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

[1] max planck institut informatik

[2] Aalto University School of Electrical Engineering

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

ECCV'16
EUROPEAN CONFERENCE ON COMPUTER VISION

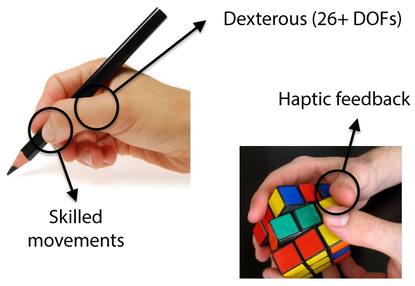
Tangible Interaction for AR/VR

AR/VR devices need **new** forms of **input**.

We need to **increase** the **expressiveness** of interaction beyond controllers and simple gestures.

Joint Hand-Object Tracking

ADVANTAGES

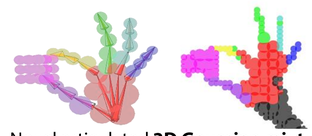


CHALLENGES

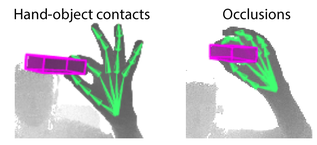
- **Extreme occlusions** by objects
- **Complex** and fast **motions**
- **Segmentation** of hand from object
- **High dimensional** problem
- **Run time** constraint

