3D Object Reconstruction from Hand-Object Interactions

Dimitrios Tzionas[1,2] and Juergen Gall[1]

[1]Computer Vision Group, University of Bonn

[2]Perceiving Systems Department, Max Planck Institute for Intelligent Systems

1 In-Hand Scanning

- Commercial RGB-D cameras enable 3d reconstruction applications
  - Kinect-Fusion-like approaches reconstruct spaces with a moving camera
  - Reconstruction of smaller objects possible with a static camera and
    - a turntable or
    - in-hand scanning

[1]

[2,3]

- Existing in-hand scanning approaches [1,2,3]
  - capitalize on high temporal continuity
  - need **stable** & **distinctive** geometry/texture features
  - **fail** in the **absence** of such **features**
  - **reject** information from the ***hands***

2 Problem Definition

- Highly **symmetric**, **textureless** objects are challenging
- Reconstruction possible currently only with intrusive ways
  - an additional trackable shape carving tool [4]
  - a robotic manipulator [5]
  - additional objects/features [6] or markers [7]
- **We** aim towards a **non-intrusive** reconstruction approach of such objects

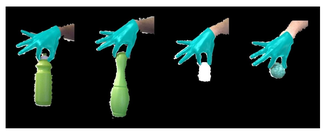
3 Proposed Idea

Σὺν Ἀθηνᾷ καὶ χεῖρα κίνει

- **Combine** in a feature-based approach
  - **visual features** (geometry/texture)
  - **contact points** based on hand-MoCap [8]

  during hand-object interaction
- **Enhance** in-hand scanning systems by re-using the rejected hand information

4 Experimental Setup



- Rotate 4 symmetric, textureless objects
- Capture RGB-D images D
- Skin-color segmentation of image D in
  - D_o for the object
  - D_h for the hand
- Hand MoCap similar to [8]:

$$E(\theta, D) = E_{model \to data}(\theta, D_h) + E_{data \to model}(\theta, D_h) + \gamma_c E_{collision}(\theta)$$

(hand pose parameters)

5 Reconstruction

Seek the rigid transformation $T=(\mathbf{R},\mathbf{t})$ that aligns the current (**source**) frame to the previous (**target**) frame

$\mathbf{R} \in SO(3)$, $\mathbf{t} \in \mathbb{R}^3$

minimize E_{feat}

Correspondences $(X, X') \in \mathcal{C}_{feat}(...)$

Corresponding **features @** source / target

Correspondence **Set**

$$E_{feat}(..., \mathbf{R}, \mathbf{t}) = \sum_{\mathcal{C}_{feat}} \|X' - (\mathbf{R}X + \mathbf{t})\|^2$$

Features:

- SIFT $\longrightarrow (X_{2d}, X'_{2d}) \in \mathcal{C}_{feat2d}(D_o) \longrightarrow E_{feat2d}(D_o, \mathbf{R}, \mathbf{t})$
- ISS3D/CSHOT $\longrightarrow (X_{3d}, X'_{3d}) \in \mathcal{C}_{feat3d}(D_o) \longrightarrow E_{feat3d}(D_o, \mathbf{R}, \mathbf{t})$

$$E_{visual}(D_o, \mathbf{R}, \mathbf{t}) = E_{feat2d}(D_o, \mathbf{R}, \mathbf{t}) + E_{feat3d}(D_o, \mathbf{R}, \mathbf{t})$$

Energy E_{visual} from traditional **visual features**

Find end-effectors @ mesh in **contact** with the object (proximity to D_o point cloud)

$(X_{hand}, X'_{hand}) \in \mathcal{C}_{hand}(\theta, D_h)$

contact correspondences

$$E(\theta, D_h, D_o, \mathbf{R}, \mathbf{t}) = E_{visual}(D_o, \mathbf{R}, \mathbf{t}) + \gamma_t E_{contact}(\theta, D_h, \mathbf{R}, \mathbf{t})$$

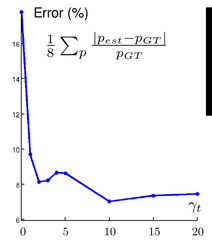
$\mathcal{C}_{feat2d}$ $\mathcal{C}_{feat3d}$ $\mathcal{C}_{hand}$ might be *noisy* → ICP refinement (Iterative Closest Point) → Align frame @ *partial* model →

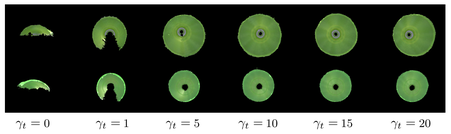
$$(X_{icp}, X'_{icp}) \in \mathcal{C}_{icp}(D_o)$$
$$E_{icp}(D_o, \mathbf{R}, \mathbf{t})$$

6 Contact Points Weight γ_t

$E = E_{visual} + \gamma_t E_{contact}$

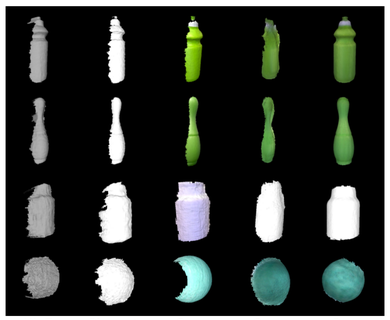
Steers influence of contact corr. $\mathcal{C}_{hand}$





