

# Towards Markerless Grasp Capture

Samarth Brahmhatt<sup>1</sup>, Charles C. Kemp<sup>1</sup>, and James Hays<sup>1,2</sup>

<sup>1</sup>Institute for Robotics and Intelligent Machines, Georgia Tech <sup>2</sup>Argo AI

samarth. robo@gatech.edu, charlie.kemp@bme.gatech.edu, hays@gatech.edu

## Abstract

Humans excel at grasping objects and manipulating them. Capturing human grasps is important for understanding grasping behavior and reconstructing it realistically in Virtual Reality (VR). However, grasp capture – capturing the pose of a hand grasping an object, and orienting it w.r.t. the object – is difficult because of the complexity and diversity of the human hand, and occlusion. Reflective markers and magnetic trackers traditionally used to mitigate this difficulty introduce undesirable artifacts in images and can interfere with natural grasping behavior. We present preliminary work on a completely marker-less algorithm for grasp capture from a video depicting a grasp. We show how recent advances in 2D hand pose estimation can be used with well-established optimization techniques. Uniquely, our algorithm can also capture hand-object contact in detail and integrate it in the grasp capture process.

## 1. Introduction

Humans, with their complex hands made of soft tissue enveloping a rigid skeletal structure, excel at grasping and manipulating objects. Capturing human grasps of household objects can enable a better understanding of grasping behavior, which can improve a wide variety of VR and human-computer interaction applications. While hand- and object-pose capture and estimation have been studied extensively in isolation, there is a lack of large-scale *grasp capture* datasets and algorithms. We define grasp capture as capture of both the hand and object pose in a scene depicting grasping. Partial occlusion of the object by the hand and vice versa make grasp capture and prediction difficult. As mentioned in Section 2, the only large scale dataset employs wired magnetic trackers taped to the hand and object [5]. However, this method has the drawback of introducing unwanted artifacts in the RGB images and potentially interfering with natural grasping behavior.

In this paper we focus on capturing the hand pose through the 6-DOF palm pose and 20 joint angles, and the 6-DOF object pose. Since a single image is often not enough



Figure 1: Grasp capture for a scene depicting a cellphone being grasped. The multi-colored hand model shows the estimated hand pose and the dark blue model shows the estimated object pose.

to estimate both the hand and object pose, we record a video of a human participant grasping a household object. The participant rotates and translates their hand in 3D space to present the grasp to an RGB-D camera from various perspectives (see Figure 1 for an example frame).

Hand-object contact is either ignored, or enforced without any ground-truth contact observation in traditional grasp capture pipelines. Observing ground truth contact has so far been very difficult, but recently Brahmhatt et al [1] created a large scale dataset of detailed hand-object contactmaps through thermal imaging. Different from other grasp capture pipelines, ours can also capture such contactmaps and utilize them improve capture accuracy.

## 2. Related Work

Hand pose estimation is a highly researched topic, and many datasets are available publicly to train models for hand pose estimation. Hand pose is captured through data gloves [9, 7], manually annotated joint locations [15], magnetic trackers [20, 22], or fitting a hand model to depth images after manual initialization [17]. These methods capture only free hands rather than hands grasping objects.

However, as mentioned in Section 1, *grasp capture* also

involves capturing the object pose. Relatively few works have addressed this problem. The First Person Hand Action Benchmark [5] is the only large scale real-world dataset capturing both hand and object pose. 3D joint locations and object pose are captured through taped magnetic trackers. In addition to limited working volume (Hasson et al [6] mention in Section 5.2 that the object poses are imprecise and result in penetration of the hand inside the object by 1.1 cm on average), taping these long wired sensors to the hand introduces artifacts in the RGB images and can potentially interfere with natural grasping behavior.

### 2.1. Learning to Predict Aspects of Hand-Object Interaction

A large number of works estimate the pose of non-grasping hands in a model-based [18, 8] or model-free [21, 14, 11] manner. Garcia-Hernando et al [5] note that hand pose estimation in images depicting grasping benefits from including such grasping images in the training dataset. Tekin et al [16] predict both the hand and object pose by predicting 3D hand joint and object bounding box locations. Hasson et al [6] predict hand parameters and approximate the object with a predicted genus-0 geometry. Relying on predicted geometry reduces applicability to grasp capture for creating datasets, where object geometry is known in detail. In addition, it is not clear how accurate such algorithms will be on images from a data collection location, which can differ significantly from their training datasets.

