Probabilistic Deformable Surface Tracking From Multiple Videos

Cedric Cagniart, Edmond Boyer, Slobodan Ilic

Probabilistic Deformable Surface Tracking from Multiple Videos

Cedric Cagniart[1], Edmond Boyer[2], and Slobodan Ilic[1]

[1] Technische Universität München
[2] Grenoble Universités - INRIA Rhône-Alpes
{cagniart,slobodan.ilic}@in.tum.de, edmond.boyer@inrialpes.fr

Abstract. In this paper, we address the problem of tracking the temporal evolution of arbitrary shapes observed in multi-camera setups. This is motivated by the ever growing number of applications that require consistent shape information along temporal sequences. The approach we propose considers a temporal sequence of independently reconstructed surfaces and iteratively deforms a reference mesh to fit these observations. To effectively cope with outlying and missing geometry, we introduce a novel probabilistic mesh deformation framework. Using generic local rigidity priors and accounting for the uncertainty in the data acquisition process, this framework effectively handles missing data, relatively large reconstruction artefacts and multiple objects. Extensive experiments demonstrate the effectiveness and robustness of the method on various 4D datasets.

1 Introduction

Inferring shapes and their temporal evolutions from image data is a central problem in computer vision. Applications range from the visual restitution of live events to their analysis, recognition and even synthesis. The recovery of shapes using multiple images has received considerable attention over the last decade and several approaches can build precise static 3D models from geometric and photometric information, sometimes in real time. However, when applied to temporal sequences of moving objects, they provide temporally inconsistent shape models by treating each frame independently hence ignoring the dynamic nature of the observed event.

Most methods interested in tracking deformable surfaces in multi-camera systems deform a reference template mesh to fit observed geometric cues as well as possible at each time frame. These cues appear in the literature as photo-consistent models, visual hulls, or even silhouette data directly. Recent works suggest that even without considering photometric information, this geometric data is in many cases sufficiently constraining [1,2,3]. It is however subject to background segmentation and reconstruction errors which needs to be handled in the tracking process. Using strong deformation priors, e.g. articulated models, can help increase robustness but does not extend well to more complex scenes involving several objects whose nature is not necessarily known beforehand. As

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part IV, LNCS 6314, pp. 326–339, 2010.

such scenes require more generic and thus weaker deformation models, it becomes necessary to look into the uncertainty of the data acquisition process and to introduce more robust algorithms modelling its errors.

In this paper, we take these uncertainties into account by embedding the shape tracking within a probabilistic framework. In this way, the need for strong priors is relaxed thus allowing for more complex scenes without sacrifying robustness. The approach considers as input a sequence of independently reconstructed surfaces and iteratively deforms a reference mesh to fit them. The problem is cast as a Bayesian maximum-likelihood estimation where the joint probability of the deformation parameters, i.e. motion, and of the observed data is to be maximized. In order to robustly handle the association between the observations and the reference mesh, latent variables are introduced to identify the mesh region each observation is drawn from, while accounting for possible outliers. We iteratively solve for the motion parameters and posterior probabilities of the latent variables using the Expectation-Maximization algorithm [4].

The remainder of this paper is organized as follows : Section 2 gives an overview previous works that deal with surface tracking in multi-camera environments. In Section 3 we detail our contribution. The corresponding results are presented in Section 4. We conclude the paper by discussing the limitations of our approach and the openings for future work.

2 Related Works

Most of the existing literature dealing with surface tracking in multi-camera environments has to do with the marker-less capture of human performances. For the common case where only one actor is captured, most methods use strong prior knowledge on the deformation of the observed object in the form of articulated models. The works by Gall et al. [5,6] use silhouette and appearance information in a particle filtering framework to infer an optimal skeletal pose. Vlasic et al. [1] first optimize for the pose using the visual hull, then refine the shape estimate from the silhouettes. The works by Mundermann, Corraza et al. [3,7] use a variant of the ICP algorithm [8] to fit an articulated model to the visual hull. The more generic framework used by Aguiar et al. [9] relies on the preservation of Laplacian coordinates of a coarse tetrahedral mesh whose deformation is guided by silhouettes and photometric information. Skeletons on one side and the preservation of volume on the other showed to be priors strong enough for these algorithms to neglect the uncertainty in the input data. However, such strong deformation priors are no longer usable when dealing with objects of arbitrary nature.

