathbf{T}_{\mathrm{k}_o^t}^t(\mathbf{x}_v^0), \sigma^t). \quad (6)$$

Pose constraining model. We constrain the fitted poses to be close to the mean pose, using the distance defined earlier (3). We embed the influence of rigidity variables in this term, by computing two versions of the distance, biased by rigidity variables C:

$$p(\mathbf{T}^t \mid \bar{\mathbf{T}}, \mathrm{C}) \ \propto \ \exp\left(- \sum_{(k,l)\in\mathcal{N}} d_{kl}^2(\bar{\mathbf{T}}, \mathbf{T}^t, \mathrm{c}_{kl}) \right), \quad (7)$$

$$\text{where } d_{kl}^2(\bar{\mathbf{T}}, \mathbf{T}^t, \mathrm{c}_{kl}) = \sum_{v\in P_k \cup P_l} \beta_{kl}(v, \mathrm{c}_{kl})\ \|\mathbf{T}_{k-l}^i(\mathbf{x}_v^0) - \mathbf{T}_{k-l}^j(\mathbf{x}_v^0)\|^2, \quad (8)$$

with $\beta_{kl}(v, \mathrm{c}_{kl})$ a uniform function over all vertices of the patch pair if $\mathrm{c}_{kl} = 1$, which encourages common rigid behavior of the two patches, and a non-uniform function encouraging more elasticity when $\mathrm{c}_{kl} = 0$:

$$\beta_{kl}(v, 0) \propto \exp(-\frac{b_{kl}(v)}{\eta \bar{D}}), \tag{9}$$

where $b_{kl}(v)$ is the distance between the vertex v and the border between P_k and P_l on the template, $\bar{D}$ is the average patch diameter and η is a global coefficient controlling the flexibility. The $\beta_{kl}(\cdot, 0)$ has larger values on the border between the patches, which allows more flexibility while enforcing continuity between the patches. The coefficients $\beta_{kl}(v, 0)$ are normalized such that $\sum_{P_k \cup P_l} \beta_{kl}(v, 0) = \sum_{P_k \cup P_l} \beta_{kl}(v, 1)$ in order to make both modes as competitive.

Mean model prior. In the absence of any prior, the mean pose is unconstrained and could theoretically have completely loose patches unrelated to each other. To avoid this and give the mean pose a plausible deformation, we consider the following a prior which expresses that the intrinsic mean pose should not significantly deviate from the original reference pose (represented by the identity transform **Id**):

$$p(\bar{\mathbf{T}}) \propto \exp(-d^2(\bar{\mathbf{T}}, \mathbf{Id})) \propto \exp\left(\sum_{(P_k, P_l) \in \mathcal{N}} \sum_{v \in P_k \cup P_l} \|\bar{\mathbf{T}}_k(\mathbf{x}_v^0) - \bar{\mathbf{T}}_l(\mathbf{x}_v^0)\|^2 \right), \tag{10}$$

3.4 Expectation-Maximization Inference

We apply Expectation-Maximization [10] to compute Maximum A Posteriori (MAP) estimates of the tracking and average shape parameters given noisy 3D measurements, using the joint probability described in (5) as described in [6]. The assignment variables K and rigidity coupling variables C are treated as latent variables, which we group by the name $Z = \{\mathrm{K}, \mathrm{C}\}$. For the purpose of clarity let us also rename all parameters to estimate as $\Theta = \{\bar{\mathbf{T}}, \mathbf{T}, \sigma\}$. Expectation-Maximization consists in iteratively maximizing the following auxiliary function Q given the knowledge of the previous parameter estimate Θ^m:

$$\Theta^{m+1} = \arg\max_{\Theta} Q(\Theta|\Theta^m) = \arg\max_{\Theta} \sum_{Z} p(Z|\mathbf{Y}, \Theta^m) \ln p(\mathbf{Y}, Z|\Theta). \tag{11}$$

