An Efficient Volumetric Framework for Shape Tracking

Benjamin Allain Jean-Sébastien Franco Edmond Boyer

Inria Grenoble Rhône-Alpes - LJK
Grenoble Universities, France
firstname.lastname@inria.fr

Abstract

Recovering 3D shape motion using visual information is an important problem with many applications in computer vision and computer graphics, among other domains. Most existing approaches rely on surface-based strategies, where surface models are fit to visual surface observations. While numerically plausible, this paradigm ignores the fact that the observed surfaces often delimit volumetric shapes, for which deformations are constrained by the volume inside the shape. Consequently, surface-based strategies can fail when the observations define several feasible surfaces, whereas volumetric considerations are more restrictive with respect to the admissible solutions. In this work, we investigate a novel volumetric shape parametrization to track shapes over temporal sequences. In constrast to Eulerian grid discretizations of the observation space, such as voxels, we consider general shape tesselations yielding more convenient cell decompositions, in particular the Centroidal Voronoi Tesselation. With this shape representation, we devise a tracking method that exploits volumetric information, both for the data term evaluating observation conformity, and for expressing deformation constraints that enforce prior assumptions on motion. Experiments on several datasets demonstrate similar or improved precisions over state-of-the-art methods, as well as improved robustness, a critical issue when tracking sequentially over time frames.

1. Introduction

The capture of shapes and their evolutions has been a very active research topic for the last decade, motivated by many applications for which dynamic shape models are useful. This ability is of interest for several fields of activity such as computer-assisted design, virtual reality, entertainment, medical imaging, gesture and sports analysis. Ever since the initial promises of free viewpoint video [9], many models of shape capture have been explored. Initially examined as a per-time reconstruction problem, e.g. [24, 14], temporal integration and tracking of the shape in the time domain have then been considered, e.g. [11, 3]. In any case, however, surface-based models, such as meshes, have been largely dominant to represent and track shapes. This is due to several factors, primarily to the fact that visual observations generally lie on the shape surface, but also to the popularity of surface-based representations in the vision and graphics communities and the availability of efficient tools to manipulate them. Yet it has been observed that certain forms of volume-preserving deformations may be beneficial to model shape deformations in graphics applications such as [1, 5], or to enforce volumetric constraints, nevertheless based on surface tesselations, in dynamic shape capture [10].

While the idea has led to interesting preliminary results, a full volumetric treatment of dynamic shape capture is still to be investigated and its benefits evaluated. Among the expected advantages of this approach are its ability to express volume conservation as well as its ability to enforce local volumetric deformation constraints. In this paper, we address this problem with a twofold contribution: we first propose a dedicated volumetric deformation model based on Centroidal Voronoi Tesselations (CVT) [13], which integrates the latest advances of recent tracking models, and second we propose an evaluation of the method based on a hybrid multi-camera and marker-based capture dataset [21].

1.1. Previous Work

A large set of techniques exist to capture moving shapes as a time independent sequence of meshes representing the object's surface [24, 14]. For this process, many volumetric parameterizations have also been devised, based on regular or hierarchical Eulerian grid discretizations [30, 20], although they are mainly dedicated to single time occupancy representation. Some approaches have taken these representations a step further, by examining short term motion characteristics of the shape using regular volume grids [33, 17, 32], yet they do not retrieve long term motion information of the sequence, nor do they embed specific motion models in the volume.

Various methods attempt leveraging time consistency to retrieve temporally coherent shape models, in the vast majority of cases manipulating a surface model. While in some cases this process is purely data-driven, by aligning surfaces across frames using sparse matching and stereo refinement [29], in most cases a deformation prior is used to drive the method toward the solution within a plausible state space. In its weakest form and without further assumptions, pure spatio-temporal continuity of the observed surface can be used [16]. At the other end of the spectrum a full kinematic rigging of a template model can be assumed, where the surface is expressed from kinematic parameters using e.g. the linear blend skinning deformation framework [23] popularized for 3D animation in the computer graphics community. These parameters can then be estimated for best fitting the model reprojections to image and silhouette data [34, 3, 18, 21]. For tracking more general subjects and situations, more generic surface deformation frameworks have been explored to bypass the rigging stage and allow for more general non-rigid motion components. Central to these methods is the idea of enforcing a cohesive behavior of the surface, such as locally rigid behavior [15], Laplacian deformation [11, 10, 6], inextensibility [25], or elasticity between piecewise-rigid surface patches [7, 6].

Among the existing surface capture methods, only a handful use volumetric representations. Some methods have proposed reparameterizing temporally aligned sequences using a volumetric cage embedding [26, 31] inspired from the animation community [27, 19]. However, no volume deformation model strong enough to solve the full tracking problem has yet emerged from these works. Among the methods successfully using volume preserving constraints, most use a Delaunay tetrahedrization of reconstructed template surface points [11, 10, 6] to enforce as-rigid-as-possible or Laplacian deformation constraints common to 3D animation techniques [1, 28]. It can be noted that the proposed decomposition is not fully volumetric as it only involves tesselating surfaces. In contrast, we propose a fully volumetric treatment of the problem, using an intrinsically volumetric tesselation, deformation model and data terms for rewarding volume alignment.