To track surfaces in less constrained scenes, it is necessary to relax the deformation priors and thus to handle the noise in the input data. Treating the task as the registration of point sets is more generic but most of the non-rigid extensions to the ICP algorithm[8] lack robustness when confronted with outliers because of the determinism in the choice of point assignments. Among the recent approaches addressing the problem in a probabilistic framework, the works by Horaud et al. address articulated tracking [10] and the registration of rigid and articulated point

sets [11], while the *Coherent Point Drift* algorithm by Myronenko et al. [12] treats arbitrary deformations by regularizing the displacement field. These approaches all use the Expectation-Maximization algorithm to iteratively re-evaluate smooth assignments between the model and the data.

The method we present in this paper uses as input 3D data acquired with a multi-camera setup. It can handle complex scenes involving numerous objects of arbitrary nature by using generic surface deformation priors. It also handles the noise inherent to visual data acquisition by modeling the uncertainty in the observation process and by using the Expectation-Maximization algorithm. The following sections detail the algorithm.

3 Method

3.1 Parametrization and Deformation Framework

In the absence of prior knowledge on the nature of the observed surface, it is challenging to use noisy and sometimes incomplete information to infer meaningful measurements of motion and deformation. A possible way of establishing rigidity priors on the surface is to use the first mesh of a sequence as reference, and then to deform it across time to fit the observed data while penalizing locally non-rigid deformations with respect to its reference pose.

The framework presented in our previous work [2] does so by arbitrarily splitting the original geometry in surface elements called patches and by creating a corresponding coarser control structure in which the reference mesh is embedded. The idea is to regularly distribute patches of a maximal fixed geodesic radius on the surface and to associate to each patch P_k a rotation matrix $\mathbf{R_k}$ and the position of it's center of mass $\mathbf{c_k}$. These parameters encode a rigid transformation with respect to the world coordinates and allow for each vertex v whose position in the reference mesh was $\mathbf{x^0}(v)$ to define its new position as predicted by P_k as:

$$\mathbf{x_k}(v) = \mathbf{R_k}(\mathbf{x^0}(v) - \mathbf{c_k^0}) + \mathbf{c_k}. \tag{1}$$

This effectively decouple the parametrization of the deformation from the complexity of the original geometry. The deformed mesh is computed by linearly blending the predictions made by different patches for each vertex as given by Eq. 2. The weighting functions α_k are simply Gaussians of the euclidean distance to the center of mass of P_k and their support is the union of P_k and its neighbouring patches N_i. They are normalised to add up to 1.

$$\mathbf{x}(v) = \sum_k \alpha_k(v)\mathbf{x_k}(v). \tag{2}$$

3.2 Problem Formulation

Given a set of observed 3D points and an estimate of the current pose of the mesh, we are faced with a parameter estimation problem where the log-likelihood of the joint probability distribution of data and model must be maximized:

$$\max_{\Theta} \ \ln \ P(\mathcal{Y}, \Theta), \tag{3}$$

where:

- $\mathcal{Y} = \{y_i\}_{i=1:m}$ is the set of observed 3D points $\{\mathbf{y_i}\}_{i=1:m}$ and their normals.
- $\Theta = \{\mathbf{R_k}, \mathbf{c_k}\}_{k=1:N_p}$ are the parameters encoding the deformation.
- N_p is the number of patches.

