2016 Fourth International Conference on 3D Vision

Computing temporal alignments of human motion sequences in wide clothing using geodesic patches

Aurela Shehu[1,*], Jinlong Yang[2,3,*], Jean-Sébastien Franco[2,3], Franck Hétroy-Wheeler[2,3] and Stefanie Wuhrer[2]
[1]Saarbrücken Graduate School of Computer Science, Germany
[2]Inria, Grenoble, France
[3]LJK, University Grenoble Alpes, Grenoble, France
{aurela.shehu, jinlong.yang, jean-sebastien.franco, franck.hetroy, stefanie.wuhrer}@inria.fr

Abstract

In this paper, we address the problem of temporal alignment of surfaces for subjects dressed in wide clothing, as acquired by calibrated multi-camera systems. Most existing methods solve the alignment by fitting a single surface template to each instant's 3D observations, relying on a dense point-to-point correspondence scheme, e.g. *by matching individual surface points based on local geometric features or proximity. The wide clothing situation yields more geometric and topological difficulties in observed sequences, such as apparent merging of surface components, misreconstructions, and partial surface observation, resulting in overly sparse, erroneous point-to-point correspondences, and thus alignment failures. To resolve these issues, we propose an alignment framework where point-to-point correspondences are obtained by growing isometric patches from a set of reliably obtained body landmarks. This correspondence decreases the reliance on local geometric features subject to instability, instead emphasizing the surface neighborhood coherence of matches, while improving density given sufficient landmark coverage. We validate and verify the resulting improved alignment performance in our experiments.*

1. Introduction

Capturing dynamic scenes, such as human body motion and performance, is a task of interest with a broad range of applications for entertainment, design, medical, and cultural heritage purposes. Typically, performance is observed with a set of calibrated cameras, yielding a set of individual raw 3D reconstructions with no temporal coherence. An important and challenging task in this context is to perform temporal alignment of the sequence, i.e. expressing the entire 3D sequence as the deformation of a single deformed template mesh. This representation has large benefits as it allows for efficient storage, transmission, reverse scene analysis and semantic characterization of the scene as one moving body.

How to obtain this alignment has been widely studied in the context of humans with tight clothing, where only the motion of humans needs to be characterized. This yields a large family of template-based surface tracking methods, which follow a similar resolution canvas. First a surface template is chosen and obtained, that can be *e.g.* a generic human model, a particular reconstruction among those observed, or a pre-obtained scan of the human subject. Second, a point-to-point correspondence scheme is devised, using point proximity in Euclidean space, in a geometric or appearance feature space, or on a learned manifold. Third, because the correspondences so obtained are often insufficiently dense, non-uniformly distributed over the body, and erroneous, a deformation model with a reduced control parameter set is used to constrain the estimation of the full body alignment and reduce the search space of deformation, typically based on human kinematics, piecewise rigid or affine.

Wide clothing acquisitions generally fail tight clothing alignment method assumptions. In particular the correspondence becomes much more challenging since such sequences exhibit much stronger geometric and topological noise, more occlusions, and non-rigid behavior, due to the inherent geometric variability of clothing. The sensitivity of correspondences to the representativeness of template topology and geometry is also drastically increased.

We address this problem with a full alignment solution, centered on a correspondence model suitable for clothes. We ground this work in two key assumptions. First, we assume that the topology of the human in clothing is fixed over time. This assumption holds for the human body itself

*Joint first authors

978-1-5090-5407-7/16 $31.00 © 2016 IEEE
DOI 10.1109/3DV.2016.27

as well as many clothing styles such as t-shirts, trousers and skirts. Second, we assume that the geometry of the model deforms in a near-isometric way. This is true when considering locomotions of the human and clothing that are not very elastic. In practice, the acquired object violates both assumptions due to aforementioned acquisition noise. That is, the acquired topology may change due to the merging of close-by body parts, which in turn completely changes the intrinsic geometry of the model. To design a method that is robust to this type of acquisition noise, we use a deformation model based on *partial* near-isometric patches. These patches are grown from a set of preselected landmarks that can be robustly found in each frame, and ensure consistent densification of correspondences over the surface. To ease correspondence over all frames, we also provide an automatic template selection method among raw 3D models of the input sequence, maximizing topological adequacy.

Our experiments demonstrate the success of the method for temporal alignment of a variety of clothing sequences. In particular we favorably compare our method against two state of the art temporal alignment methods [1, 2] based on more restrictive locally rigid deformation assumptions, on both pretracked and real datasets. Results show that our method outperforms previous work for human characters with wide or layered clothing.

2. Related work

Most methods to temporally align sequences of 3D data require a prior on the shape to track. More general techniques without prior knowledge on the geometry also exist, but they assume a strong temporal coherence. These methods use nearest neighbor techniques as a prior to aid the alignment. They usually suffer from drift and cannot follow fast motion [7]. Dou *et al*. [9] reduce drifts errors by detecting loop closures and distributing the alignment error over the loop. Newcombe *et al*. [15] reconstruct a non-rigidly deforming surface while estimating a per frame volumetric warp field that transforms this surface into the live frame. Tevs *et al*. [21] solve the problem of following fast motion by first computing a few landmark correspondences and then propagating to a dense matching, assuming isometric deformation. Landmarks are computed as geometric feature points. Contrary to [21], our landmarks are not computed as geometric features on the cloth surface. Instead, we estimate the body shape under the cloth and then automatically transfer the anatomical landmarks from the body shape to the cloth. Many methods for human body shape estimation under clothing have been proposed [11, 23, 14, 17]. We use the recent method of Yang *et al*. [24] since it works for moving shapes and is fully automatic.