1.2. Approach Overview

We formulate the tracking problem as the MAP estimation of multiple poses of a given geometric template model, non-rigidly adjusted to a set of temporally inconsistent shape measurements. In multi-view camera systems, these measurements typically take the form of time independent 3D mesh reconstructions obtained from a visual hull or multi-view stereo method, which is what we assume here. To efficiently make use of volumetric information, we need to express volume conservation and overlapping constraints from the template to the observed shape volumes. For representational and computational efficiency, we thus need a proper discretization of the interior of the shape. While uniformly located in the volume, regular grids are inherently anisotropic and biased toward the axis of the template basis. Furthermore, their intersection with the object surface yields boundary voxels of irregular shape (Fig. 1(a)). On the other hand, the Constrained Delaunay tetrahedrization of the boundary vertices, previously used in [11, 10, 6], yields a set of highly non-uniform tetrahedra spanning the whole interior of the volume, whose cardinality is not controlled but imposed by the surface discretization (Fig. 1(b)). Taking instead the Voronoi diagram of a uniform set of samples of the interior volume decorrelates the cardinality of the decomposition from the geometry, but still yields cells of irregular shape (Fig. 1(c)). Rejection sampling may statistically impose additional regularity, but this would only come with asymptotic guaranties attainable at high computational cost. We therefore propose to use CVT (Fig. 1(d)), informally a Voronoi tesselation where the samples are iteratively repositioned to coincide with the center of mass of their cell, which achieves the desired properties [13]: isotropy, rotational invariance, uniform cells of compact and regular form factor, regular intersection of boundary cells and surface, independent cardinality and practical computation.

After introducing how to define and compute CVTs in the context of our approach (§2), we show how this discretization can be used to define adequate volumetric deformation (§3) and observation (§4) models in the form of Bayesian prior and likelihoods. The MAP estimation proposed on this basis in §5 is evaluated in §6.

2. Centroidal Voronoi Tessellation (CVT)

Definitions. To tesselate the shape, we manipulate Voronoi diagrams that are restricted, or clipped, to its inner volume. More formally, let $\mathcal{S}$ be a set of 3D point samples of a volumetric domain Ω, either the template to be fitted or the observed shape for our purposes. The *Clipped Voronoi diagram* of $\mathcal{S}$ in Ω is defined as the intersection of the Voronoi diagram of $\mathcal{S}$ in $\mathbb{R}^3$ with the domain Ω. Thus the Voronoi cell Ω_s of a sample s is the set of points from Ω that are closer to s than to any other sample:

$$\Omega_s = \{\mathbf{x} \in \Omega \mid \forall s' \in \mathcal{S} \setminus \{s\} \;\; \|\mathbf{x} - \mathbf{x}_s\| < \|\mathbf{x} - \mathbf{x}_{s'}\|\}, \quad (1)$$

where cells Ω_s are mutually exclusive and define a partition of Ω:

$$\bigcup_{s \in \mathcal{S}} \overline{\Omega_s} = \overline{\Omega} \, , \quad (2)$$

where $\overline{\Omega_s}$ and $\overline{\Omega}$ denote topolgical set closures. If the border $\partial\Omega$ of Ω is a polyhedral surface, then each cell also has a polyhedral border.

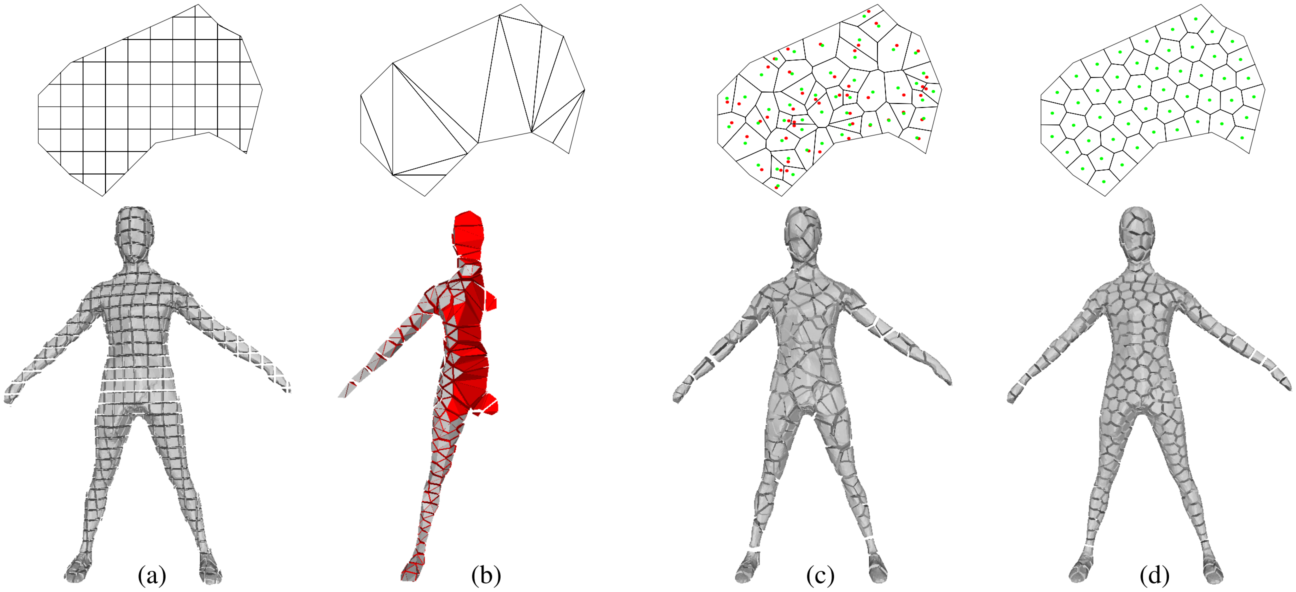

Figure 1. Possible volumetric decompositions of the template and observed shapes. (top) 2D schematic view. (bottom) 3D decomposition example. (a) Voxels on a regular grid. (b) A sliced Constrained Delaunay tetrahedrization showing the elongated inner tetrahedra generated. (c) Voronoi cells with random centroids shown in red, center of mass of each cell in green. (d) Centroidal Voronoi tesselation cells, where the center of mass and cell centroid coincide.