The **E-Step** consists in computing the posterior distribution $p(Z|\mathbf{Y}, \Theta^m)$ of latent variables given observations and the previous estimate. It can be noted given the form of (5) that all latent variables are individually independent under this posterior according to the D-separation criterion [6], thus following the

factorization of the joint probability distribution:

$$p(\mathbf{Y}, Z|\Theta^m) = \prod_{t\in\mathcal{T}} \left(\prod_{(k,l)\in\mathcal{N}} p(\mathrm{c}_{kl}|\Theta^m) \prod_{o\in\mathcal{O}_t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m) \right), \tag{12}$$

$$\text{where } \; p(\mathrm{c}_{kl}|\Theta^m) = a \cdot \exp\left(- \sum_{v\in P_k\cup P_l} -d_{kl}^2(\mathbf{T}^{t,m}, \bar{\mathbf{T}}^m, \mathrm{c}_{kl}) \right) \tag{13}$$

$$\text{and } \; p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m) = b \cdot \mathcal{N}(\mathbf{y}_o^t \mid \mathbf{T}_{\mathrm{k}_o^t}^{t,m}(v), \sigma^{t,m}), \tag{14}$$

where a, b are normalization constants ensuring the respective distributions sum to 1, and v is the closest vertex on patch k. Equations (13) and (14) are the E-step updates that need to be computed at every iteration for every latent variable. (13) corresponds to a reevaluation of probabilities of rigid coupling between patches, based on the previous m-th estimates of temporal and mean poses. (14) corresponds to the probability assignment table of time t's observation o to each patch in the model. This corresponds to the soft matching term commonly found in EM-ICP methods [14].

The **M-Step** maximizes expression (11), which can be shown to factorize similarly to (5) and (12), in a sum of three maximizable independent groups of terms, leading to the following updates:

$$\mathbf{T}^{t,m+1} = \arg\min_{\mathbf{T}^t} \sum_{(k,l)\in\mathcal{N}} \sum_{\mathrm{c}_{kl}} p(\mathrm{c}_{kl}|\Theta^m)\; d_{kl}^2(\bar{\mathbf{T}}^m, \mathbf{T}^t, \mathrm{c}_{kl}) + \sum_{o\in\mathcal{O}_t} \sum_{\mathrm{k}_o^t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m) \|\mathbf{y}_o^t - \mathbf{T}_{\mathrm{k}_o^t}^t(\mathbf{x}_v^t)\|^2, \tag{15}$$

$$\sigma^{t,m+1\,2} = \frac{1}{3} \frac{\sum_{o\in\mathcal{O}_t} \sum_{\mathrm{k}_o^t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m)\; \|\mathbf{y}_o^t - \mathbf{T}_{\mathrm{k}_o^t}^{t,m+1}(\mathbf{x}_v^t)\|^2}{\sum_{o\in\mathcal{O}_t} \sum_{\mathrm{k}_o^t} p(\mathrm{k}_o^t|\mathbf{Y}, \Theta^m)}, \tag{16}$$

$$\bar{\mathbf{T}}^{m+1} = \arg\min_{\bar{\mathbf{T}}} \; d^2(\bar{\mathbf{T}}, \mathbf{Id}) + \sum_{t\in\mathcal{T}} \sum_{(k,l)\in\mathcal{N}} \sum_{\mathrm{c}_{kl}} p(\mathrm{c}_{kl}|\Theta^m) d_{kl}^2(\bar{\mathbf{T}}, \mathbf{T}^{t,m+1}, \mathrm{c}_{kl}). \tag{17}$$

Expression (15) corresponds to simultaneous updates of all patch transformations for a given time t, weighed by E-step probabilities. (16) updates the per-time frame noise parameter with an E-step weighed contribution of each observation. (17) computes the mean pose, accounting for all poses in the time sequence. Note that, for ease of resolution, we decouple the estimation of $\mathbf{T}^{t,m+1}$ and $\bar{\mathbf{T}}^{m+1}$, which is why (17) uses the result $\mathbf{T}^{t,m+1}$. We solve both systems with Gauss-Newton iterations, using a parametrization of the rigid transforms as a rotation matrix and translation.

4 Experiments

We evaluate the proposed generative model using 3D sequences reconstructed from real human performances captured by multiple view videos. We propose

two datasets, GoalKeeper and Dancer, which provide two different actions and clothing situations with high resolution inputs. These were processed by extracting visual hull reconstructions, and two neutral topology frames were selected to provide the template model after smoothing and simplifying the obtained mesh down to $5k$ vertices. Additionally, we also validate using two public datasets made available by the community. The Free [25] dataset consists of a photocoherent mesh sequence of a dancer with approximately $135k$ vertices per frame, exhibiting particularly fast and difficult dancing motion. The Marker dataset [19] provides another type of challenging situation with a two-person sequence of reconstructions, with martial art motions. It also provides markers on one of the persons which we will use for quantitative evaluation. For both these public sequences, we use the templates provided downsampled to $5k$ vertices.