The majority of works for temporal alignment of human motion sequences make use of a template showing a specific subject in a specific type of clothing. The main advantage of such methods is that they are particularly robust to severe acquisition artifacts. However, a template is required for tracking. For example, the method of Bradley *et al*. [5], which is designed for the temporal alignment of a moving cloth, constructs the template from a photograph of the garment. Aguiar *et al*. [8, 20] create the template by a full-body laser scan of the subject in its current clothes before performing capture. A physically-based cloth model is then used for temporal alignment. Budd *et al*. [7] introduce a way to align frames that are not necessarily temporally adjacent, but most similar according to a volume-based shape similarity measure, which makes the template-based alignment more robust to large motions between adjacent frames. Allain *et al*. [1] perform shape tracking assuming a locally rigid deformation model, in order to compute both the mean pose and correspondences over time. The template is manually selected as one of the meshes of the input sequence. While this method is surface-based, the follow-up method [2] proposes a volumetric parametrization of the tracking which shows better results than surface-based methods. The shape is represented by a centroidal Voronoi tessellation, which enables volume conservation. This method also assumes a template is provided. Our method is template-based for robustness purposes. However, no additional information nor user interaction is necessary to build the template since it is automatically selected among the meshes of the sequence, using a method similar in spirit to the one of Letouzey and Boyer [12].

Most temporal alignment methods are well adapted to human characters in tight clothes. Methods handling loose clothing usually assume near-isometric deformations [5, 21]. This is because most clothes do not usually stretch during motion [10]. However, computing approximate isometries can be computationally expensive [6, 19]. In our method, we reduce the computation cost by selecting anatomically-motivated landmarks to discover approximate isometries.

3. Method overview

Given an unstructured temporal sequence showing a clothed human performing a motion, our method produces a consistently deforming model. To be robust to acquisition and topological noise our method has three steps as shown in Figure 1: a human body model is used to find good initial starting points for the partial near-isometries; point trajectories are computed using a partial near-isometric deformation model; and the point trajectories are used as assignments in a template-based deformation to find a coherent model that deforms over time. We now provide more detail for each step.

First, we estimate the undressed human body shape in motion for the given sequence. This step reduces significantly the search space of the problem by providing a good

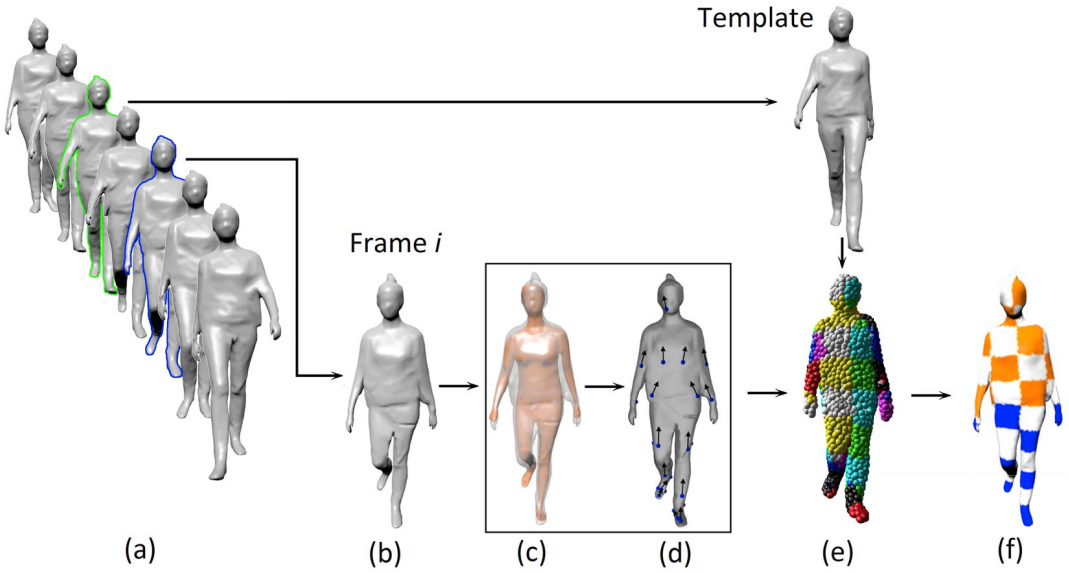

Figure 1. Method overview. Given an input sequence of 3D meshes shown in (a), each frame is processed in three steps, which are shown for the frame indicated in blue and shown in (b). First, a statistical model of undressed body shape is fitted to the frame (c). Pre-marked anatomical points on the fitted human body model are mapped to the input frame (d). Second, these anatomical markers are used to guide a partial near-isometric correspondence computation between an automatically selected template (green model in (a), and shown in top row) and the frame (e). Third, the resulting correspondences are used as assignments to deform the template to the input frame (f).

initialization for the computation of near-isometric partial matches. This step takes advantage of recent robust methods that use statistical human body models to estimate the naked human body shape under clothing [11, 23, 14, 17, 24]. In our implementation, we use an automatic method that estimates the naked body shape in motion under clothing based on a 3D input sequence [24]. For the specific frame that is shown in blue in Figure 1(a) and enlarged in (b), the estimated human body shape is shown in Figure 1(c). The input frame is shown in grey and the estimated human body under clothing in brown. Next, we transition from the naked human body to surface of the clothing. We once manually select a small set of anatomical points on the statistical model and automatically transfer these points on the clothed sequence as shown in Figure 1(d).