A *Centroidal Voronoi tessellation* is a clipped Voronoi tessellation of Ω for which each sample s is the center of mass of its (clipped) Voronoi cell Ω_s. CVT cells are of regular size and shapes, and also define a regular connectivity of the samples set, two samples being connected if and only if their respective CVT cells share a face. This connectivity thus encodes the shape volume and topology, a property we will use in the following sections.

Computing a CVT. It has been proven [13] that local minima of the energy

$$E(\mathcal{S}) = \sum_{s \in \mathcal{S}} \int_{\mathbf{x} \in \Omega_s} \|\mathbf{x} - \mathbf{x}_s^t\|^2 dV \tag{3}$$

define CVTs on Ω. Thus a CVT can be obtained by iteratively estimating the sample locations that minimize (3), with a quasi-Newton method such as the L-BFGS algorithm [22], for a sample population of desired size and uniform initial position.

3. Deformation Model

3.1. Principle

Once a regular, anisotropic volumetric decomposition of the template shape is obtained, we can use it as a fundamental building block to build a volumetric deformation model of the shape, which will constrain the estimation. Botsch *et al.* [5] show that a non-linear elastic deformation energy can be devised using small volumetric deformations, typically voxels. While such a reasoning could be advantageously transposed to the CVT discretization proposed, eliminating the grid orientation bias, doing so comes at a high computational cost. Cagniart et al. [7] show that the complexity of the deformation model is best decorrelated from the geometry itself, in their case by using rigid surface patches in lieu of the original surface vertices. Recent works have shown a way to improve the quality and temporal stability using a similar surface decomposition [2], by inferring a mean pose and sequence rigidity behaviour.

We extend the latter ideas to the case of a complete volumetric treatment of the deformation problem. In so doing, we cluster together groups of CVT cells in rigid volume patches P_k using a k-medoids algorithm. Note that such patches can lie either on the surface or completely inside the template shape's volume, which is of particular interest to express non-rigid deformation of the model while preserving the local volume and averting over-compression or dilation. We associate to each patch a rigid transform $\mathbf{T}_k^t \in SE(3)$ at every time t. Each position $\mathbf{x}_{k,q}$ of a mesh vertex or inner sample is indiscriminately labeled as a point q. Its position can be written as a transformed version of its template position $\mathbf{x}_q^0$ as follows, once the rigid transform of its patch is applied:

$$\mathbf{x}_{k,q} = \mathbf{T}_k(\mathbf{x}_q^0). \tag{4}$$

This makes it possible to define a *pose* of the shape as the the set of patch transforms $\mathbf{T} = \{\mathbf{T}_k\}_{k \in \mathcal{K}}$, which expresses a given volumetric shape deformation.

3.2. Formulation

To prevent patch poses of the shape from being arbitrary, we constrain the shape to be close to a sequence rest pose $\bar{\mathbf{T}}$ and to follow constant rigidity characteristics C over the sequence. These rigid characteristics are defined for neighboring patch pairs in the volume (P_k, P_l), as a binary valued property c_{kl}, whose value in $\{0,1\}$ reflects wether the relative motion between patches P_k and P_l is respectively articulated or rigid. To define the rest pose $\bar{\mathbf{T}}$ we rely on the following measure [5, 2] of the relative deformation energy between two arbitrary poses $\mathbf{T}^i$ and $\mathbf{T}^j$ of the template shape, given a rigidity configuration C:

$$\mathcal{E}(\mathbf{T}^i, \mathbf{T}^j|\mathrm{C}) = \sum_{(P_k,P_l)\in\mathcal{N}} \mathcal{E}_{kl}(\mathbf{T}^i, \mathbf{T}^j|\mathrm{c}_{kl}), \quad \text{with} \tag{5}$$

$$\mathcal{E}_{kl}(\mathbf{T}^i,\mathbf{T}^j|\mathrm{c}_{kl}) = \sum_{q\in P_k\cup P_l} \beta_{kl}(q,\mathrm{c}_{kl}) \|\mathbf{T}^i_{k-l}(\mathbf{x}^0_q) - \mathbf{T}^j_{k-l}(\mathbf{x}^0_q)\|^2,$$

where $\mathbf{T}^i_{k-l} = \mathbf{T}^{i\;-1}_l \circ \mathbf{T}^i_k$ 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 energy measures the rigid deviation from pose i to j of every neighboring patch pair, as the sum over each of the samples s of the pair, of the discrepancy in relative positions of the vertex as displaced by P_k on one hand, and P_l on the other. If the two patches are rigidly linked ($\mathrm{c}_{kl} = 1$), then the discrepancy of all samples of the pair should be equally penalized, therefore $\beta_{kl}(s, 1)$ is chosen to be constant over all samples s of the pair. On the other hand, if the patch pair is articulated ($\mathrm{c}_{kl} = 0$), only samples that lie near the boundary between the two patch volumes should be penalized for deviating relative positions: those samples materialize the locus of the inter-patch articulation, whereas samples that aren't close the inter-patch boundary can move more freely. We express this using $\beta_{kl}(q, 0) \propto \exp(-\frac{b_{kl}(s)}{\eta\bar{D}})$ where $b_{kl}(s)$ is the distance between the sample s and the boundary 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.