Contributions

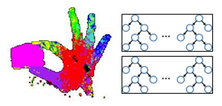
NOVEL FRAMEWORK FOR REAL-TIME HAND-OBJECT TRACKING



Novel articulated **3D Gaussian mixture alignment** for pose optimization



Energy terms **motivated** by physics of **grasps**

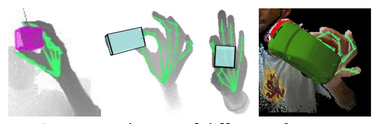


Multi-layered random forests for hand-object **segmentation**

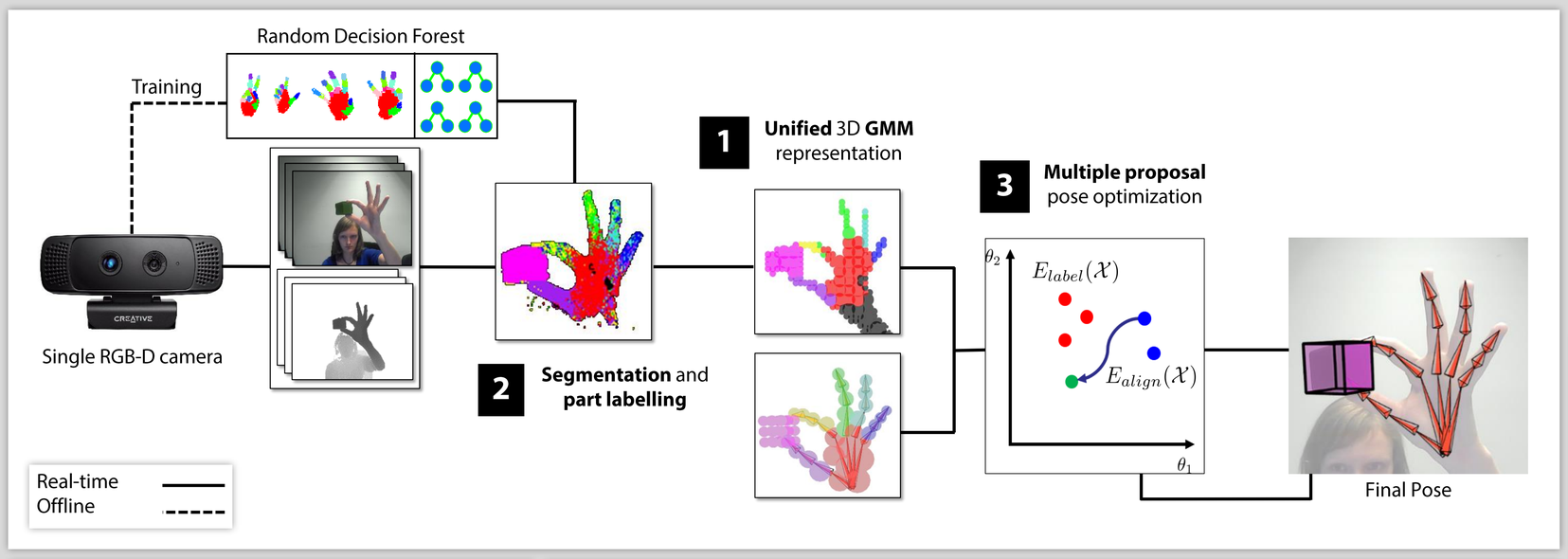
PRACTICAL

Single RGB-D **camera**

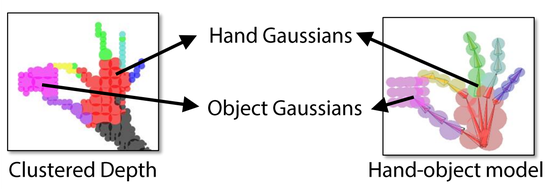
Real-time @ **30 FPS**



Supports objects of different **shapes, sizes**, and **colors**.

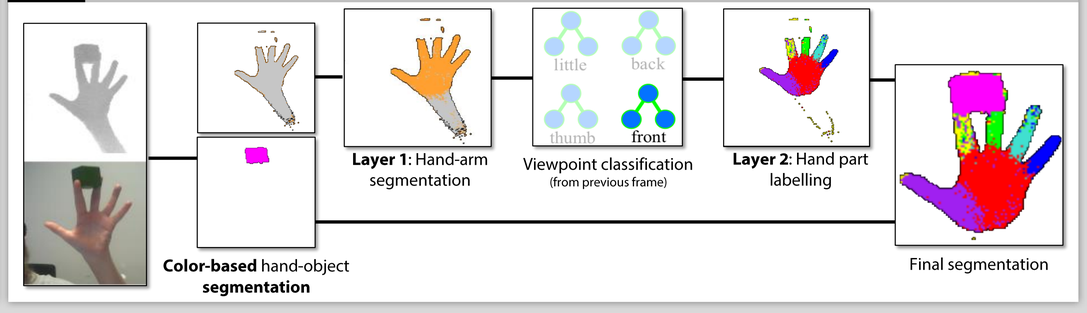


1 GMM Representation



$$\mathcal{M}(\mathbf{x}) = \sum_{i=1}^{K} w_i \mathcal{G}_i(\mathbf{x}; \sigma_i, \mu_i),\ \mathbf{x} \in \mathcal{R}^3$$

2 Segmentation and Part Labelling



3 Multiple Proposal Pose Optimization

OBJECTIVE 1:

$$E_{align}(\mathcal{X}) = \overbrace{E_a}^{\text{Spatial alignment}} + \overbrace{w_p\, E_p}^{\text{Anatomical plausibility}} + \overbrace{w_t\, E_t}^{\text{Temporal smoothness}} + \overbrace{w_c\, E_c + w_o\, E_o}^{\text{Contact and occlusion terms}}$$

OBJECTIVE 2:

$$E_{label}(\mathcal{X}) = \overbrace{E_a}^{\text{Spatial alignment}} + \overbrace{w_s\, E_s}^{\text{Semantic alignment}} + \overbrace{w_p\, E_p}^{\text{Anatomical plausibility}}$$

SPATIAL ALIGNMENT

$$E_a(\mathcal{X}) = \int_{\Omega} \Big[\big(\mathcal{M}_{d_h}(\mathbf{x}) - \mathcal{M}_h(\mathbf{x})\big)^2 + \big(\mathcal{M}_{d_o}(\mathbf{x}) - \mathcal{M}_o(\mathbf{x})\big)^2 \Big] d\mathbf{x}$$