## 3. Grasp Capture

### 3.1. Data Collection Protocol

As mentioned in Section 1, our aim is to capture both the hand pose (6-DOF palm pose and joint angles) and 6-DOF object pose from a video of a human participant grasping a household object. The objects in our experiments are 3D printed at real-life scale using detailed mesh models downloaded from online repositories. Our data collection protocol builds on the protocol from ContactDB [1], in which participants hold the object for 5 s and then place it on a turntable, where it is scanned with a calibrated RGB-D-Thermal camera rig. We propose to utilize the object holding time for grasp capture.

**Stage 1:** The object is first placed on the turntable, where its 6-DOF pose  ${}^wT_o$  is estimated using the depth camera point-cloud and the known object 3D model.

**Stage 2:** The grasp video recording starts when the participant reaches for the object. The participants are instructed to hold their joints steady after a transient phase (termed *grasp adjustment*) in which they pick the object up and settle into a comfortable grasp. Frames of the video after this instant are used to detect 2D hand joints using the OpenPose library [14] (Figure 2). These 2D detections  $\mathbf{x}^{(i)}$

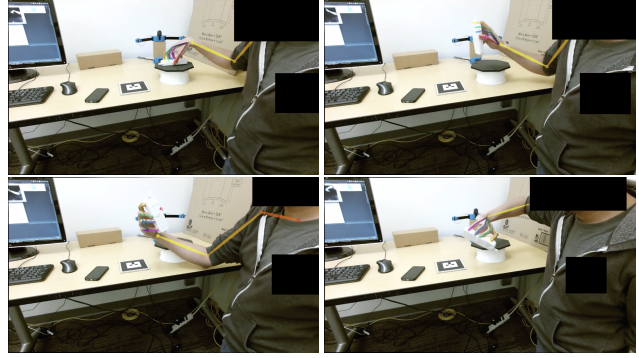


Figure 2: OpenPose [14] is used to detect 2D joint locations in the grasp video frames.

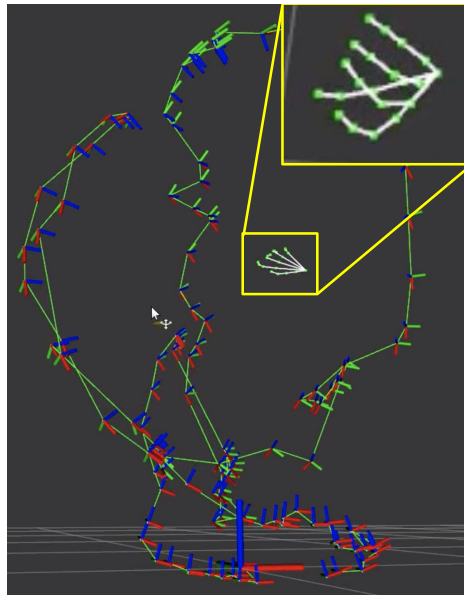


Figure 3: Structure from Motion (SfM) is used to recover 3D joint locations (green dots connected by white lines), and virtual camera poses (3D axes with blue pointing along the camera axis). Green lines connect consecutive camera poses.

are treated as observations from a mobile virtual camera that is observing a stationary hand (the problem is inverted; in reality we have a stationary camera and a moving hand). A Structure from Motion (SfM) problem is setup using these 2D detections, and optimized using the GTSAM library [3] to recover the 3D joint locations  $\mathbf{X}$  as well as virtual camera poses  $\{wT_c^{(i)}\}_{i=1}^N$ , with  $wT_c^{(1)}$  anchored to the origin (Figure 3). SfM minimizes the following re-projection error:

$$L\left(\mathbf{X}, \left\{wT_c^{(i)}\right\}_{i=1}^N\right) = \sum_{i=1}^N \|\mathbf{x}^{(i)} - \pi\left(\mathbf{X}; wT_c^{(i)}, K\right)\|_2^2 \quad (1)$$

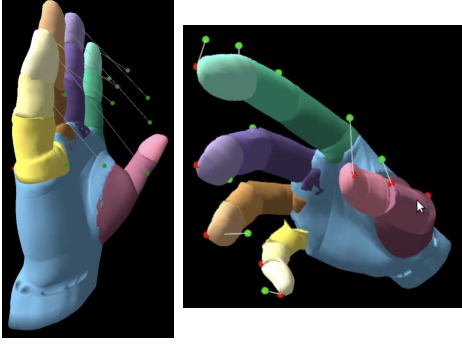


Figure 4: Fitting a hand model to the 3D joint locations. Left: Hand transformed by the palm pose. Right: Hand after inverse kinematics fitting. Green dots: target 3D joint locations (from SfM), Red dots: corresponding locations on the hand model.

