Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input

Srinath Sridhar[1] Franziska Mueller[1] Michael Zollhöfer[1]
Dan Casas[1] Antti Oulasvirta[2] Christian Theobalt[1]

[1]Max Planck Institute for Informatics [2]Aalto University
{ssridhar,frmueller,mzollhoef,dcasas,theobalt}@mpi-inf.mpg.de
{antti.oulasvirta}@aalto.fi

Supplementary Document

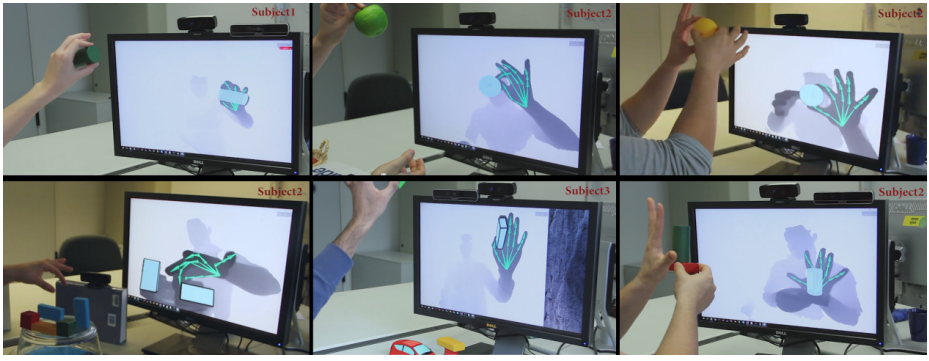

Fig. 1. Live tracking results for three different subjects.

In this document we take a deeper look at our articulated Gaussian mixture alignment strategy and show more qualitative results of our live capture setup that allows to track hand-object interactions at frame rate. In addition, we provide details on our benchmark dataset and the error metric used in the ground truth evaluation. Finally, we give the gradients of all components of our objective function. For further results, i.e. influence of the different components and video footage of live tracking sessions, we refer to the supplemental video.

1 Alignment Objective

In this section, we take a deeper look at the design of our alignment objective E_a and explore its connection to point set registration methods that are based on Gaussian mixtures [1]. Note, the alignment objective is just a small component of our complete energy function that also includes novel contact and occlusion handling constraints. Let us assume the model as well as the input depth data

are represented each as a Gaussian mixture:

$$\mathcal{M}(\mathbf{x}) = \sum_{i \in \mathbf{M}} w_i \mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_i, \sigma_i),\ \mathcal{I}(\mathbf{x}) = \sum_{i \in \mathbf{I}} w_i \mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_i, \sigma_i)\ .$$

Here, the set $\mathbf{M}$ contains the indices of all model Gaussians and the set $\mathbf{I}$ of all image Gaussians, respectively. Each Gaussian is isotropic with standard deviation $\sigma_i \in \mathbb{R}$ and mean $\boldsymbol{\mu}_i \in \mathbb{R}^3$. For simplicity let us assume all mixing weights to be one ($w_i = 1$). We then define an ℓ_2-dissimilarity measure between the two Gaussian mixtures, also see [1] for more details:

$$E_a = \int_\Omega \left[\mathcal{M}(\mathbf{x}) - \mathcal{I}(\mathbf{x})\right]^2 d\mathbf{x}\ .$$

The expansion of Equation 1 splits the objective in three distinct parts:

$$\begin{aligned} E_a &= \int_\Omega \left[\mathcal{M}(\mathbf{x}) - \mathcal{I}(\mathbf{x})\right]^2 d\mathbf{x} \\ &= \int_\Omega \left[\mathcal{M}(\mathbf{x})^2 - 2\mathcal{M}(\mathbf{x})\mathcal{I}(\mathbf{x}) + \mathcal{I}(\mathbf{x})^2\right] d\mathbf{x} \\ &= \underbrace{\int_\Omega \mathcal{M}(\mathbf{x})^2\ d\mathbf{x}}_{(a)} - 2\underbrace{\int_\Omega \mathcal{M}(\mathbf{x})\mathcal{I}(\mathbf{x}) d\mathbf{x}}_{(b)} + \underbrace{\int_\Omega \mathcal{I}(\mathbf{x})^2 d\mathbf{x}}_{(c)}\ . \end{aligned}$$