Aligns model and input 3D Gaussian mixtures

SEMANTIC ALIGNMENT

$$E_s(\mathcal{X}) = \sum_{i=1}^{N_s} \sum_{j=1}^{N_d} \alpha_{i,j} \cdot ||\boldsymbol{\mu}_i - \boldsymbol{\mu}_j||_2^2$$

Aligns GMMs with semantic label information

CONTACT POINTS

$$E_c(\mathcal{X}) = \sum_{(k,l,t_d) \in \mathcal{T}} \left(||\boldsymbol{\mu}_k - \boldsymbol{\mu}_l||^2 - t_d^2 \right)^2$$

Enforces contact between fingertip and object Gaussians that are close

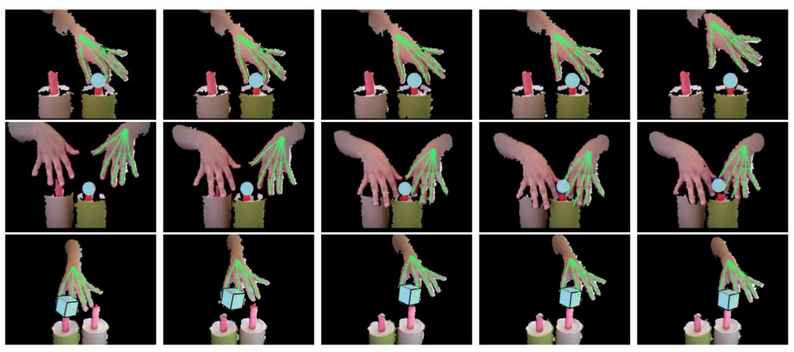
OCCLUSION HANDLING

$$E_o(\mathcal{X}) = \sum_{i=0}^{N_h} \sum_{j \in \mathcal{H}_i} (1 - \hat{f}_i) \cdot ||x_j - x_j^{old}||_2^2$$

Forces occluded parts of the hand to move consistently with other parts

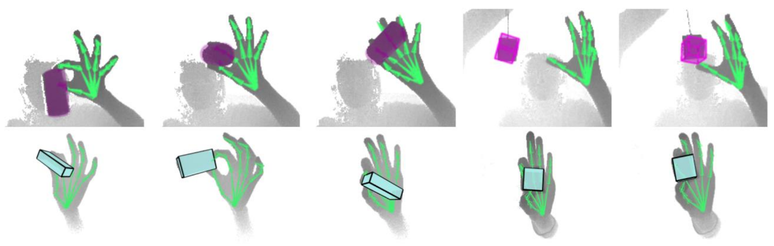
Results

RESULTS ON DATA FROM *TZIONAS ET AL.* (IJCV 2016)



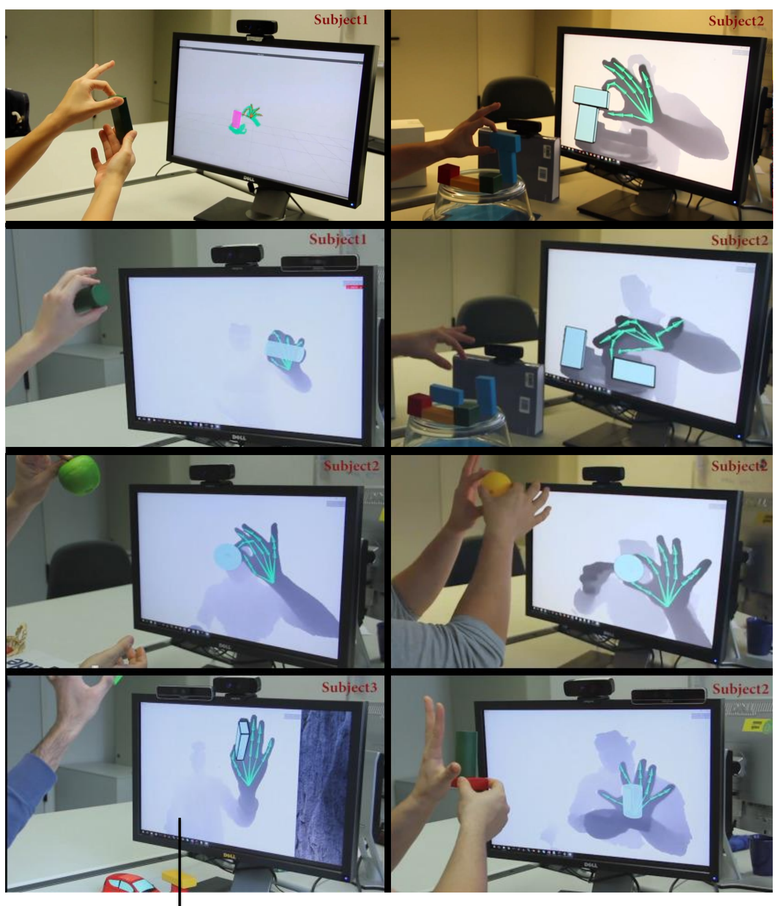
RESULTS ON DATA FROM *TZIONAS AND GALL* (ICCV 2015)