3.3. Resulting Pose Likelihoods

The relative pose energy described in (5) makes it possible to express the expected behavior of the estimated models as a prior and likelihood over the poses:

$$p(\bar{\mathbf{T}}) \;\propto\; \exp(-\mathcal{E}(\bar{\mathbf{T}}, \mathbf{Id})), \tag{6}$$

$$p(\mathbf{T}^t \mid \bar{\mathbf{T}}, \mathrm{C}) \;\propto\; \exp\left(-\mathcal{E}(\mathbf{T}^t, \bar{\mathbf{T}}|\mathrm{C})\right). \tag{7}$$

$p(\bar{\mathbf{T}})$ is the prior over the rest pose, which should minimize the relative displacement energy to the default template pose (transformed by identity $\mathbf{Id}$). This terms ensures minimal cohesion of the volume patches of the rest pose model, as it enforces mutual patch elasticity.

$p(\mathbf{T}^t|\bar{\mathbf{T}}, \mathrm{C})$ is the likelihood of a given tracked pose at time t, which should minimize the relative displacement energy with respect to the sequence rest pose $\bar{\mathbf{T}}$ given a current rigidity state C. This ensures the inter-patch cohesion of pose $\mathbf{T}^t$ as well as a general proximity to the rest pose, which stabilizes the resulting pose estimates. In turn the rest pose will be simultaneously estimated as the pose which minimizes the relative deformation energy to all poses in the sequence.

4. Observation Model

4.1. Probabilistic Shape Fitting

The observed shape Ω^t at time t is described by the point cloud $\mathbf{Y}^t = \{\mathbf{y}^t_o\}_{o\in\mathcal{O}_t}$. To describe how a deformed template can explain the observed shape, we propose a generative data term following EM-ICP, expressing how a given deformed template point predicts the position of an observed shape point o. A set of cluster association variables k^t_o is therefore instantiated for every observed point in time, indicating which cluster generates this point. For simplicity, each observation o is associated to its cluster k^t_o via the best candidate q of cluster k^t_o. The best candidate is chosen as the closest compatible sample in the cluster during iterative resolution. We consider that each cluster P_k generates observations perturbed by a Gaussian noise with isotropic variance σ^2:

$$p(\mathbf{y}^t_o \mid \mathrm{k}^t_o, \mathbf{T}^t_k, \sigma) \;=\; \mathcal{N}(\mathbf{y}^t_o \mid \mathbf{T}^t_{\mathrm{k}}(\mathbf{x}^0_q), \sigma). \tag{8}$$

Note that o indiscriminately refers to surface or volume sample points of the observed shape, as the principles we describe here apply to both, with the restriction that observed surface points only associate to surface template points, and volume samples are associated to volume samples of the template. As often proposed in ICP methods, we additionally filter associations using a compatibility test, described in the following sections. The compatibility test is specific to the nature (surface or volume) of the observed point and is detailed in the next paragraphs. If there is no compatible candidate in the cluster, then we set conditional probability density (8) to zero. We deal with outliers by introducing an outlier class among values of k, which generate observations with a uniform probability density over the scene.

4.2. Compatibility Tests

Compatibility tests are useful for pruning the association graph for obvious mismatches that would perturb and otherwise slow down convergence. We use two compatibilty tests respectively designed for surface fitting and volumetric fitting.

Surface Observations. While surface points may be matched based on position only, obvious orientation incompatibilites can be filtered by detecting large discrepancies between the normal of the deformed template candidate point v, and the normal of surface observation vertex o:

$$\vec{\mathbf{n}}_o^t \cdot \mathbf{R}_k^t(\vec{\mathbf{n}}_v^0) \geq \cos(\theta_{\max}), \tag{9}$$

where $\vec{\mathbf{n}}_o^t$ is the surface normal of observation o, $\vec{\mathbf{n}}_v^0$ is the surface normal of the template at vertex v, $\mathbf{R}_k^t$ is the rotation component of $\mathbf{T}_k^t$, and $\theta_{\max}$ is an arbitrary threshold.

Volume Observations. We introduce a compatibility test specific to volumetric fitting, by assuming that the distance of inner surface points to the shape's surface remains approximately constant under deformation. Let us define the distance between an inner shape point x and the shape's surface by:

$$d(x, \partial\Omega) = \min_{p \in \partial\Omega} d(x, p). \tag{10}$$

In our observation model, this hypothesis can be leveraged by the following compatibility test: a volumetric observation o can be associated to a template point s only if

$$d(\mathbf{x}_s^0, \partial\Omega^0) = d(\mathbf{y}_o^t, \partial\Omega^t). \tag{11}$$

To account for small deviations to this assumption, which might occur under e.g. slight compression or dilation of the perceived shape, we relax the equality constraint up to a precision ϵ, where ϵ accounts for the distance-to-surface inconsistency caused by the discrete sampling of the template. Using the triangular inequality, it can be shown that this error is bounded by the maximum cell radius over the set of the template's CVT cells. This leads to the following compatibility test:

$$d(\mathbf{y}_o^t, \partial\Omega^t) - \epsilon \leq d(\mathbf{x}_s^0, \partial\Omega^0) \leq d(\mathbf{y}_o^t, \partial\Omega^t) + \epsilon \tag{12}$$

For the particular case of silhouette-base observed shapes, it can be noted that reconstruction algorithms based on the visual hull inherently provides inflated estimates of the true shape. This phenomenon results in an overestimation of the distance to the surface when computed on the reconstructed shape. Hence, we only impose a volumetric inclusion constraint instead of complete depth correspondance, i.e. we only keep the right inequality from expression (12) in this case:

$$d(\mathbf{x}_s^0, \partial\Omega^0) \leq d(\mathbf{y}_o^t, \partial\Omega^t) + \epsilon. \tag{13}$$

Contrary to the surface compatibility test, this test does not depend on pose parameters $\mathbf{T}$, consequently it is robust to convergence failures of inference algorithms.

