Volumetric 3D Tracking by Detection Supplementary Material

Chun-Hao Huang *[1], Benjamin Allain *[2], Jean-Sébastien Franco[2], Nassir Navab[1], Slobodan Ilic[1], and Edmond Boyer[2]

[1]Technische Universität München
[2]Inria, LJK, Univ. Grenoble Alpes
{huangc,ilics,navab}@in.tum.de, {firstname.lastname}@inria.fr

The supplementary material for the paper *Volumetric 3D Tracking by Detection* consists of this document and the accompanying video. It provides more quantitative tracking results, discussions, and a table of notations (Table 1) used in the main paper.

1. Probabilistic inverse mapping g^{-1}

In § 4.1.3 of the main paper, we align the topologies of different templates so that forests are learned based in a topology-consistent domain. To this end, we utilize skinning weights $\mathbf{w}$ and develop a mapping g that maps each cell s on a subject-specific template $\mathcal{S}^\mu$ to a cell $\hat{s}$ on the common template $\hat{\mathcal{S}}$:

$$g^\mu(s) = \arg\min_{\hat{s}\in\hat{\mathcal{S}}} \|\mathbf{w}_{\hat{s}} - \mathbf{w}_s\|_2, \forall s \in \mathcal{S}^\mu. \qquad (1)$$

During tracking, forests predict the correspondences that lie on the generic template $\hat{\mathcal{S}}$ and one has to revert it back to the cell index on the subject-specific template $\mathcal{S}^\mu$.

We assume users know the tracking subject, *i.e.* μ is known. Since g^μ is constructed by nearest neighbor search, leading to a many-to-one function, the inverse mapping $(g^\mu)^{-1}$ is by nature ill-defined. We therefore resort to a probabilistic formulation. Specifically, given a predicted cell $\hat{s}$ on the common template $\hat{\mathcal{S}}$, all the possible $s \in \mathcal{S}^\mu$ being mapped to $\hat{s}$ are taken into account. When they are used to construct the least-square-based energy formulations, we weight them differently according to their distances to $\hat{s}$ in the skinning-weight space. This strategy fits naturally to the EM-ICP framework presented in § 4.2 of the main paper.

*The first two authors contribute equally to this paper.

2. Quantitative tracking evaluation

We also evaluate our tracking approach with two different metrics. On one hand, evaluation with marker-based motion capture evaluates the correctness of the surface pose, but only for a sparse set of surface points. On the other hand, the silhouette overlap error evaluates the shape estimate, but it does not evaluate the estimated pose. Hence these metrics are complementary.

2.1. Marker-based motion capture

The *Ballet/Seq2* sequence has marker-based motion capture data: fifty markers were attached to the body of the subject. The 3D tracking of the markers provides a sparse ground truth for surface tracking. First, each marker is associated to a surface vertex of the template. Then, for each marker, in each temporal frame, we measure the distance between the marker location and the estimated vertex location. Statistics on the distance are reported on Table 2. We observe that our approach gives slightly better performances than a state of the art ICP-based approach, and outperforms a learning-based tracking approach which mostly fails to correctly register the legs of the subject.

method	mean (mm)	stddev. (mm)
Proposed	26.37	16.67
Huang *et al.* [2]	124.02	200.16
Allain *et al.* [1]	27.82	18.39

Table 2: Statistics of surface registration error at marker locations, on the *Ballet/Seq2* sequence.

2.2. Silhouette overlap error

We evaluate the tracking approach by computing the overlap error between the ground truth silhouette and the

notations	descriptions
Ω	A volumetric domain; $\Omega \subset \mathbb{R}^3$.
$\partial\Omega$	A surface (the border of a volume).
$\mathcal{S}$	A set of indices representing CVT cells/centroids.
$(\cdot)_M, (\cdot)_Y$	Subscripts representing variables of templates (M) and observations (Y) respectively.
s, i	Indices of CVT cells on templates (s) and observations (i); $s \in \mathcal{S}_M, i \in \mathcal{S}_Y$.
$\mathbf{M}, \mathbf{Y}$	Sets of 3D locations of CVT centroids; $\mathbf{M} \subset \Omega_M, \mathbf{Y} \subset \Omega_Y$. Note that locations $\mathbf{M}, \mathbf{Y}$ and index sets $\mathcal{S}_Y$ are time dependent variables, while index set $\mathcal{S}_M$ is constant during tracking.
$\mathbf{x}$	3D locations of one CVT centroid on the template; $\mathbf{x}_s \in \mathbf{M}, s \in \mathcal{S}_M$.
K	Number of clusters of CVT cells; $K = 150$ for *Ballet* and *Goalkeeper*; $K = 250$ for *Thomas*.
L	Number of layers clusters for feature computation; $L = 8$.
B	Number of bones for skinning-weight computation; $B = 17$.
μ	Denoting $\mathcal{S}$ of different templates; $\mu = 1 \cdots U$.

Table 1: Notations and the setting of parameters.

projection of the estimated surface. The metric we use is the pixel error (number of pixels that differ). Statistics are computed on all frames of all cameras.

method	mean	stddev.	median	max
Proposed	15221	6843	14754	57748
Huang *et al.* [2]	19838	14260	15607	109428
Allain *et al.* [1]	14773	6378	14355	43359

Table 3: Silhouette pixel error on sequence *Goalkeeper/UpJump*. Image size is 2048×2048.

method	mean	stddev.	median	max
Proposed	2620	1041	2557	8967
Huang *et al.* [2]	5427	2809	4863	39559
Allain *et al.* [1]	2606	1008	2571	7642

Table 4: Silhouette pixel error on sequence *Ballet/Seq2*. Image size is 1920×1080.

method	mean	stddev.	median	max
Proposed	9991	7089	7968	78242
Huang *et al.* [2]	28731	23421	22991	354293
Allain *et al.* [1]	10199	7379	8022	81649

Table 5: Silhouette pixel error on sequence *Thomas/Seq2*. Image size is 2048×2048.

Discussion. As discussed in § 5.1 of the main paper, the high memory footprint of voxel-based volume in [2] limits the allowed training variations. Consequently, they choose to align the orientations for both training and input data such that forests only need to learn the pose variations of one single subject. They rely on the skeletal poses of previous frames to re-orient the input data of the current frame. This leads to not fully frame-independent forest predictions and makes tracking subject to the potential risk of drifting. On the other hand, our approach attempts to incorporate rotational, pose, and even shape variations during training, yielding completely frame-wise forest predictions. To facilitate a fully 3D tracking-by-detection framework, the information of previous frames is preferred no to participate in the discriminative correspondence estimation.

Therefore, we do not re-orient meshes when implementing our method and [2], to draw fair comparisons. As reported in Fig. 5 of the main paper, without canceling rotational variations, the accuracies of correspondences drop substantially on the testing sequences for the method in [2]. This means that voxel-based framework and the corresponding features do not generalize well to unseen rotations. When deployed in tracking applications, such unreliable associations eventually result in tracking failure. In particular, one can observe in Table 5 that [2] attains really high silhouette overlap discrepancy, due to the fact that the subject rotates himself in many different orientations and thus confuses the forest.

References

[1] B. Allain, J.-S. Franco, and E. Boyer. An efficient volumetric framework for shape tracking. In *CVPR*. IEEE, 2015.

[2] C.-H. Huang, E. Boyer, B. do Canto Angonese, N. Navab, and S. Ilic. Toward user-specific tracking by detection of human shapes in multi-cameras. In *CVPR*. IEEE, June 2015.