Second, we use a partial near-isometric deformation model that is robust with respect to topological acquisition noise to compute point trajectories on the input sequence. Following Letouzey and Boyer [12], we proceed by first taking advantage of the assumption that the real topology stays fixed over time to automatically select a template frame from the sequence based on topological and geometric criteria. The automatically selected template is shown in green in Figure 1(a) and enlarged in the first row. Second, for each frame of the sequence, we independently compute correspondence information between the template and the frame. This computation finds partial near-isometric matches between two frames [6]. To increase robustness and make the algorithm efficient, we use the previously computed anatomical points on each frame for initialization. The partial matches are then merged into a global correspondence. The computed correspondence between template and a frame is shown in Figure 1(e). This correspondence information between the template and every frame of the sequence results in point trajectories over time.

Third, we find a consistent topology that deforms over time by extending a standard template fitting technique [3] to take advantage of the previously computed point trajectories and operate on motion sequences. Since the previous step computes the correspondence information between the template and every other frame independently, the resulting point trajectories may be interrupted and the correspondence of a particular point on the template may not be known on every frame. To remedy this we add this final step that deforms the template to the entire motion sequence using previously computed point trajectories as assignments. The output of this step is shown in Figure 1(f).

4. Clothed human alignment

This section provides details of the three steps of our proposed method.

4.1. Human body estimation and mapping anatomical points to clothed human

We start by estimating the human body shape for the given input sequence. The result of this step allows to map

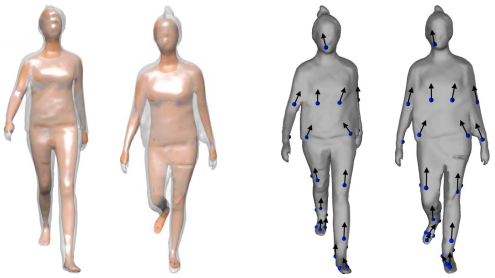

Figure 2. Left: estimated human body $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$ shown in brown overlaid with two input frames. Right: oriented anatomical points on two meshes of a sequence. The points are shown in blue, and their orientations in black.

a sparse set of anatomical points to the input frames of a human in wide clothing.

Human body estimation Given a human motion sequence in wide clothing, we start by estimating the undressed human body shape for the entire sequence using a recent automatic method [24]. This method takes advantage of a learned statistical human body model [16]. In particular, the statistical body model represents the variation in identity across different people in a 100-dimensional shape space learned using principal component analysis, and the variations in posture for a fixed identity using a linear blend skinning that is found using a rigging tool [4]. This model allows to generate a human model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta})$, where $\boldsymbol{\beta}$ denotes the parameters controlling identity and $\boldsymbol{\Theta}$ for posture.

The method proceeds by finding a single $\boldsymbol{\beta}$ along with $\boldsymbol{\Theta}_i$ for each input frame by minimizing the energy

$$E_{body} = w_l E_l + w_n E_n + w_w E_w \tag{1}$$

consisting of a weighted sum of three terms, where E_l is an energy that aligns a small set of automatically detected landmarks, E_n is an energy that pulls each vertex on $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$ to its nearest neighbor on the i-th frame, and E_w is designed to cope with wide clothing by encouraging $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$ to be located inside all observed frames. The weights w_l, w_n, and w_w allow to trade off the influence of the different energy terms, and we follow the weight schedule proposed by Yang *et al*. [24].

Figure 2 shows the estimated human body shape and posture for two frames of a sequence. Note that all meshes $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta})$ generated this way share the same vertex ordering and mesh topology.

Mapping anatomical points from human body to surface of clothing We manually select a small set of oriented anatomical points on the statistical model $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta})$, where an oriented point is a 3D location on the surface along with a direction in its tangent plane. An oriented point is in practice selected by choosing two points: a starting point and a close-by neighbor that defines the direction in the tangent plane. Note that these points only need to be chosen once for any $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$, as all body models share the same mesh structure.

To automatically map an oriented anatomical point $\mathbf{t}$ on $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$ to the clothing surface of the i-th frame $\mathcal{S}_i$, we intersect the line through $\mathbf{t}$ along the normal direction of $\mathbf{t}$ with $\mathcal{S}_i$. If there's any intersection outside of $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$, and the distance of the one closest to $\mathbf{t}$ is within a threshold τ_o, this intersection point is considered a valid mapping. Otherwise we look for the intersection that lies inside $\mathcal{M}(\boldsymbol{\beta}, \boldsymbol{\Theta}_i)$, and chose the closest one to be a valid mapping if the distance to $\mathbf{t}$ is smaller than τ_i. In case no valid mapping is found, $\mathbf{t}$ is removed from consideration for $\mathcal{S}_i$.

An example of all valid oriented anatomical points on two frames of a sequence is shown in Figure 2, where points are shown in blue and orientations in black.