We introduce prior knowledge on the range of possible shape deformations in the form of $E_r(\Theta) = -\ln P(\Theta)$. This energy is modelled by a simple term penalizing local non-rigid deformations of the surface with respect to a reference pose. It is directly linked to the patch-based representation and simply tries to enforce the predicted positions $\mathbf{x_k}(v)$ and $\mathbf{x_l}(v)$ of a vertex v by two neighbouring patches P_k and $P_l \in N_k$ to be consistent.

$$E_r(\Theta) = \frac{1}{2} \sum_{P_l} \sum_{P_k \in N_l} \left[\sum_{v \in P_k \cup P_l} (\alpha_k(v) + \alpha_l(v)) \|\mathbf{x_k}(v) - \mathbf{x_l}(v)\|^2 \right]. \quad (4)$$

Eq.3 can be rewritten using the fact that $P(\mathcal{Y}, \Theta) = P(\mathcal{Y}|\Theta)P(\Theta)$ and leads to solving the following optimization problem:

$$\min_{\Theta} E_r(\Theta) - \ln\ P(\mathcal{Y}|\Theta). \quad (5)$$

3.3 Bayesian Model

We approximate the pdf $P(\mathcal{Y}|\Theta)$ with a mixture of distributions parametrized by a common covariance σ^2, where each component corresponds to a patch. This requires to introduce latent variables z_i for each observation $y_i \in \mathcal{Y}$, where $z_i = k$ means that y_i was generated by the mixture component associated with P_k. We also increase the robustness of our model to outliers by introducing a uniform component in the mixture to handle points in the input data that could not be explained by the patches. This uniform component is supported on the scene's bounding box and we index it with $N_p + 1$.

$$P(y_i|\Theta, \sigma) = \sum_{k=1}^{N_p+1} \Pi_k P(y_i|z_i = k, \Theta, \sigma), \quad (6)$$

where the $\Pi_k = p(z_i = k|\Theta, \sigma)$ represent probabilities on the latent variables marginalized over all possible values of y_i. In other words they are prior probabilities on model-data assignments. We define them as constants $p(z_i = k)$ that add up to 1, using the expected proportion of outlier surface in the observations and the ratios of patch surfaces in the reference mesh.

The patch mixture component with index k must encode a distance between the position $\mathbf{y_i}$ and the patch P_k while accounting for the alignment of normals. For computational cost reasons, we model this distance by looking for each patch P_k in its different predicted poses (this means the positions $\{\mathbf{x_l}(v)\}_{l \in \{k\} \cup N_k, v \in P_k}$ and corresponding normals as shown in Fig. 1) for the closest vertex with a

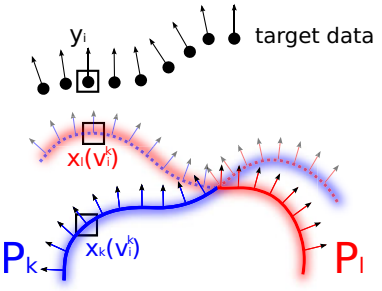


Fig. 1. A point/normal y_i with position $\mathbf{y_i}$ from the observed data is associated to v_i^k, the closest vertex with a compatible normal among all the predictions for the patch P_k. In this case v_i^k is selected because of its position and normal in the prediction made by the neighbouring patch P_l.

compatible normal v_i^k. We consider two points and normals to be compatible when their normals form an angle smaller than a threshold.

$$\forall k \in [1, N_p], \quad P(y_i|z_i = k, \Theta, \sigma) \sim \begin{cases} \mathcal{N}(\mathbf{y_i}|\mathbf{x}(v_i^k), \sigma) & \text{if } v_i^k \text{ exists} \\ \epsilon & \text{otherwise,} \end{cases} \tag{7}$$

where ϵ encodes for a negligible uniform distribution defined on the scene's bounding box.