In all visualizations, we render mesh poses by computing vertex position $\mathbf{x}_v^t$ at time t as a linear blend of positions $\mathbf{x}_k^t$ of expression (1), weighed by a set of Gaussian weights $\alpha_k(v)$ materializing the region of influence of patch P_k on the mesh. These weights are maximal at the center of mass of P_k and their sum over all non-zero patch influences are normalized to 1 for a given vertex v:

$$\mathbf{x}_v^t = \sum_k \alpha_k(v)\, \mathbf{x}_k^t \ . \tag{18}$$

We visualize the rigidity coupling probabilities over the surface with heat-colored probabilities, by diffusing this probability over vertices of influence of patch pairs to obtain a smooth rendering. We provide a supplemental video[3] with the processed results for these datasets.

4.1 Tracking Evaluation

We first evaluate the tracking performance of the algorithm. Full sequences may be processed but because of the motion of subjects in the sequence, all poses of the sequence cannot be initialized with a single static pose, as this would surely be susceptible to local minima. We thus process the four datasets using a sliding window strategy for $\mathcal{T}$, where processing starts with a single pose, then additional poses are introduced in the time window after the previous window converges. We provide tracking results with sliding window size 20 which corresponds to approximately one second of video. We show the resulting poses estimated by our algorithm on the four datasets in Fig. 4, Fig. 5a and Fig. 5b. Runtime is approximately 15 seconds per time step on a recent workstation and can be further improved.

We also provide a comparison with state of the art methods Liu *et al.* [19] and with a purely patch-based strategy [9], on the Marker dataset. We reproduce [9] results by neutralizing mean updates and rigid coupling updates from our method, which corresponds to removing these terms from the energy and closely mimics [9]. Note that [19] is a kinematic tracking strategy, where both subjects are rigged to a kinematic skeleton providing a strong, fixed and dataset specific

[3] `http://hal.inria.fr/hal-01016981`

rigidity prior. On the other hand, [9] only use patch rigidity and inter-patch elasticity priors, that are weaker than [19] and our method. The MARKER dataset provides sparse marker positions, at which we estimate geometric positional error with respect to the surface. To this purpose we match the closest vertex on the template model provided, and follow it with the different methods, computing geometric errors in position with respect to the corresponding marker's position in these frames. The average errors are shown in Table 1. We also provide a temporal error graph for our method and [9] in Fig. 2.

Table 1. Mean error and standard deviation over the sequence of the MARKER dataset.

method	mean error (mm)	standard deviation (mm)
no coupling, no mean pose [9]	55.11	48.02
our method	43.22	29.58
Liu *et al.* [19]	29.61	25.50

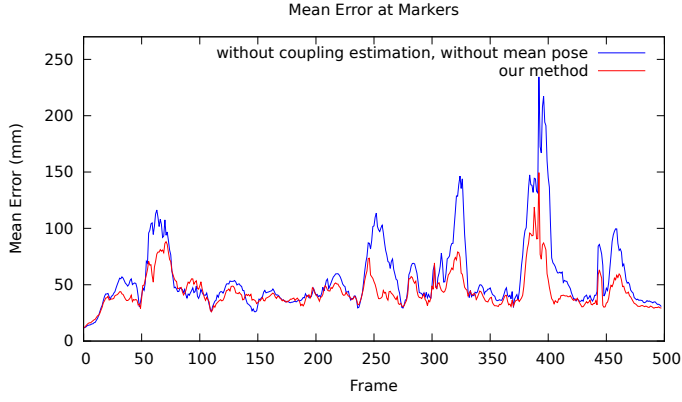
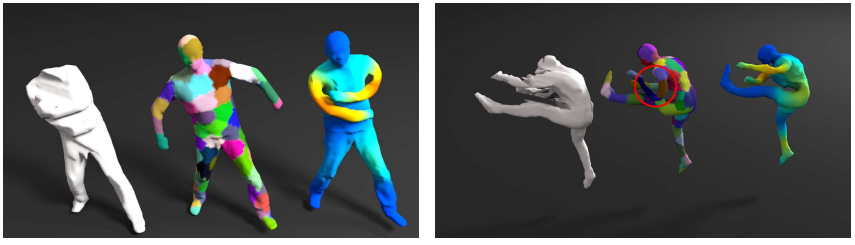


Fig. 2. Mean error for temporal evolution over MARKER dataset.