5. Inference

The model proposed with (6), (7) and (8), defines a joint likelihood over the rest pose, the rigidity configuration, all temporal poses, the observed points and their selection variables, and prediction noise σ:

$$p(\bar{\mathbf{T}}) \prod_{t \in \mathcal{T}} \left(p(\mathbf{T}^t | \bar{\mathbf{T}}, \mathrm{C}) \prod_{o \in \mathcal{O}_t} p(\mathbf{y}_o^t \mid \mathrm{k}_o^t, \mathbf{T}^t, \sigma^t) \right), \tag{14}$$

It can be shown that this likelihood can be maximized using an Expectation Maximization algorithm [2, 12, 4], yielding maximum a posteriori estimates of the pose parameters $\bar{\mathbf{T}}$, $\mathbf{T}$ and prediction noise σ. This results in an algorithm iterating between two steps.

Intuitively, The **E-Step** computes all observation cluster assignment probabilities over K, based on the distance to the predicted template positions under the currently estimated poses. Compatibility rules are applied at this stage. Probabilities over inter-cluster rigid links C are also estimated based on the current deformation energy of the poses. The **M-Step** updates the rest pose $\bar{\mathbf{T}}$, all poses $\mathbf{T}$, and prediction noise σ, using the assignment and rigid link probabilities to weigh individual observation contributions to each cluster transform estimate.

6. Experiments

6.1. Datasets

We validate our framework using four synchronized and calibrated multiple-camera datasets, labeled GOALKEEPER-13, DANCER [2], MARKER [21], and the newly proposed BALLET, whose content reflect a wide variety of shape tracking situations. DANCER is a long sequence (1362 frames, 2048 × 2048 resolution, 48 viewpoints) showing slow and medium speed dance moves, and thus offers good opportunity to verify the tracking stability. GOALKEEPER-13 (2048×2048, 150 frames, 48 viewpoints) illustrates a specific soccer goalie plunging move, of particular interest when the goalie is on the floor, where the reconstruction data is of poor quality due to grazing camera angle and challenges the tracking performance. Both previous sequences otherwise have very high quality and detailed reconstructions. BALLET is a more challenging full HD (1920 × 1080) sequence we have acquired with fewer cameras (9 viewpoints and 500 frames) and thus coarser reconstructions, consisting in a number of ballet moves with various levels of difficulty, including fast moves, spinning and crossing legs. MARKER (1296 × 972, 500 frames, 12 viewpoints) is a sequence with two actors performing karate moves, illustrating the robustness to several subjects, and which was captured simultaneously with a set of sparse markers offering a reference and comparison basis

method	std. dev. (L)	
	MARKER	BALLET
Cagniart *et al.* 2010 [8]	3.85	1.22
Allain *et al.* 2014 [2]	4.32	1.20
our method	**2.24**	**0.95**

Table 1. Variation of the estimated volume over the sequence for MARKER and BALLET datasets.

with [21]. The reconstructions are of coarser quality due to relatively noisy inputs and occasional jumps where actor heads get clipped.

6.2. Experimental Protocol

We first select a template among the better frames with correct reconstruction topology, then compute a CVT using §2 with 5'000 samples per person (10'000 for the two-person MARKER sequence) and 250 clusters per person (500 for MARKER), as illustrated in Fig. 3. Each shape reconstruction is obtained from a silhouette-based reconstruction algorithm [14] and CVTs are also extracted (1 minute/frame). We then apply the algorithm described using a sliding window strategy over a 10 frame window, where the rest position is computed for each time slice to locally stabilize the estimation. The sequences are initially solved for frames in the vicinity of the template pose, then the solution is classically propagated to future windows as initialization. Convergence has been achieved for all frames, typically in a few hundred iterations, with a convergence time per frame of the order of a minute to a few minutes. The provided supplemental video[1] illustrates the results obtained with these sequences.

6.3. Quantitive Analysis

Volume Stability. We verify here the assertion that the volumetric parameterization of the tracking produces poses with stable volumes. As we use silhouette based reconstructions, it is not relevant to compare the estimated volumes with the observed shape volumes. Instead, we compute the standard deviation to this volume, in Table 1 and provide a comparison of these results with best runs of two state of the art methods [8, 2]. This comparison supports the initial intuition of volumetric stability in the sequence, as the standard deviation of the estimated volumes is significantly lower for our method.