4.2. Computation of point trajectories

Here we describe how to compute point trajectories given the input sequence and oriented anatomical points. First, we automatically select a template from all frames of the sequence. Second, we establish a dense correspondence between the template and each frame of the sequence with the help of a partial near-isometric deformation model.

Automatic template selection Often in an acquired sequence the topology changes over time due to acquisition noise. In particular, tiny holes frequently appear and large contacts appear when limbs are close-by, see Figure 3 for an example. To remedy this problem during alignment, we automatically detect in the sequence a template $\mathcal{T}$ and register $\mathcal{T}$ to every other frame $\mathcal{S}$ of the sequence.

The template selection is based on topological and geometric criteria. Since we assume that the scene has a fixed topology and that observed topology changes come from acquisition artifacts, it follows that observed topological properties of the shape can only grow [12]. That is, observed splits are accepted as changes of the real shape while observed merges are ignored from consideration as the real shape cannot merge. To account for acquisition artifacts, splits are only considered if they are persistent over time and small components are filtered. As observed splits are accepted, we first select as candidate templates all frames that have a maximum number of components. As observed splits and merges are directly linked to the genus number, we further select from the candidate templates the frames with minimum genus of the largest component.

The aforementioned topological criteria might provide several candidate templates. The final selection is based on a geometric quality criterion. From the candidate templates, we select the one with minimum area ratio (maximum point area/minimum point area, where a point area is the area of

Figure 3. Some frames of a sequence and the computed template $\mathcal{T}$ highlighted in green.

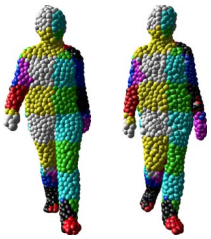

Figure 4. Color-coded correspondence information computed between $\mathcal{T}$ (right) and $\mathcal{S}$ (left) using a partial near-isometric deformation model.

the Voronoi region around a vertex) of the largest component, as the rest of our pipeline benefits from a template whose vertices are as uniform in area as possible. Figure 3 shows the selected template $\mathcal{T}$ for one of our test sequences.

Dense correspondence computation To align $\mathcal{T}$ to a frame $\mathcal{S}$, we use a partial near-isometric deformation model that is robust to topological acquisition noise. Near-isometric models have previously been used successfully for cloth modeling [18]. Ideally, $\mathcal{T}$ and $\mathcal{S}$ should be mapped by a global near-isometric mapping. In practice, due to acquisition noise, such a global mapping may not exist.

To remedy this, we use a partial near-isometric model to account for topological acquisition noise as in Brunton *et al*. [6]. To this end, we consider every frame $\mathcal{S}$ to be a set of smooth, orientable 2-manifolds embedded in three-dimensional space. In practice, $\mathcal{S}$ is discretized by a set of points that are connected by a neighborhood graph. We denote the geodesic distance between two points $\mathbf{s}_i, \mathbf{s}_j \in \mathcal{S}$ by $d_{\mathcal{S}}(\mathbf{s}_i, \mathbf{s}_j)$.

Consider a mapping $f : \mathcal{U} \rightarrow \mathcal{S}$, where $\mathcal{U}$ is a subset of $\mathcal{T}$. The mapping function f is a near-isometry if:

$$|d_{\mathcal{U}}(\mathbf{t}_i, \mathbf{t}_j) - d_{\mathcal{S}}(f(\mathbf{t}_i), f(\mathbf{t}_j))| \leq \epsilon \qquad (2)$$

where $\mathbf{t}_i, \mathbf{t}_j$ are vertices in $\mathcal{U}$ and ϵ refers to the allowed stretching threshold. Since $\mathcal{U} \subset \mathcal{T}$ refers to a shape part, we seek parts of $\mathcal{T}$ that can be mapped to parts of $\mathcal{S}$ without much stretching.

Brunton *et al*. [6] use the partial near-isometry model for pairwise frame alignment. They show that a correspondence between an oriented point on $\mathcal{T}$ and an oriented point on $\mathcal{S}$ is sufficient to recover an isometric mapping. In the following, we use $\mathbf{t}$ and $\mathbf{s}$ to denote both points and oriented points on $\mathcal{T}$ and $\mathcal{S}$, respectively.

Due to the acquisition noise in the acquired data, it is necessary to define several corresponding oriented point pairs to recover a full alignment between $\mathcal{T}$ and $\mathcal{S}$. It is known that the automatic computation of corresponding oriented point pairs is a computational bottleneck for the original method that renders the processing of motion sequences impractical.

We overcome this problem by using the automatically computed oriented anatomical points on each of the frames. For these oriented points, the correspondence information is known across frames. We use them as starting points for the partial near-isometric correspondence computation between $\mathcal{T}$ and $\mathcal{S}$. To speed up the computation time, we minimize point correspondence information coming from many overlapping partial near-isometries that is discovered from nearby starting points. For this, we stop when the discovered near-isometric part has an intrinsic radius bigger than a distance threshold τ_d. To increase robustness, we ignore from consideration near-isometric parts that have an intrinsic radius smaller than a threshold τ_s.

After finding multiple near-isometric parts between $\mathcal{T}$ and $\mathcal{S}$, we merge the parts into a global alignment by assigning to each $\mathbf{t}$ in $\mathcal{T}$ that is mapped to at least one point on $\mathcal{S}$ the geodesic average of all computed assignments on $\mathcal{S}$. Figure 4 shows a color-coded alignment computed between $\mathcal{T}$ and $\mathcal{S}$ for one of the test sequences.