3.4 Expectation-Maximization

The variables z_i can not be observed but we can use their posterior distributions (Eq. 8) in the EM algorithm first presented by Dempster et al.[4].

$$P(z_i = k|y_i, \Theta, \sigma) = \frac{\Pi_k P(y_i|z_i = k, \Theta, \sigma)}{\sum_{l=1}^{N_p+1} \Pi_l P(y_i|z_i = l, \Theta, \sigma)}. \tag{8}$$

The idea is to replace $P(\mathcal{Y}|\Theta, \sigma)$ with the marginalization over the hidden variables of the joint probability.

$$\ln P(\mathcal{Y}|\Theta, \sigma) = \ln \sum_Z q(Z) \frac{P(\mathcal{Y}, Z|\Theta, \sigma)}{q(Z)}, \tag{9}$$

where $q(Z)$ is a positive real valued function who sums up to 1. The concavity of the log function allows to write a bound on the function of interest:

$$-\ln P(\mathcal{Y}|\Theta, \sigma) \leq -\sum_Z q(Z) \ln \frac{P(\mathcal{Y}, Z|\Theta, \sigma)}{q(Z)}. \tag{10}$$

It can be shown that given a current estimate (Θ^t, σ^t), it is is optimal to choose $q(Z) = P(Z|Y, \Theta^t, \sigma^t)$ in that the bounding function then touches the bounded

function at (Θ^t, σ^t). This means that the bounding function should be the expected complete-data log-likelihood conditioned by the observed data:

$$-\ln P(\mathcal{Y}|\Theta,\sigma) \leq const - E_Z[\ln P(\mathcal{Y}, Z|\Theta,\sigma)|Y]. \tag{11}$$

We rewrite $P(\mathcal{Y}, Z|\Theta,\sigma)$ by making the approximation that the observation process draws the y_i's in $\mathcal{Y}$ from the distribution in an independent identically distributed way:

$$P(\mathcal{Y}, Z|\Theta,\sigma) = \prod_{i=1}^{m} P(y_i, z_i|\Theta,\sigma) \tag{12}$$

$$= \prod_{k=1}^{N_p+1} \prod_{i=1}^{m} \left[P(y_i, z_i = k|\Theta,\sigma)\right]^{\delta_k(z_i)}. \tag{13}$$

The choice made for $q(z)$ then allows to write:

$$-\ln P(\mathcal{Y}|\Theta,\sigma) \leq const - \sum_{k=1}^{N_p+1} \sum_{i=1}^{m} P(z_i = k|y_i, \Theta^t, \sigma^t) \ln P(y_i|z_i = k, \Theta, \sigma). \tag{14}$$

We use the Expectation-Maximization algorithm to iteratively re-evaluate the (Θ, σ) and the posterior probability distributions on the latent variables $\{z_i\}$.

In the E - Step step the posterior $P(z_i|y_i, \Theta^t, \sigma^t)$ functions are evaluated using the current estimation Θ^t, σ^t and the corresponding predicted local deformations of the mesh. They represent weights in the soft assignments of the data to the model. The process amounts to the computation of a $m \times (N_p + 1)$ matrix whose lines add up to 1. This is an extremely parallel operation as all the elements of this matrix can be evaluated independently, except for the normalization step that has to be done by line.

The M - Step requires to minimize the bounding function obtained by evaluating the data-model assignment weights in the E-Step:

$$\Theta^{t+1}, \sigma^{t+1} = \text{argmin} \left[const + E_r(\Theta) - \sum_{k=1}^{N_p+1} \sum_{i=1}^{m} P(z_i = k|y_i, \Theta^t, \sigma^t) \ln P(y_i|z_i = k, \Theta, \sigma) \right] \tag{15}$$

In this bounding function, both data terms and rigidity terms are squared distances between 3D points. Instead of completely minimizing the bounding function, we just run one iteration of the Gauss-Newton algorithm, which amounts to minimizing the quadratic approximation of the objective function around Θ^t.