Table 1 shows our method achieves comparable tracking performance to state of the art surface tracking techniques. The slightly higher error with respect to [19] is not unexpected given that they use a stronger kinematic skeleton prior. Regarding [9], the graph and table show a small advantage in error for our method along the sequence, as well as a smaller variance of the error, showing the better constraining provided by our framework. The graph also shows significantly higher error values with [9] than with our method around frames 60, 250, 325, 390 and 460. These error peaks are imputable to difficult segments of the input sequence where [9] loses track of limbs (see Fig. 3a and Fig. 3b) while our method does not. The high error values around frame 390 are due to ambiguous

input meshes where the head of the second character (not seen in Fig. 3b) is out of the field of view. Around this frame, our method still outperforms [9] which misaligns an arm (see Fig. 3b). These results substantiate stronger robustness for our method over [9].

Regarding limitations, the model may fall into local minima when the noise level of inputs is too high similarly to all patch-based methods but this was not a strong limitation on the datasets. As the model favours rigidity and isometric surface deformations, the surface sometimes overfolds in non-rigid sections (as sometimes seen in video), which we will address in future work.



(a) Frame 325 (b) Frame 390

Fig. 3. Input mesh (left), tracked mesh with [9] (middle) and with our method (right).

4.2 Mean Pose and Rigidity Estimation

Fig. 5a shows tracking results with color coded rigidity coupling probabilities with sliding window size 20. The method accurately reports instantaneous rigidity deviation, such as when the subject folds his elbows or shoulders. Blue regions correspond to regions of the mesh that have no non-rigid distortion with respect to the estimated mean pose. Fig. 5b shows estimates of mean poses for full sequences, colored with the estimated rigidity coupling probabilities over full sequence (no sliding window). It can be noted that the method accurately reports where the most common deviations occur.

The supplemental video shows mean pose sequences for several sliding window sizes. We observe a temporal smoothing of the initial deformation: fast deformation is filtered out. This effect is stronger with wide windows. We interpret this phenomenon as follows: when the temporal window slides along the sequence, it produces a mean pose sequence analogous to the convolution of the estimated pose sequence with a gate function, with the same size as the window size. This process can be seen as a low-pass filtering of the sequence poses.

We also observe that the mean pose is not affected by global rigid motion of the shape (noticeable with the DANCER dataset). This is an expected consequence of using a pose distance that is invariant under global rigid transforms in (2).

Fig. 4. Tracking excerpts from the DANCER dataset. Colors code patches.

(a) Tracking Excerpts.

(b) Mean poses computed on full sequences.

Fig. 5. Tracking excerpts from GOALKEEPER, MARKER and FREE datasets. Best viewed in color. Please watch supplemental video for more visualizations.

5 Conclusions

We present a novel methodology for the analysis of complex object shapes in motion observed by multiple cameras. In particular, we propose a generative model of 3D temporal sequences using a probabilistic framework that simultaneously learns local surface rigidity probabilities and estimates a mean pose over temporal sequence. Hence, rigidity-based surface segmentation can be achieved using local deformation properties, while motion synthesis or surface alignment for compression or morphing applications can be achieved using a mean pose as a sequence keyframe or a cluster prototype.

Our model can also perform surface tracking with state of the art performance, and does not require a priori rigid (kinematic) structure, nor prior model learning from a database. Surface tracking and rigidity variable probabilities are obtained by solving an Expectation-Maximization inference problem which alternatively minimizes two nested fixed point iterations.

To our knowledge, this is the first model that achieves simultaneous estimation of mean pose, local rigidity, and surface tracking. Experimental results on real datasets show the numerous potential applications of the proposed framework for complex shape analysis of 3D sequences.

Acknowledgements

This work was funded by the Seventh Framework Programme EU project RE@CT (grant agreement no. 288369). It was also supported in part by the INRIA-JSPS Bilateral Program AYAME 146121400001 and the JSPS WAKATE B 26730089.