After this step, we have computed pairwise correspondence information between $\mathcal{T}$ and every frame $\mathcal{S}$ of the sequence. This allows us to follow the trajectory of a particular point on $\mathcal{T}$ over time.

4.3. Consistent topology alignment

In the previous step, we computed frame correspondences between $\mathcal{T}$ and every other frame $\mathcal{S}$. To find a consistent topology that deforms over time we extend a standard template fitting technique [3] to take advantage of the previously computed point trajectories. That is, the previously computed frame correspondence serves as assignment information to guide the template deformation to allow for

fast motion and reduce drift.

In particular, let $\mathcal{T}$ contain $N_{\mathcal{T}}$ vertices, and let the motion sequence contain N_f frames $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_{N_f}$. Without loss of generality, let $\mathcal{T}$ be frame $\mathcal{S}_j$. We fit the frames of the sequence starting at $\mathcal{T}$ by processing adjacent frames in (backward and forward temporal) order as $\mathcal{S}_{j-1}, \ldots, \mathcal{S}_1$ and $\mathcal{S}_{j+1}, \ldots, \mathcal{S}_{N_f}$. After $\mathcal{S}_i$ has been fitted, we update $\mathcal{T}$ to the fitting result of $\mathcal{S}_i$. This initializes the location and shape of the template to be close to the next frame to be fitted, thereby allowing for a simple template fitting scheme.

The template fitting follows a standard technique [3], and fits $\mathcal{T}$ to $\mathcal{S}$ by optimizing the following energy

$$E_{template} = w_c E_c + w_d E_d + w_s E_s + w_t E_t, \quad (3)$$

which consists of the weighted linear combination of four simple energy terms: correspondence energy E_c, nearest neighbor energy E_d, deformation smoothness energy E_s, and temporal smoothness energy E_t, which are weighted by w_c, w_d, w_s, and w_t, respectively. To model the deformation, each vertex $\mathbf{t}_i$ of $\mathcal{T}$ is expressed in homogeneous coordinates and transformed using a 4×4 matrix $\boldsymbol{A}_i$ that represents an affine transformation in $\mathbb{R}^3$.

The first energy term modifies the marker energy proposed by Allen *et al.* [3] to use the previously computed correspondences from $\mathcal{T}$ to $\mathcal{S}$ as assignments. The energy is expressed as

$$E_c = \frac{1}{\sum_{i=1}^{N_{\mathcal{T}}} w_{corr,i}} \sum_{i=1}^{N_{\mathcal{T}}} w_{corr,i} \|\boldsymbol{A}_i \mathbf{t}_i - \mathbf{s}_c(\mathbf{t}_i)\|_2^2, \quad (4)$$

where $w_{corr,i}$ is a weight equal to 1 if $\mathbf{t}_i$ has a correspondence on $\mathcal{S}$ and equal to 0 otherwise, $\mathbf{s}_c(\mathbf{t}_i)$ is the point of $\mathcal{S}$ that corresponds to $\mathbf{t}_i$ and $\|\cdot\|_2$ is the Euclidean distance. Using this energy guides the deformation in case of large motion between adjacent frames and reduces drift. Note that unlike the marker term, our energy does not rely on any manually provided information on $\mathcal{S}$, but takes advantage of the correspondences found using the partial near-isometric deformation model.

The second energy term is a standard data term that pulls each $\mathbf{t}_i$ to its nearest neighbor on $\mathcal{S}$ as

$$E_d = \frac{1}{\sum_{i=1}^{N_{\mathcal{T}}} w_{NN,i}} \sum_{i=1}^{N_{\mathcal{T}}} w_{NN,i} \|\boldsymbol{A}_i \mathbf{t}_i - \mathbf{s}_n(\mathbf{t}_i)\|_2^2, \quad (5)$$

where $w_{NN,i}$ is a weight equal to 1 if the nearest neighbor is valid and 0 otherwise, and $\mathbf{s}_n(\mathbf{t}_i)$ is the point on $\mathcal{S}$ that is the nearest neighbor of $\mathbf{t}_i$. We consider the nearest neighbor valid if the surface normals at each point are less than 90° apart and the distance between two points is at most $0.2m$. When used without the correspondence energy (Equation (4)), this nearest neighbor energy is known to suffer from drift [7]. To avoid this problem, we only activate this energy once the deformed template is close to $\mathcal{S}$, thereby reducing drift.

The third energy term is the standard deformation smoothness energy

$$E_s = \frac{1}{\sum_{i=1}^{N_{\mathcal{T}}} |\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \|\boldsymbol{A}_i - \boldsymbol{A}_j\|_F^2, \quad (6)$$

where $\mathcal{N}(i)$ is the 1-ring neighborhood of $\mathbf{t}_i$ and $\|\cdot\|_F$ is the Frobenius norm. This term encourages a smooth deformation field across $\mathcal{T}$.

As the fourth term, we add a simple temporal smoothness energy to prevent very large deformations between adjacent frames as

$$E_t = \frac{1}{N_{\mathcal{T}}} \sum_{i=1}^{N_{\mathcal{T}}} \|\boldsymbol{A}_i - \boldsymbol{I}\|_2^2, \quad (7)$$

where $\boldsymbol{I}$ is the identity matrix. This energy discourages large displacements of individual vertices between adjacent frames and helps to prevent jittering in case of slightly inaccurate correspondences between $\mathcal{T}$ and $\mathcal{S}$.