Silhouette reprojection error. We evaluate the silhouette reprojection error as the symmetric pixel difference between the reprojected and the silhouette projection of the reconstructions used as observed input shapes. We then express this value as a percentage of the area of the silhouette region in each view. Table 2 shows favorable comparisons to state of the art methods [8, 2]. In particular, the mean error and maximum error achieved by our method over the sequences is significantly lower, and exhibits lower variance. Additionally we test the influence of the volumetric data term by comparing the results with a run where it is disabled (surface data-term only), all other parameters being equal. Interestingly, the method still achieves better mean error than state of the art, but with less stable behavior.

method	mean	stddev.	median	max
Cagniart *et al.* [8]	5.74	1.88	5.48	15.20
Allain *et al.* [2]	5.81	1.70	5.61	13.77
Ours, no vol. fitting	4.62	1.94	**4.28**	17.20
Ours	**4.56**	**1.21**	4.43	**11.00**

Table 2. Mean and statistics of silhouette reprojection error over BALLET dataset, expressed in percentage of silhouette area.

method	mean	std. dev.
Cagniart *et al.* [8]	55.11	48.02
Allain *et al.* [2]	43.22	29.58
Proposed, no surface fitting	42.60	29.32
Proposed	**38.41**	**26.70**
Liu *et al.* [21]	**29.61**	**25.50**

Table 3. Mean marker error and std.dev. (mm), MARKER dataset. Note that Liu *et al.* assume a rigged skeleton is associated to the template, a stronger and more restrictive assumption.

Marker reference error. We use the MARKER sequence provided by Liu *et al.* [21] to sparsely compare the output quality of our method against state of the art methods. This comparison is illustrated in Table 3 and plotted against time in the sequence in Fig. 2. Again we illustrate turning off the surface data term, in which case the estimation is slightly worse. The method proposed performs consistently better than comparable surface-based state of the art. Note that Liu *et al.* fully rig a skeleton to the template, which provides slightly better mean results than ours thanks to the stronger assumption. On the other hand, our method is generic and can be applied to arbitrary objects.

6.4. Qualitative Assessment

To illustrate the benefits of the approach, in particular where the improvements can be seen, we provide excerpts of the datasets (see supplemental video for further examples). Fig. 4 shows the improvements of the method over surface-based methods [8, 2] in poses of strong contortion, such as an elbow or knee bending gesture. Because of their use of the elastic energy on the surface, these methods tend to dilute error compensation over a smooth and extended location of the folds, yielding curvy elbow and knee shapes in the tracking. A usual side effect here is the local decrease of

[1] Video available at https://hal.inria.fr/hal-01141207

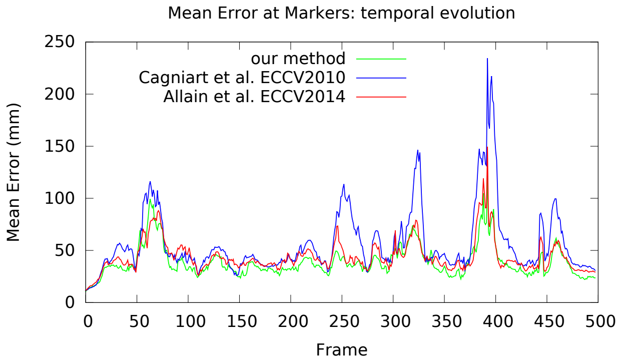

Figure 2. Mean error for temporal evolution over MARKER dataset.

the shape volume in the vicinity of the fold. In contrast, our method being volumetrically constrained, it penalizes such local volume changes and prefers to focus the bending energy on fewer volumetric patch articulations, yielding more consistent and accurate pose estimates. The GOALKEEPER-13 dataset illustrates the increased robustness in the presence of highly perturbed reconstructions thanks to the volumetric constraints, where other methods yield random results. The reconstructed visual hull input is very ambiguous on the shown frame because of the presence of concavities and the strong topology mismatch creates errors for surface-based methods.

7. Discussion

We have presented a novel volumetric approach to shape tracking based on CVT volume decomposition. The approach leverages CVT desirable properties to build suitable volumetric deformable constraints, while formulating a discrete volume assignment scheme as data term through the uniform cell centroid coverage of the volume. Currently, the volumetric clustering proposed for volumes yields uniform sizes over the entire template shape, which can be a limitation for parts that are thinner than the cluster size, such as arms. We will address this in future work with adaptive cluster densities, ensuring the volumetric prior is equally efficient regardless of thickness. Numerical analysis nevertheless shows significant improvement over state of the art tracking methods, both in terms of tracking error over the surface and silhouette reprojection. The framework is also shown to conform to initial intuition in being more stable in terms of the errors and volume measures of the fitted template shapes. We believe the approach paves the way for proper use of volumetric priors in any shape tracking framework.

Acknowledgements

This work was funded by the Seventh Framework Programme EU project RE@CT (grant agreement no. 288369). We thank Li Wang and Franck Hetroy for providing the CVT computation code.

References

[1] M. Alexa, D. Cohen-Or, and D. Levin. As-rigid-as-possible shape interpolation. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '00, pages 157–164, 2000. 1, 2

[2] B. Allain, J.-S. Franco, E. Boyer, and T. Tung. On mean pose and variability of 3d deformable models. In *ECCV*, 2014. 3, 4, 5, 6, 8

[3] L. Ballan and G. M. Cortelazzo. Marker-less motion capture of skinned models in a four camera set-up using optical flow and silhouettes. In *3DPVT*, 2008. 1, 2

[4] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 5

[5] M. Botsch, M. Pauly, M. Wicke, and M. Gross. Adaptive space deformations based on rigid cells. *Comput. Graph. Forum*, 26(3):339–347, 2007. 1, 3, 4

[6] C. Budd and A. Hilton. Temporal alignment of 3d video sequences using shape and appearance. *Conference on Visual Media Production*, pages 114–122, 2010. 2

[7] C. Cagniart, E. Boyer, and S. Ilic. Free-from mesh tracking: a patch-based approach. In *CVPR*, 2010. 2, 3