References

1. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. ACM Transactions on Graphics 27(3) (2008)
2. Allard, J., Ménier, C., Raffin, B., Boyer, E., Faure, F.: Grimage: Markerless 3d interactions. SIGGRAPH - Emerging Technologies (2007)
3. Anguelov, D., Srinivasan, P., Koller, D., Thrun, S., Rodgers, J., Davis, J.: Scape: Shape completion and animation of people. ACM Transactions on Graphics 24(3) (2005)
4. Ballan, L., Cortelazzo, G.M.: Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. 3DPVT (2008)
5. Baran, I., Popović, J.: Automatic rigging and animation of 3D characters. ACM Transactions on Graphics 26(3), 72:1–72:8 (2007)
6. Bishop, C.M.: Pattern Recognition and Machine Learning (Information Science and Statistics). Springer-Verlag New York, Inc., Secaucus, NJ, USA (2006)
7. Botsch, M., Pauly, M., Wicke, M., Gross, M.: Adaptive space deformations based on rigid cells. Comput. Graph. Forum 26(3), 339–347 (2007)
8. Bronstein, A.M., Bronstein, M.M., Kimmel, R.: Efficient computation of isometry-invariant distances between surfaces. SIAM Journal on Scientific Computing 28 (2006)

9. Cagniart, C., Boyer, E., Ilic, S.: Probabilistic deformable surface tracking from multiple videos. ECCV (2010)
10. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the Royal Statistical Society, series B (1977)
11. Franco, J., Boyer, E.: Learning temporally consistent rigidities. CVPR (2011)
12. Franco, J., Menier, C., Boyer, E., Raffin, B.: A distributed approach for real-time 3d modeling. CVPR Workshop (2004)
13. Fréchet, M.: Les éléments aléatoires de nature quelconque dans un espace distancié. Annales de l'institut Henri Poincaré 10, 215–310 (1948)
14. Granger, S., Pennec, X.: Multi-scale EM-ICP: A fast and robust approach for surface registration. ECCV 4, 6973 (2002)
15. Hasler, N., Ackermann, H., Rosenhahn, B., Thormählen, T., Seidel, H.P.: Multilinear pose and body shape estimation of dressed subjects from image sets. CVPR (2010)
16. Hufnagel, H., Pennec, X., Ehrhardt, J., Ayache, N., Handel, H.: Generation of a Statistical Shape Model with Probabilistic Point Correspondences and EM-ICP. IJCAR 2(5) (2008)
17. J.Gall, C.Stoll, de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. CVPR (2009)
18. Kendall, D.: Shape manifolds, procrustean metrics, and complex projective spaces. Bulletin of the London Mathematical Society 16(2), 81–121 (1984)
19. Liu, Y., Gall, J., Stoll, C., Dai, Q., Seidel, H.P., Theobalt, C.: Markerless motion capture of multiple characters using multi-view image segmentation. PAMI (2013)
20. Matsuyama, T., Nobuhara, S., Takai, T., Tung, T.: 3d video and its applications. Springer (2012)
21. Mémoli, F., Sapiro, G.: Comparing Point Clouds. SGP (2004)
22. Ovsjanikov, M., Mrigot, Q., Mmoli, F., Guibas, L.J.: One point isometric matching with the heat kernel. Comput. Graph. Forum 29(5) (2010)
23. Sahilliogu, Y., Yemez, Y.: 3D Shape correspondence by isometry-driven greedy optimization. CVPR (2010)
24. Starck, J., Hilton, A.: Model-based multiple view reconstruction of people. ICCV (2003)
25. Starck, J., Hilton, A.: Spherical matching for temporal correspondence of non-rigid surfaces. ICCV (2005)
26. Straka, M., Hauswiesner, S., Ruether, M., Bischof, H.: Simultaneous shape and pose adaption of articulated models using linear optimization (2012)
27. Sumner, R.W., Popović, J.: Deformation Transfer for Triangle Meshes. ACM Transactions on Graphics 23(3) (2004)
28. Tung, T., T.Matsuyama: Topology dictionary for 3d video understanding. PAMI 34(8), 1645–1647 (2012)
29. Tung, T., T.Matsuyama: Intrinsic characterization of dynamic surfaces. CVPR (2013)
30. Vlasic, D., Baran, I., Matusik, W., Popovic, J.: Articulated mesh animation from multi-view silhouettes. ACM Transactions on Graphics 27(3) (2008)
31. Windheuser, T., Schlickewei, U., Schmidt, F., Cremers, D.: Geometrically consistent elastic matching of 3d shapes: A linear programming solution. ICCV (2011)
32. Zeng, Y., Wang, C., Gu, X., Samaras, D., Paragios, N.: A Generic Deformation Model for Dense Non-Rigid Surface Registration: a Higher-Order MRF-based Approach. ICCV (2013)