Figure 5 shows an example of a template-fitted frame for one of our sequences, and Figure 6 shows the template fitting with and without E_c, for an example where the target frame is only four frames away from $\mathcal{T}$.

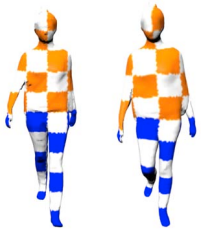

Figure 5. Template $\mathcal{T}$ (left) and template-fitted frame $\mathcal{S}$ (right). Correspondence is color-coded.

4.4. Implementation Details

For the human body estimation under clothing, we use the code of Yang *et al.* [24]. We manually select 40 oriented anatomical points on the statistical model and map these

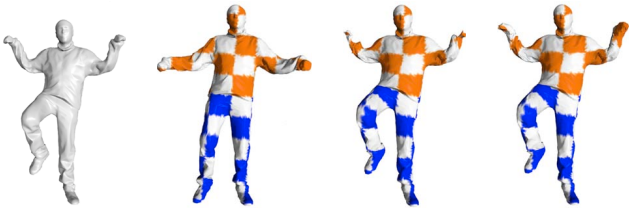

Figure 6. Our E_c energy term prevents loss of tracking. From left to right: $\mathcal{S}$, $\mathcal{T}$, template deformation without E_c, and with E_c.

points to the clothed sequence. To map points from human body to clothed human we set $\tau_i = 0.2m$ and $\tau_o = 0.05m$.

To compute $\mathcal{T}$ over a given sequence we do not count as components shells with an area smaller than $0.1m^2$. To compute dense correspondence we use the code of Brunton *et al*. [6] with $\epsilon = 0.25m$. To increase robustness and computational efficiency, we set $\tau_s = 0.1m$ and $\tau_d = 0.25m$.

The weights (w_c, w_d, w_s, w_t) in the template fitting step are fixed by solving the optimization problem in several stages with different energy weights at each stage as proposed by Allen *et al*. [3]. So that when the template is far away from the target frame, correspondence and smoothness terms can lead the deformation and we set the weights to $(1, 400, 0, 0.1)$. Now the template is already closely enough to the target, we can use nearest neighbor information. We decrease the weight of spatial smoothness and increase the data term and set the weights to $(1, 200, 10, 0.1)$. Next, we turn off the weights of the dense correspondence term and optimize the energy based on the other three weights. This is to address inaccurate dense correspondence information. We set the weights to $(0, 100, 10, 0.1)$ and finally to $(0, 50, 10, 0.1)$. The nearest neighbors are recomputed for each vertex every 20 iterations of optimization. The energy $E_{template}$ is optimized using a quasi-Newton method [13].

5. Experiments

For better visualization, please refer to the supplemental video*.

Pretracked dataset We have tested our method against the methods by Allain *et al*. [1, 2] on four sequences (`crane`, `march2`, `samba`, `squat1`) from Vlasic *et al*. [22]. Since this dataset is temporally registered we can use it as ground truth for our evaluation. Our method takes as input the original pretracked dataset and we compute as error the Euclidean distance between our estimated vertex position and the ground truth. Both methods by Allain *et al*. require preprocessing of the data, which involves manual selection of the template and downsampling of the input mesh. We first map the processed template to the pretracked template by a nearest neighbor approach and then proceed with the evaluation as described above. The preprocessing is done in favor of the reference algorithms.

All three methods share similar average quantitative results. The average vertex error over all sequences is 3.6cm, 3.9cm and 3.9cm for [1], [2] and our method respectively. The percentage of the vertices within 10cm error is 96.22, 94.68, and 96.83 respectively. Figure 7 shows the per-frame error on each sequence and for each method. Our method is

*`https://hal.inria.fr/hal-01367791`

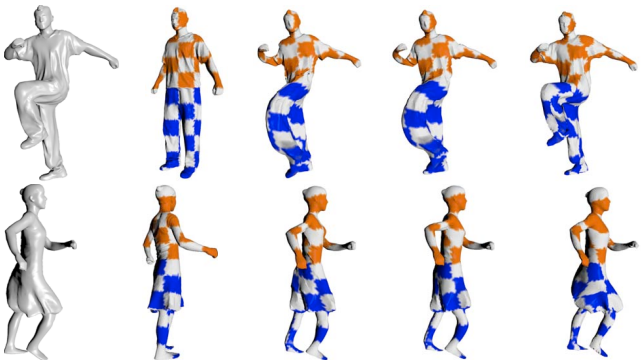

Figure 8. Results on sequences `march2` (top row, frame #55) and `samba` (bottom row, frame #90) from [22]. From left to right: $\mathcal{S}$, $\mathcal{T}$, result with [1], result with [2], and our result. See also the supplementary video.