[8] C. Cagniart, E. Boyer, and S. Ilic. Probabilistic deformable surface tracking from multiple videos. In *ECCV*, 2010. 6, 8

[9] J. Carranza, C. Theobalt, M. A. Magnor, and H.-P. Seidel. Free-viewpoint video of human actors. In *ACM SIGGRAPH 2003 Papers*, pages 569–577, 2003. 1

[10] E. de Aguiar, C. Stoll, C. Theobalt, N. Ahmed, H.-P. Seidel, and S. Thrun. Performance capture from sparse multi-view video. *ACM Transactions on Graphics*, 27(3), 2008. 1, 2

[11] E. de Aguiar, C. Theobalt, C. Stoll, and H.-P. Seidel. Marker-less deformable mesh tracking for human shape and motion capture. In *CVPR*, 2007. 1, 2

[12] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society, series B*, 1977. 5

[13] Q. Du, V. Faber, and M. Gunzburger. Centroidal voronoi tessellations: Applications and algorithms. *SIAM review*, 41:637–676, 1999. 1, 2, 3

[14] J.-S. Franco and E. Boyer. Exact polyhedral visual hulls. In *British Machine Vision Conference (BMVC'03)*, volume 1, pages 329–338, Norwich, United Kingdom, Sept. 2003. 1, 6

[15] Y. Furukawa and J. Ponce. Dense 3d motion capture from synchronized video streams. *CVPR*, 2008. 2

[16] B. Goldlücke and M. Magnor. Space-time isosurface evolution for temporally coherent 3D reconstruction. In *CVPR*, 2004. 2

[17] L. Guan, J.-S. Franco, E. Boyer, and M. Pollefeys. Probabilistic 3d occupancy flow with latent silhouette cues. In *CVPR*, 2010. 1

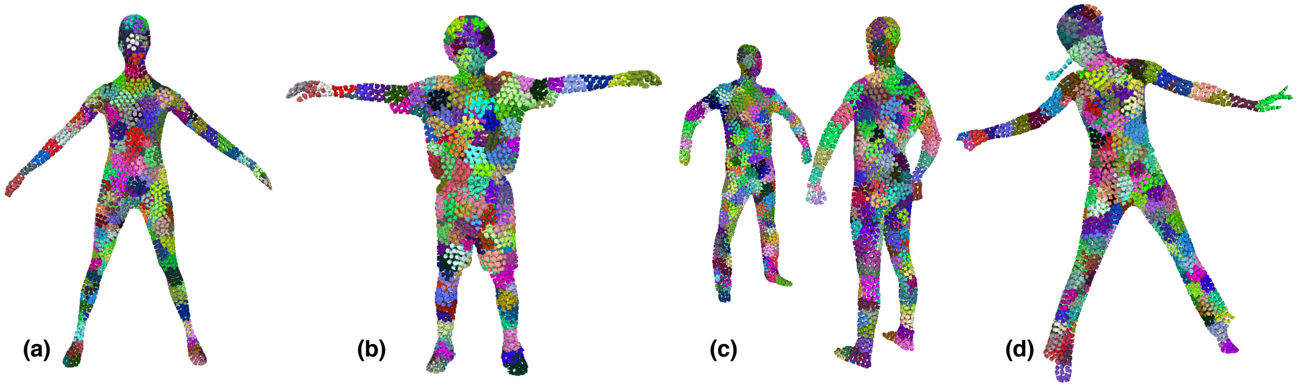

Figure 3. Patched volumetric decompositions of the template for sequences (a) BALLET, (b) GOALKEEPER-13, (c) MARKER and (d) DANCER. A color has been assigned to each patch for a visualization purpose.