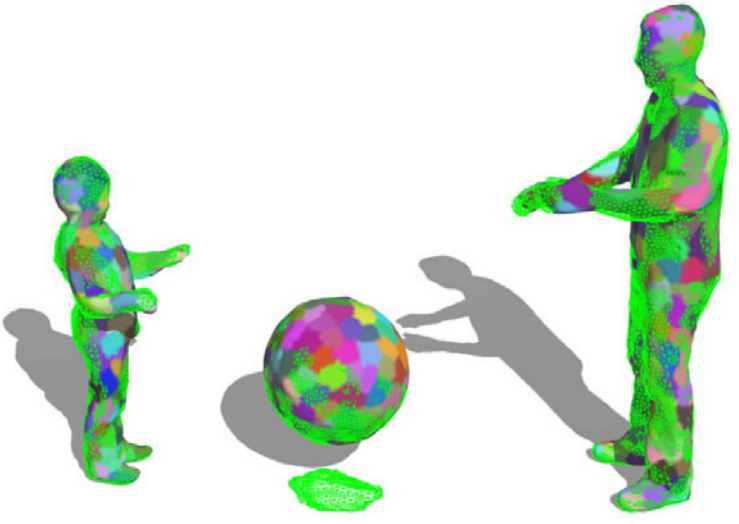


Fig. 2. Ball Sequence involving multiple objects. Note that the wrong geometry on the floor, coming from the shadows, does not affect the tracking results. It is classified as outlier by EM and the ball is not attracted to it.

4 Results

4.1 Multi-object Tracking and Outlier Rejection

The algorithm presented in more generic than the available state of the art methods and allows to track surfaces in complex scenes. We show our results on two of these sequences to demonstrate the clear advantages of our approach. We also provide timing estimates in Table 1 to give a rough idea of its computational complexity.

Ball Sequence. The first of these sequence is the *ball* dataset from INRIA-Perception. It consists of 275 photo-consistent meshes. It involves three distinct object and can not be treated with articulated models. The significant overlap in the silhouettes makes it necessary to run a 3D reconstruction and use point clouds as input data to reduce ambiguity. In Figure 2 we show a particularly difficult frame in which the wrong segmentation of shadows in the original images has resulted in the creation of outlying geometry. The data term presented in [2] does not account for this possibility and simply tries to minimize the distance between two point clouds. Our approach in contrast handles the outlying geometry by progressively reducing its weight in the function optimized by the M-Step.

BasketBall Sequence. We recorded the Basketball sequence in our own multi-camera studio. It is 1364 frames (about 55sec) long and consists of meshes independently reconstructed by a voxel-carving method. It displays a basketball

player dribbling a ball. The interactions between the two objects are fast and complex as the ball bounces between the legs and is sometimes held close to the body for many frames. The results presented in Figure 3 and the accompanying video show two things : firstly, our algorithm can recover these difficult motions and deformations. Secondly, it can cope with the numerous artefacts in the input data : missing limbs, occlusions and self intersecting geometry.

4.2 Human Performance Capture

We also ran our algorithm on standard datasets available to the community to compare it to previous works. We used as input the results of a precise 3D reconstruction algorithm in one case, and noisy voxel carving in the other. As we show in this section, our algorithm performs consistently well in both these situations.

Tracking Using Photo-consistent Meshes As Input. The Surfcap Data from University of Surrey consists of a series of temporally inconsistent meshes obtained by the photo-consistency driven graph-cut method of Starck et al.[13]. Except for some rare reconstruction artefacts in a couple of frames, these are overall very clean and smooth meshes. Because of their extremely high resolution, these meshes were down-sampled to roughly 10k vertices and fed to our algorithm. We present in this paper and the associated video our results on six sequences. They show a hip-hop dancer whose moves are very challenging to track because they contain fast motions and large deformations. In Figure 4, our results on the *Flashkick* dataset show that we can cope with extremely fast deformations such as a backflip. In Figure 5 we present our results on the *Pop* sequence in which the intricate and ambiguous motion of crossing arms is handled properly. Additionally Figure 7 shows a quantitative evaluation of the overlap error between the reprojected silhouettes from our result and the original silhouettes. The error is given as the ratio of erroneous pixels and total number of pixels in the original silhouette. In the presented results we performed an additional optimization that minimizes this reprojection error and keeps it approximately at a constant value of 5%.