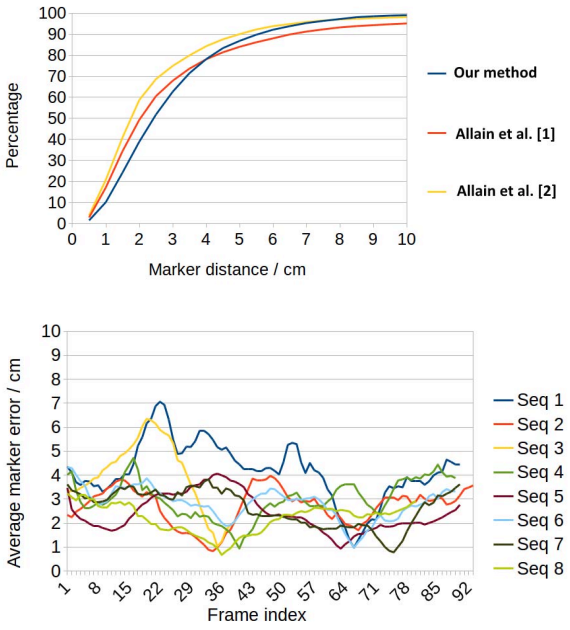

Figure 9. Quantitative results on sparse markers (dataset of [24]). Top: cumulative error curves for [1], [2] and our method. Bottom: average marker error per frame of our method for all eight sequences.

generally more robust in the sense the maximum error for a sequence is lower. As shown in Figure 8, our method also better follows the deformation of wide clothes and makes good use of the prior knowledge of the underlying human whereas the two other methods loose tracking of the knee.

Dataset with sparse markers We have also tested our method on eight sequences from the dataset of Yang *et al*. [24]. These sequences include two subjects, one male

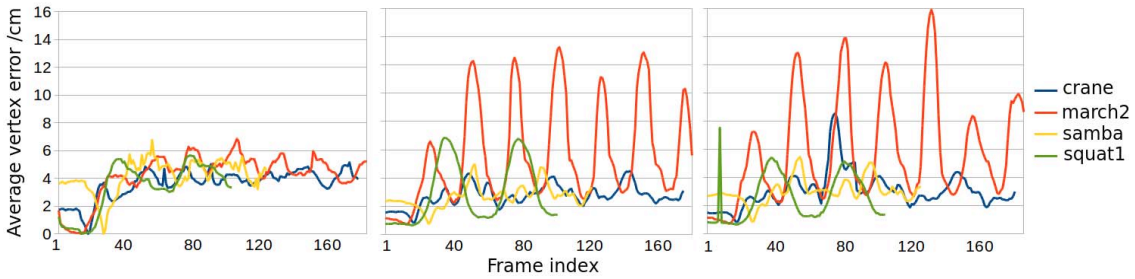

Figure 7. Per-frame result on 4 sequences from [22]. From left to right: our method, Allain et al. [1] and Allain et al. [2].

and one female, wearing either layered or wide clothes, either performing a walk or rotating the body. The mesh sequences were obtained using visual hull reconstructions from a 68-camera-system. Reconstruction artifacts like holes are present. Along with the mesh sequence, marker trajectories are also provided. Each subject has 14 markers placed on anatomically significant locations: forehead, shoulders, elbows, wrists, belly, knees, feet and heels. If these locations are covered by cloth while the subject is at resting pose then markers are placed on the clothes. We map each marker position to the nearest template vertex and use this mapping and correspondence information to compute the distance error of the alignment. We compute as error the Euclidean distance between our estimated location and the ground truth location of the marker.

Figure 9 shows a quantitative evaluation of our method on these sparse markers. The cumulative error curve, from which the template frame is excluded, shows that our method performs as well as the methods by Allain *et al*. The bottom curves show that our method does not drift in time, because our dense correspondence is calculated from the template to each frame independently and is based on a partial near-isometric deformation model.

Figure 10 shows alignment results on representative sequences of the dataset of [24]. As for the pretracked dataset, these results also demonstrate that our method is more robust than [1] and [2], where limbs may be switched or the tracking of clothing may get lost. The color-coding shows that our method does not suffer from significant drift. Artifacts produced in the visual hull reconstruction even have a chance to be repaired such as the tunnel appearing on the left leg for the last example.

6. Conclusions

In this paper we have presented a method for the temporal alignment of motion sequences of humans in wide clothing. In order to be robust to geometric and topological artifacts, we use reliable body landmarks which are mapped to the clothed surface and serve as starting points to grow partial near-isometric patches. These patches allow to compute pairwise correspondence information between

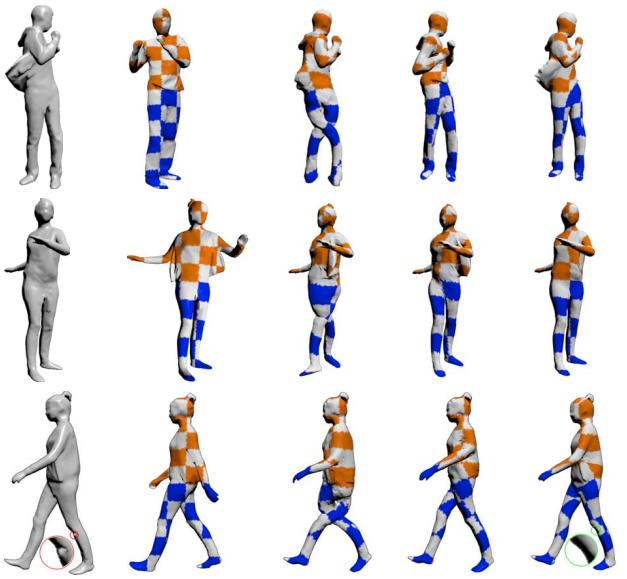

Figure 10. Alignment result on three representative sequences of [24]. From left to right: $\mathcal{S}$, $\mathcal{T}$, result with [1], result with [2] and our result.

an automatically selected template frame, based on topological and geometric criteria, and every other frame of the sequence. Correspondences are then used to guide the deformation from the template to the other frames.