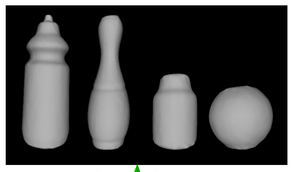
- $\gamma_t = 0$ degenerate reconstruction & high error
- $\gamma_t > 0$ better reconstruction & abrupt error drop
- $\gamma_t = 15$ our choice

7 Comparisons / Results



Reconstruction **without hands** similar across different systems (**degenerate reconstruction**)

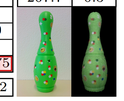
Hand MoCap incorporation in reconstruction plays a **vital role**



Final Results

8 Quantitative Results

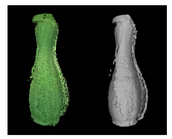
Dimensions Comparison	G.Truth	Ours $\gamma_t = 15$		KinFu		Skanect		Detect.Baseline		Enriched Texture	
		Capture	Diff.	Capture	Diff.	Capture	Diff.	Capture	Diff.	Capture	Diff.
Water-bottle diameter	73	82.3	9.3	66.2	6.8	64.3	8.7	86.6	13.6		
Water-bottle height	218	225.4	7.4	195.7	22.3	222.1	4.1	237.4	19.4		
Bowling-pin head diameter	50	50.8	0.8	54.1	4.1	39.0	11.0	48.7	1.3	49.8	0.2
Bowling-pin body diameter	82	90.0	8.0	70.9	11.1	63.8	18.2	93.2	11.2	89.4	7.4
Bowling-pin height	268	275.2	7.2	239.3	28.7	270.9	2.9	272.4	4.4	267.7	0.3
Small-bottle diameter	52	57.7	5.7	45.6	6.4	39.5	12.5	61.6	9.6		
Small-bottle height	80	89.5	9.5	78.1	1.9	84.9	4.9	95.0	15.0		
Sphere diameter	70	71.4	1.4	46.9	23.1	43.8	26.2	72.2	2.2		
Average			6.1625		13.05		11.0625		9.5875		
Sphere volume	179503	190490	10987	53988	125515	43974	135529	196965	17462		



Average error approximately $6mm$

9 Is ICP needed?

- **Omit ICP** stage E_{icp}
  Results in registration **artifacts**
- **ICP** enforces **consistency** with the **partial model** during reconstruction



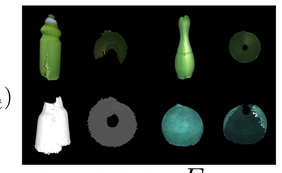
10 Is 3D Hand Pose the best way?

- Replace **contact points** with Hough-forest **contact detector** E_{detect} (top-down) VS $E_{contact}$ (bottom-up)
- Correspondences based on inner-bounding-box coordinates
- Annotate 2 points per end-effector for frame-pairs (X_{gt}, X'_{gt})
- Measure transformation error $\|X'_{gt} - (\mathbf{R}X_{gt} + \mathbf{t})\|$

Energy	mean	st.dev.
$E_{contact}+E_{visual}$	1.67	0.95
$E_{contact}$	1.64	0.88
$E_{detector}+E_{visual}$	1.73	1.08
$E_{detector}$	1.80	1.12

mm

Pose-based contact points **more accurate**



Contact detector E_{detect} leads to registration **artifacts**

11 References

Project Website **Dataset Code**

[1] S. Rusinkiewicz, O. Hall-Holt, and M. Levoy. *Real-time 3d model acquisition.* TOG 2002

[2] T. Weise, B. Leibe, and L. Van Gool. *Accurate and robust registration for in-hand modeling.* CVPR 2008

[3] T. Weise, T. Wismer, B. Leibe, and L. Van Gool. *Online loop closure for real-time interactive 3d scanning.* CVIU 2011

[4] D. Michel, X. Zabulis, and A. A. Argyros. *Shape from interaction.* MVA 2014

[5] M. Krainin, P. Henry, X. Ren, and D. Fox. *Manipulator and object tracking for in-hand 3d object modeling.* IJRR 2011

[6] R. A. Newcombe, S. Izadi, O. Hilliges, D. Molyneaux, D. Kim, A. J. Davison, P. Kohli, J. Shotton, S. Hodges, and A. Fitzgibbon. *KinectFusion: Real-Time Dense Surface Mapping and Tracking.* ISMAR 2011

[7] R.-G. Mihalyi, K. Pathak, N. Vaskevicius, and A. Birk. *Uncertainty estimation of ar-marker poses for graph-slam optimization in 3d object model generation with rgbd data.* IROS 2013

[8] D Tzionas, A Srikantha, P Aponte, and J Gall. *Capturing hand motion with an RGB-D sensor, fusing a generative model with salient points.* GCPR 2014

tzionas@informatik.uni-bonn.de

Project Page - http://files.is.tue.mpg.de/dtzionas/In-Hand-Scanning

gall@informatik.uni-bonn.de