Tracking Using Voxel Carving As Input. We used the multi-view image data made public by the MIT CSAIL group to run a very simple voxel carving algorithm. The resulting visual hulls, although only a coarse approximation of the true shape, were enough to drive the deformation of the provided template mesh through the sequences. We ran our algorithm on four of the available sequences and refined the result using silhouette fitting. We compared the silhouette reprojection error to the meshes obtained by Vlasic et al. in [1] and display our results in Figure 8. We also show our results after silhouette fitting on the *Samba* dataset. In this specific sequence, a woman in a skirt dances. Skirts are difficult to handle for methods deforming a reference mesh as the interpolated surface between the bottom of the skirt and the legs does not exist and has to undergo severe compression and stretching. We show in Figure 6 that our approaches still manages to produce visually convincing results.

Fig. 3. Results on the Basketball Sequence. Note that wrong geometry, missing data and fast motion have a limited impact on our tracking algorithm.

Fig. 4. The Flashkick sequence exhibits very fast motion

Fig. 5. The Pop sequence involves a very ambiguous situation when the arms cross

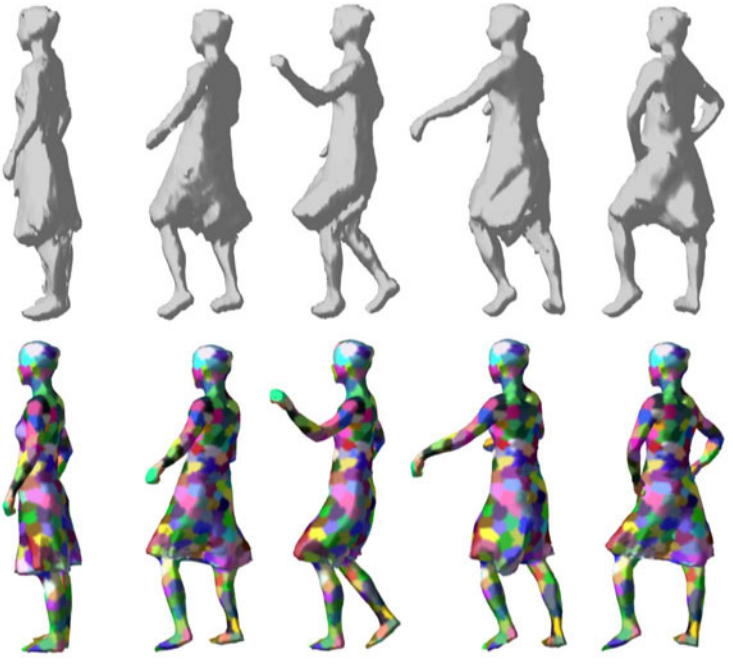
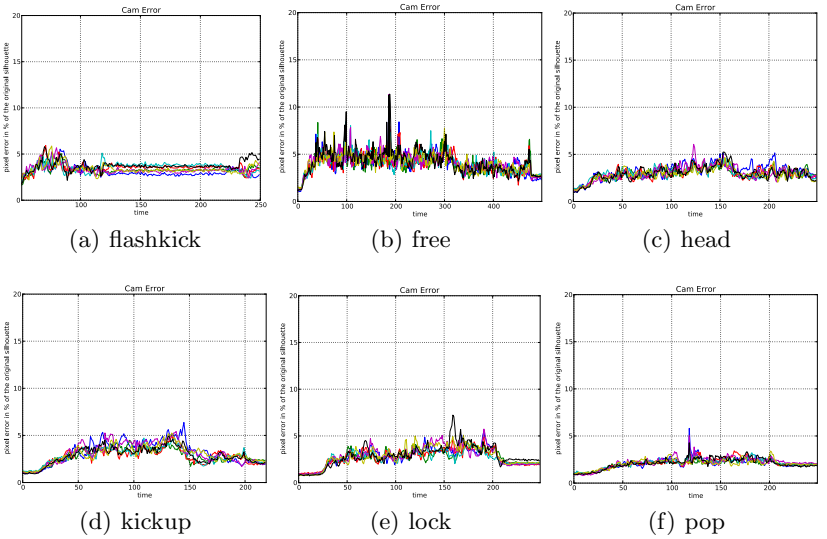