Note, (c) is constant in the presented tracking scenario, since we only optimize for the positions of the model Gaussians. The terms (a) and (b) are integrals over products of Gaussian Mixtures. Let us first consider (b):

$$\begin{aligned} \int_\Omega \mathcal{M}(\mathbf{x})\mathcal{I}(\mathbf{x}) d\mathbf{x} &= \int_\Omega \Big(\sum_{i \in \mathbf{M}} \mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_i, \sigma_i)\Big)\Big(\sum_{j \in \mathbf{I}} \mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_j, \sigma_j)\Big) d\mathbf{x} \\ &= \int_\Omega \Big[\sum_{i \in \mathbf{M}} \sum_{j \in \mathbf{I}} \mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_i, \sigma_i)\mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_j, \sigma_j)\Big] d\mathbf{x} \\ &= \sum_{i \in \mathbf{M}} \sum_{j \in \mathbf{I}} \underbrace{\Big[\int_\Omega \mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_i, \sigma_i)\mathcal{G}(\mathbf{x}|\boldsymbol{\mu}_j, \sigma_j)\Big] d\mathbf{x}}_{\mathcal{S}_{i,j}}\ . \end{aligned}$$

Since $\mathcal{S}_{i,j}$ is the integral over a product of Gaussians, it has a closed form expression [2]:

$$\mathcal{S}_{i,j} = \frac{(2\pi)^{\frac{3}{2}}(\sigma_i^2\sigma_j^2)^{\frac{3}{2}}}{(\sigma_i^2 + \sigma_j^2)^{\frac{3}{2}}} \exp\left(-\frac{||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||_2^2}{2(\sigma_i^2 + \sigma_j^2)}\right)\ .$$

Its gradient can be easily derived in closed form; the same holds for (a).

Fig. 2. We are able to track complex shapes like a toy car. Our contact points term (contacts are circled in blue) makes fingers hold the car even in the presence of severe occlusion.

2 Live Tracking Results

Our real-time approach uses the color and depth data from a single *Creative Senz3D* time-of-flight (TOF) sensor. Note, we also support other depth sensors like the *Intel RealSense*, *Kinect* and *Primesense Carmine*. The used color and depth resolutions are 640×480 and 320×240, both captured at 30 Hz. We show compelling live tracking results for three different subjects in a close interaction range of 15 to 100 cm away from the camera, see Fig. 1. In addition, Fig. 2 presents a tracking result of a complex object (toy car). Tracking is robust even if hands closely interact with objects due to the proposed contact and occlusion constraints. Our approach is robust even if a second hand is visible. This enables interesting and new interaction possibilities as shown in Fig. 1. For additional live footage, we refer to the supplemental video.

3 Error Measure

We provide a new benchmark with 3014 frames (6 sequences) with ground truth annotations to evaluate hand-object tracking methods, see Fig 3. For each frame, we annotated 8 distinct landmarks (5 fingertip positions and 3 corners of the object). If a location is not visible, the corresponding landmark is set to be invalid and is not considered in the error measure. For the object (cuboid), the 3 landmarks span a coordinate system along the cuboid's two dominant axes. This uniquely defines the cuboid with respect to an axis of symmetry. For evaluation, we employ the following error metric to compare our tracking results with the ground truth annotations:

$$E = \frac{1}{|\mathcal{V}| + \mathbb{1}_{\mathcal{M}}} \left[\sum_{i \in \mathcal{V}} ||X_i - G_i|| + \frac{\mathbb{1}_{\mathcal{M}}}{3} \sum_{m \in \mathcal{M}} ||X_m - G_m|| \right],$$

where $\mathcal{V}$ denotes the set of all un-occluded fingertip positions in the ground truth, $\mathcal{M}$ denotes *matched* cuboid corners, and X and G denote estimated and ground truth positions, respectively. The indicator function $\mathbb{1}_{\mathcal{M}}$ is 1 if $|\mathcal{M}| = 3$ and 0 otherwise. Fingertip positions are compared with the corresponding landmarks based on the distance in 3D Euclidean space. To this end, the 2D annotations are back-projected based on depth and inverse camera intrinsics. *Matched* cuboid