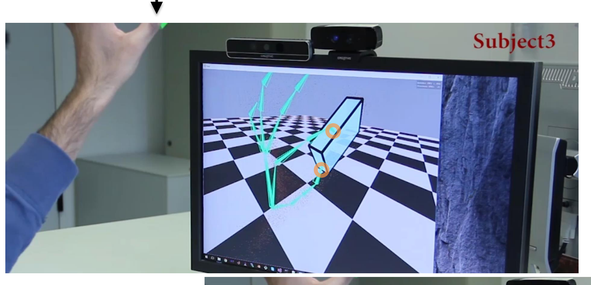
REAL-TIME TRACKING



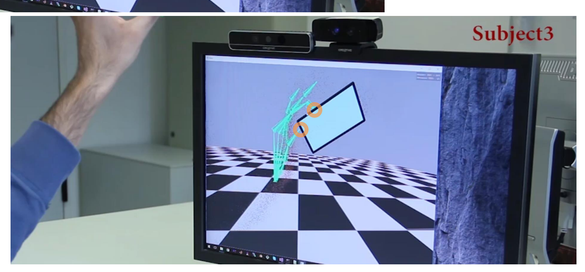
Tablet with video goes here

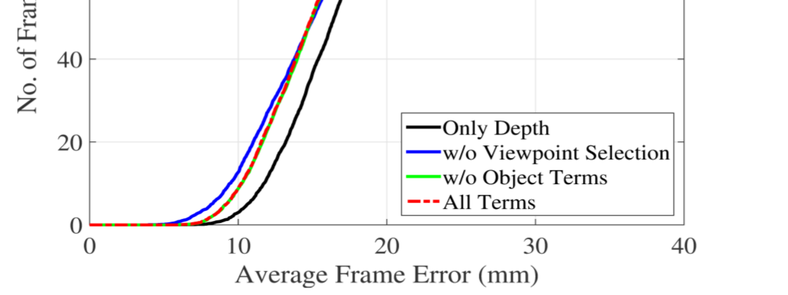
LIVE CAPTURE SETUP





HAND-OBJECT CONTACT CONSTRAINTS





Tracking consistency for the best, worst and average case in Dexter+Object. We are consistently below 30 mm.

Ablative analysis with different terms disabled shows robustness of our method.

- Runs at **30 FPS**
- Objects of **various shapes and sizes**
- Works with **different users**
- Average joint position error of **15.7 mm** on Dexter+Object
- 3D Gaussian mixture alignment improves hand-only tracking results on Dexter by **2 mm**

Dexter+Object dataset available!

handtracker.mpi-inf.mpg.de

Graphics Vision & Video

For access to a wide range of human shape and performance capture datasets, please visit:

gvvperfcapeva.mpi-inf.mpg.de

Supported by the ERC Starting Grant projects CapReal (335545) and COMPUTED (637991), and the Academy of Finland.

We would like to thank Christian Richardt.