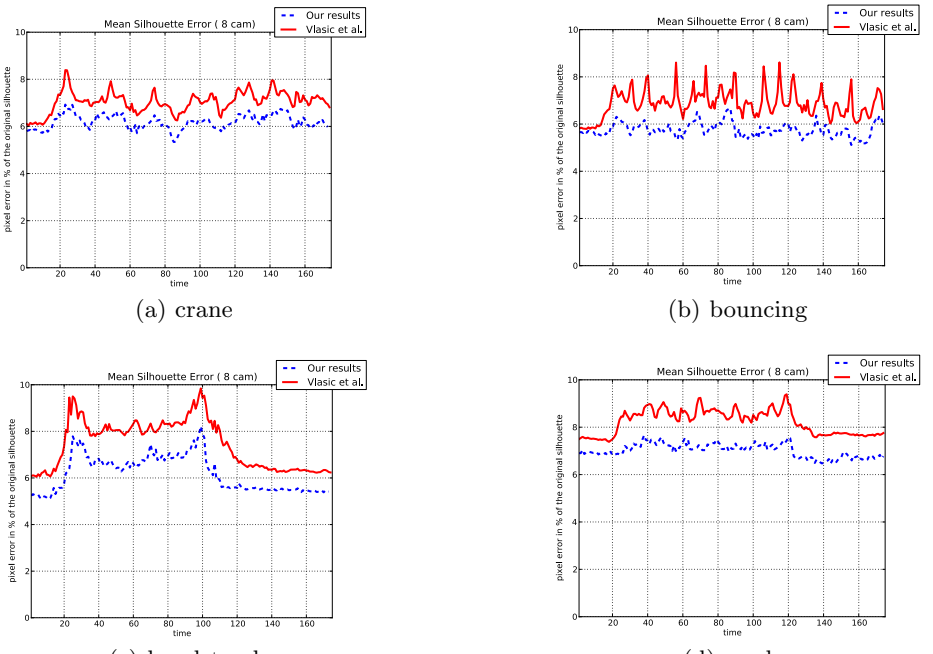
Fig. 6. Results on the Samba sequence show the tracking of a skirt using visual hull reconstructions



(a) flashkick (b) free (c) head

(d) kickup (e) lock (f) pop

Fig. 7. Silhouette reprojection error of our deformed model in percentage of the original silhouette area. Each color represents a camera.



(a) crane (b) bouncing

(c) handstand (d) samba

Fig. 8. Comparison of our numerical results with the method of Vlasic et al.[1]. Although we perform numerically better, it should be noted that their results are temporally smoothed, which can explain the difference in performance.

Table 1. Average timings on standard sequences for the EM procedure (without silhouette refinement), obtained on a 2.5Ghz quad-core machine with target point clouds of roughly 10k vertices. These measurements were obtained by looking at times when files were written to the hard-drive and do not constitute a precise performance evaluation. However they give a rough idea of the computational complexity of our method.

Sequence	Length	Reference Mesh Vertex Count	Average Time Per Frame
Flashkick	200	5445	24 sec
Free	500	4284	25 sec
Head	250	5548	29 sec
Kickup	220	5580	23 sec
Lock	250	5301	24 sec
Pop	250	5596	16 sec
Handstand	174	5939	29 sec
Bouncing	174	3848	29 sec
Crane	174	3407	11 sec
Samba	150	5530	12 sec

5 Discussion

The prediction mechanism for neighbouring patches in the computation of associations described in subsection 3.3 is the key to our method, as it encodes for multiple hypothesis on the position of the patch. More specifically, it gives a chance to the surface to locally quickly return to its rest pose by propagating the information from correctly registered parts of the mesh to parts where the current approximation of the deformation is erroneous.