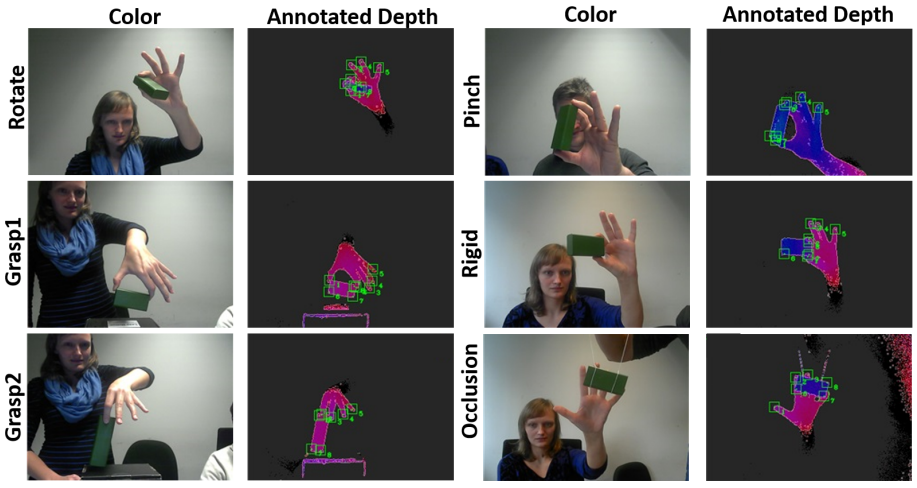

Fig. 3. The six sequences of our novel ground-truth hand-object benchmark.

corners refers to corners in the estimated cuboid that are closest to the ground truth. If one of the cuboid corners is occluded, then the set $\mathcal{M}$ is empty as the cuboid cannot be uniquely positioned.

4 Gradients

Here, we give analytical expressions for the gradients of all energy terms. The used mathematical notation is defined in the main document.

Spatial Alignment Term E_a:

$$\frac{\partial E_a}{\partial x_k} = \sum_{i \in \mathbf{M}} \sum_{j \in \mathbf{M}} \left[\mathcal{S}_{i,j} \cdot \left(- \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma_i^2 + \sigma_j^2} \right) \cdot \left(\frac{\partial \boldsymbol{\mu}_i}{\partial x_k} - \frac{\partial \boldsymbol{\mu}_j}{\partial x_k} \right) \right] - 2 \cdot \sum_{i \in \mathbf{M}} \sum_{j \in \mathbf{I}} \left[\mathcal{S}_{i,j} \cdot \left(- \frac{\boldsymbol{\mu}_i - \boldsymbol{\mu}_j}{\sigma_i^2 + \sigma_j^2} \right) \cdot \frac{\partial \boldsymbol{\mu}_i}{\partial x_k} \right] .$$

Semantic Alignment Term E_s:

$$\frac{\partial E_s}{\partial x_k} = 2 \cdot \sum_{i \in \mathbf{M}} \sum_{j \in \mathbf{I}} \alpha_{i,j} \cdot (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) \cdot \frac{\partial \boldsymbol{\mu}_i}{\partial x_k} .$$

Anatomical Plausibility Regularizer E_p:

$$\frac{\partial E_p}{\partial x_k} = \begin{cases} 0 & \text{if } x_k^l \leq x_k \leq x_k^u \\ 2 \cdot (x_k - x_k^u) & \text{if } x_k > x_k^u \\ 2 \cdot (x_k - x_k^l) & \text{if } x_k < x_k^l \end{cases} .$$

Temporal Smoothness Regularizer E_t:

$$\frac{\partial E_t}{\partial x_k} = 2 \cdot (x_k^{(t)} - 2x_k^{(t-1)} + x_k^{(t-2)}) \ .$$

Contact Points Term E_c:

$$\frac{\partial E_c}{\partial x_k} = \sum_{(j,l,t_d)\in\mathcal{T}} 4 \cdot (||\boldsymbol{\mu}_j - \boldsymbol{\mu}_l||_2^2 - t_d^2) \cdot (\boldsymbol{\mu}_j - \boldsymbol{\mu}_l) \cdot \Big(\frac{\partial \boldsymbol{\mu}_j}{\partial x_k} - \frac{\partial \boldsymbol{\mu}_l}{\partial x_k}\Big) \ .$$

Object Occlusion Term E_o:

$$\frac{\partial E_o}{\partial x_k} = 2 \cdot \sum_{i\in\mathcal{H}_i} (1 - \hat{f}_i) \cdot (x_k - x_k^{old}) \ .$$

References

1. Jian, B., Vemuri, B.C.: Robust point set registration using gaussian mixture models. Pattern Analysis and Machine Intelligence, IEEE Transactions on 33(8), 1633–1645 (2011)
2. Stoll, C., Hasler, N., Gall, J., Seidel, H., Theobalt, C.: Fast articulated motion tracking using a sums of gaussians body model. In: Proc. IEEE ICCV. pp. 951–958 (2011)