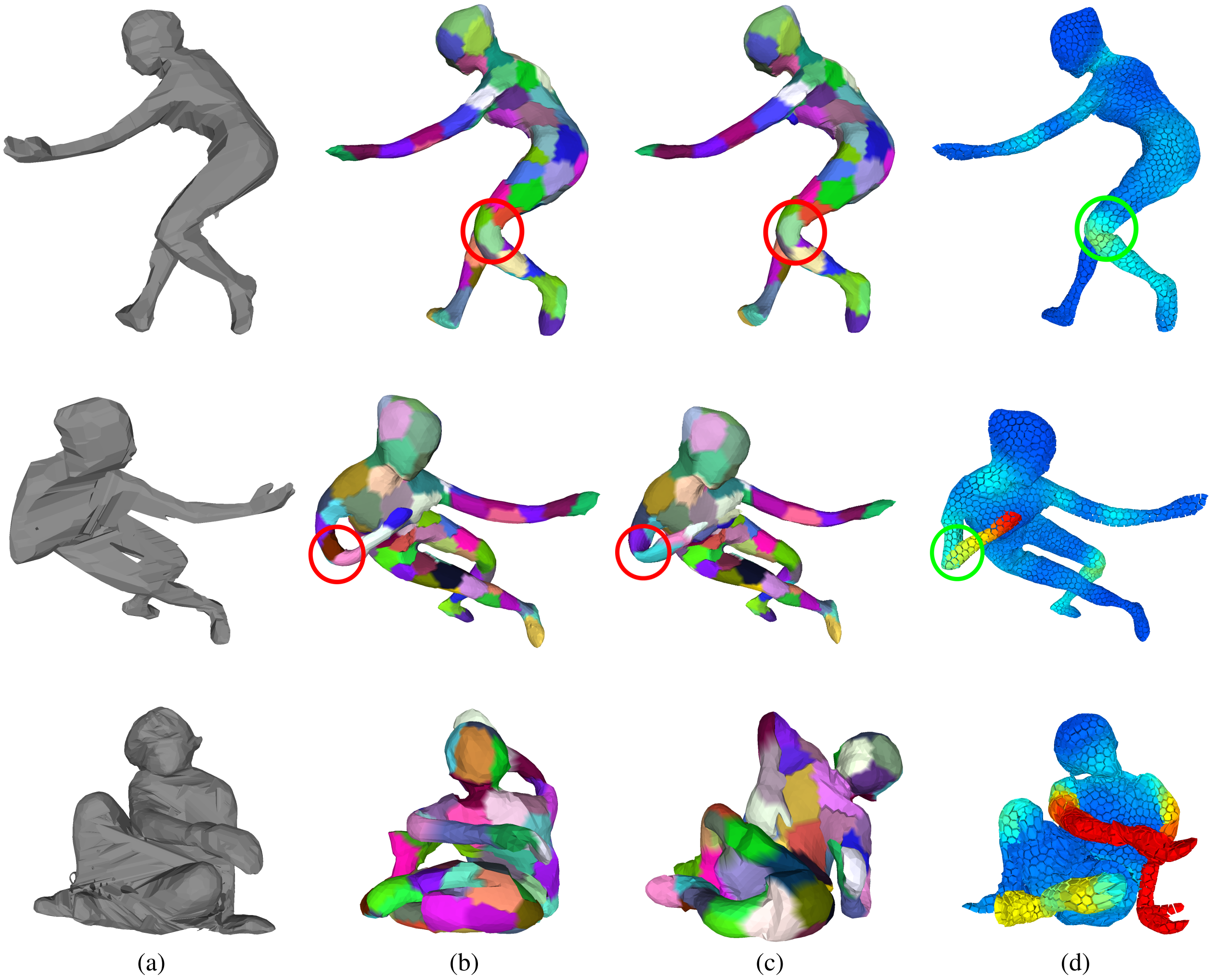

Figure 4. Frames of the BALLET (top and middle) and GOALKEEPER-13 datasets (bottom). (a) Visual hull input. (b) Tracking result of Cagniart *et al*. [8]. (c) Allain *et al*. [2]. (d) Our method. Note the improved angular shapes on the dancer's knee (top) and elbow (middle), and the improved robustness (bottom).

[18] J.Gall, C.Stoll, E. de Aguiar, C. Theobalt, B. Rosenhahn, and H.-P. Seidel. Motion capture using joint skeleton tracking and surface estimation. In *CVPR*, 2009. 2

[19] P. Joshi, M. Meyer, T. DeRose, B. Green, and T. Sanocki. Harmonic coordinates for character articulation. In *ACM SIGGRAPH 2007 Papers*, 2007. 2

[20] K. Kolev, T. Brox, and D. Cremers. Fast joint estimation of silhouettes and dense 3d geometry from multiple images. *PAMI*, 2012. 1

[21] Y. Liu, J. Gall, C. Stoll, Q. Dai, H.-P. Seidel, and C. Theobalt. Markerless motion capture of multiple characters using multi-view image segmentation. *PAMI*, 2013. 1, 2, 5, 6

[22] Y. Liu, W. Wang, B. Lévy, F. Sun, D.-M. Yan, L. Liu, and C. Yang. On centroidal voronoi tessellation - energy smoothness and fast computation. *ACM Transactions on Graphics*, 28(101), 2009. 3

[23] N. Magnenat-Thalmann, R. Laperrière, and D. Thalmann. Joint-dependent local deformations for hand animation and object grasping. In *Proceedings on Graphics Interface '88*, pages 26–33, Toronto, Ont., Canada, Canada, 1988. Canadian Information Processing Society. 2

[24] W. Matusik, C. Buehler, and L. McMillan. Polyhedral visual hulls for real-time rendering. In S. Gortler and K. Myszkowski, editors, *Rendering Techniques 2001*, Eurographics, pages 115–125. Springer Vienna, 2001. 1

[25] M. Salzmann, F. Moreno-Noguer, V. Lepetit, and P. Fua. Closed-form solution to non-rigid 3d surface registration. In *ECCV*, 2008. 2

[26] Y. Savoye and J.-S. Franco. Cage-based tracking for performance animation. In *ACCV*, 2010. 2

[27] T. W. Sederberg and S. R. Parry. Free-form deformation of solid geometric models. In *Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '86, pages 151–160, 1986. 2

[28] O. Sorkine, D. Cohen-Or, Y. Lipman, M. Alexa, C. Rössl, and H.-P. Seidel. Laplacian surface editing. In *Proceedings of the 2004 Eurographics/ACM SIGGRAPH Symposium on Geometry Processing*, SGP '04, pages 175–184, 2004. 2

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

[30] R. Szeliski. Rapid octree construction from image sequences. *CVGIP: Image Underst.*, 58(1):23–32, 1993. 1

[31] J.-M. Thiery, J. Tierny, and T. Boubekeur. Cager: Cage-based reverse engineering of animated 3d shapes. *Computer Graphics Forum*, 31(8):2303–2316, 2012. 2

[32] A. O. Ulusoy, O. Biris, and J. L. Mundy. Dynamic probabilistic volumetric models. In *ICCV*, 2013. 1

[33] S. Vedula, S. Baker, S. M. Seitz, and T. Kanade. Shape and motion carving in 6d. In *CVPR*, 2000. 1

[34] D. Vlasic, I. Baran, W. Matusik, and J. Popovic. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3), 2008. 2