Topology changes. Although this framework assumes very little on the nature of the tracked objects, it can not handle variations in the topological nature of the reference surface. The reference frame has to be topologically suitable, that is it has to be split wherever the surface might split during the sequence. In other terms, a small amount of geometry disappearance (self-intersection) can be handled, but there can't be any creation of geometry.

The i.i.d. assumption can be considered as problematic in that the observation process is a multi-camera setup in which parts of the surface, thus patches occlude each other. This clearly biases the drawing of samples in the distribution of 3D data. For example in Figure 3, when the arms and body are joined, the local density of points in the input data doesn't double, which clearly indicates that the data generation by two overlapping patches on the arm and the body is not independent. In that sense our method and Equation 12 are only approximations.

6 Conclusion

We proposed a probabilistic method for temporal mesh deformation which can effectively cope with noisy and missing data. We deform a reference mesh and fit it to independently reconstructed geometry obtained from multiple cameras. The imperfection of background segmentation and reconstruction algorithms results in the creation of wrong or missing geometry. Using generic local rigidity priors on the tracked surface, we propose a Bayesian framework which takes into account uncertainties of the acquisition process. We perform a maximum-likelihood estimation where the joint probability of the deformation parameters and the observed data is maximized using the Expectation-Maximization algorithm. We showed on a large number of multi-view sequences that our method is robust to reconstruction artefacts and numerically as precise as state of the art methods based on skeletal priors. Moreover, this effectiveness is achieved with a much more generic deformation model that allows to process complex sequences involving several objects of unknown nature.

References

1. Vlasic, D., Baran, I., Matusik, W., Popović, J.: Articulated mesh animation from multi-view silhouettes. In: ACM SIGGRAPH 2008 (2008)
2. Cagniart, C., Boyer, E., Ilic, S.: Free-from mesh tracking: a patch-based approach. In: IEEE CVPR (2010)

3. Mundermann, L., Corazza, S., Andriacchi, T.P.: Accurately measuring human movement using articulated icp with soft-joint constraints and a repository of articulated models. In: CVPR (2007)
4. 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)
5. Gall, J., Stoll, C., de Aguiar, E., Theobalt, C., Rosenhahn, B., Seidel, H.P.: Motion capture using joint skeleton tracking and surface estimation. In: IEEE CVPR 2009 (2009)
6. Gall, J., Rosenhahn, B., Brox, T., Seidel, H.P.: Optimization and filtering for human motion capture. IJCV 87 (2010)
7. Corazza, S., Mündermann, L., Gambaretto, E., Ferrigno, G., Andriacchi, T.P.: Markerless motion capture through visual hull, articulated icp and subject specific model generation. IJCV 87 (2010)
8. Besl, P.J., McKay, N.D.: A method for registration of 3-d shapes. IEEE PAMI 14 (1992)
9. de Aguiar, E., Stoll, C., Theobalt, C., Ahmed, N., Seidel, H.P., Thrun, S.: Performance capture from sparse multi-view video. In: ACM SIGGRAPH 2008 (2008)
10. Horaud, R.P., Niskanen, M., Dewaele, G., Boyer, E.: Human motion tracking by registering an articulated surface to 3-d points and normals. IEEE PAMI 31 (2009)
11. Horaud, R.P., Forbes, F., Yguel, M., Dewaele, G., Zhang, J.: Rigid and articulated point registration with expectation conditional maximization. IEEE PAMI (2010)
12. Myronenko, A., Song, X.: Point-set registration: Coherent point drift. IEEE PAMI (2010)
13. Starck, J., Hilton, A.: Surface capture for performance based animation. In: IEEE Computer Graphics and Applications 27(3) (2007)