Given a statistical human model with a sparse set of annotated anatomical points, our method is fully automatic. It relies on a few assumptions, which are valid in most practical cases. In particular, cloth is supposed to be inelastic and not to slip much along the body. The topology of the scene is supposed to stay constant, and at least one frame in the sequence needs to capture the correct topology.

Acknowledgments

This work has been supported by the ANR through the ACHMOV project (ANR-14-CE24-0030). We would like to thank Benjamin Allain for providing comparison code, Daniel Vlasic for providing the pre-tracked dataset and the Kinovis platform[†] for the sparse marker dataset.

[†]http://kinovis.inrialpes.fr

References

[1] B. Allain, J. Franco, E. Boyer, and T. Tung. On mean pose and variability of 3d deformable models. In *European Conference on Computer Vision*, pages 284–297, 2014.

[2] B. Allain, J.-S. Franco, and E. Boyer. An efficient volumetric framework for shape tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 268–276, 2015.

[3] B. Allen, B. Curless, and Z. Popović. The space of human body shapes: Reconstruction and parameterization from range scans. *ACM Transactions on Graphics*, 22(3):587–594, 2003.

[4] I. Baran and J. Popović. Automatic rigging and animation of 3D characters. *ACM Transactions on Graphics*, 26(3):#72:1–8, 2007.

[5] D. Bradley, T. Popa, A. Sheffer, W. Heidrich, and T. Boubekeur. Markerless garment capture. *ACM Transactions on Graphics*, 27(3):#99:1–9, 2008.

[6] A. Brunton, M. Wand, S. Wuhrer, H.-P. Seidel, and T. Weinkauf. A low-dimensional representation for robust partial isometric correspondences computation. *Graphical Models*, 76:70–85, 2014.

[7] C. Budd, P. Huang, M. Klaudiny, and A. Hilton. Global non-rigid alignment of surface sequences. *International Journal of Computer Vision*, 102:256–270, 2013.

[8] 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):#98,1–10, 2008.

[9] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi. 3d scanning deformable objects with a single rgbd sensor. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 493–501, 2015.

[10] R. Goldenthal, D. Harmon, R. Fattal, M. Bercovier, and E. Grinspun. Efficient simulation of inextensible cloth. *ACM Transactions on Graphics*, 26(3):#49:1–8, 2007.

[11] A. Jain, T. Thormählen, H.-P. Seidel, and C. Theobalt. MovieReshape: tracking and reshaping of humans in videos. *ACM Transactions on Graphics*, 29(6):#148:1–10, 2010.

[12] A. Letouzey and E. Boyer. Progressive shape models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 190–197, 2012.

[13] D. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, 45:503–528, 1989.

[14] A. Neophytou and A. Hilton. A layered model of human body and garment deformation. In *International Conference on 3D Vision*, pages 171–178, 2014.

[15] R. A. Newcombe, D. Fox, and S. M. Seitz. Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 343–352, 2015.

[16] L. Pishchulin, S. Wuhrer, T. Helten, C. Theobalt, and B. Schiele. Building statistical shape spaces for 3d human modeling. Technical Report 1503.05860, arXiv, 2015.

[17] G. Pons-Moll, J. Romero, N. Mahmood, and M. Black. DYNA: a model of dynamic human shape in motion. *ACM Transactions on Graphics*, 34(4):#120:1–14, 2015.

[18] T. Popa, Q. Zhou, D. Bradley, V. Kraevoy, H. Fu, A. Sheffer, and W. Heidrich. Wrinkling captured garments using space-time data-driven deformation. *Computer Graphics Forum*, 28(2):427–435, 2009.

[19] E. Rodolà, S. R. Bulò, T. Windheuser, M. Vestner, and D. Cremers. Dense non-rigid shape correspondence using random forests. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4177–4184, 2014.

[20] C. Stoll, J. Gall, E. de Aguiar, S. Thrun, and C. Theobalt. Video-based reconstruction of animatable human characters. *ACM Transactions on Graphics*, 29(6):#139:1–10, 2010.

[21] A. Tevs, A. Berner, M. Wand, I. Ihrke, M. Bokeloh, J. Kerber, and H.-P. Seidel. Animation cartography: Intrinsic reconstruction of shape and motion. *ACM Transactions on Graphics*, 31(2):#12:1–15, 2012.

[22] D. Vlasic, I. Baran, W. Matusik, and J. Popović. Articulated mesh animation from multi-view silhouettes. *ACM Transactions on Graphics*, 27(3):#97:1–10, 2008.

[23] S. Wuhrer, L. Pishchulin, A. Brunton, C. Shu, and J. Lang. Estimation of human body shape and posture under clothing. *Computer Vision and Image Understanding*, 127:31–42, 2014.

[24] J. Yang, J. Franco, F. Hétroy-Wheeler, and S. Wuhrer. Estimation of human body shape in motion with wide clothing. In *European Conference on Computer Vision*, page to appear, 2016.