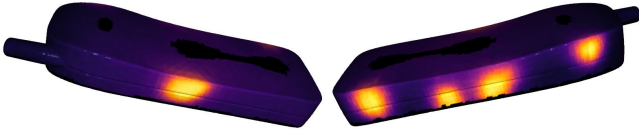


Figure 5: The contactmap for the grasp depicted in Figure 2, captured using ContactDB [1].

where  $\pi(\cdot)$  is the camera projection function.

**Stage 3:** The hand pose is estimated by fitting a hand model (we use HumanHand20DOF from GraspIt! [10]) to 3D joint locations  $\mathbf{X}$ , in two stages (Figure 4): 1) palm pose  ${}^wT_p$  is recovered from the locations of the 6 rigid hand points (wrist base + base of 5 fingers) through the Umeyama transform [19], which estimates a 3D similarity matrix, and 2) joint angles are recovered through inverse kinematics after the hand is transformed by  ${}^wT_p$ .

**Stage 4:** The participant places the object back on the turntable, which starts rotating and the object is scanned by the RGB-D-Thermal camera to construct the contact map according to the ContactDB [1] protocol (Figure 5).

**Stage 5:** The contact map from Stage 4 can be used to further refine the grasp capture by enforcing the observed contact relation i.e. attracting the closest hand segment to contacted points, repelling it away from non-contacted points, and penalizing intersection of the hand and object (Figure 1). We follow the grasp optimization stage of the ContactGrasp algorithm [2] to perform this refinement.

The virtual camera poses estimated in Stage 2 can be used to propagate the object and palm pose to all frames of the grasp video:  ${}^cT_o^{(i)} = [{}^wT_c^{(i)}]^{-1} T_{adj} {}^wT_o$ ,  ${}^cT_p^{(i)} = [{}^wT_c^{(i)}]^{-1} {}^wT_p$ . Here,  $T_{adj}$  is the change in the object pose during grasp adjustment mentioned in Stage 2. To summarize, the proposed algorithm captures the hand- and object-

pose for all frames in a video depicting a grasp from various perspectives, without requiring gloves, reflective markers or magnetic trackers.

### 3.2. Caveats

Grasp adjustment (Stage 2), which involves in-hand manipulation [4], introduces an unknown change  $T_{adj}$  in the object pose. We plan to estimate  $T_{adj}$  through ICP initialized at  ${}^wT_o$ . Another caveat is that OpenPose requires a visible head and shoulder to initialize the hand detector. Since it is not desirable to record the participants’ body and face for privacy reasons, we plan to develop a hand detector by skin color segmentation.

### 4. Future Work

We plan to improve the grasp capture algorithm described in Section 3 in two aspects:

- Utilizing a more expressive hand model (e.g. MANO [13]) will allow a better fit to individual hand characteristics. Currently, the only identity-dependent parameter in our algorithm the scale estimated during palm fitting (Stage 3). Research in hand modeling has shown [13] that many more parameters are needed to capture the diversity of human hands.
- Integrating hand-fitting into the SfM problem (Eq. 1) will reduce the number of stages in the pipeline and make it less brittle. Denoting hand parameters by  $\Phi$ , we plan to recover hand pose and virtual camera poses jointly by minimizing the following cost function:

$$L\left(\Phi, \left\{{}^wT_c^{(i)}\right\}_{i=1}^N\right) = \sum_{i=1}^N \left\| \mathbf{x}^{(i)} - \pi\left(J(\Phi); {}^wT_c^{(i)}, K\right) \right\|_2^2 \quad (2)$$

where  $J(\Phi)$  gives the 3D joint locations from  $\Phi$ .

### 5. Conclusion

In summary, this paper presents preliminary work on a completely markerless grasp capture algorithm that utilizes well-established geometric optimization techniques and recent advances in 2D hand keypoint detection. In addition to the hand- and object-pose, our algorithm also captures detailed hand-object contact, which is an important component of grasping. We also discuss ways to improve the proposed algorithm. Markerless grasp capture can enable a better understanding of human grasping behavior, and can generate datasets for training models to predict various aspects of grasping like physically plausible hand pose, occurrence of contact at object and hand locations, and potentially even locations and directions of forces being applied to the object [12]. These models have applications in VR and human-computer interaction.

## References

- [1] Samarth Brahmabhatt, Cusuh Ham, Charles C. Kemp, and James Hays. ContactDB: Analyzing and predicting grasp contact via thermal imaging. *2019 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun 2019. 1, 2, 3
- [2] Samarth Brahmabhatt, Ankur Handa, James Hays, and Dieter Fox. ContactGrasp: Functional multi-finger grasp synthesis from contact. *arXiv preprint arXiv:1904.03754*, 2019. 3
- [3] Frank Dellaert. Factor graphs and gtsam: A hands-on introduction. Technical report, Georgia Institute of Technology, 2012. 2
- [4] Charlotte E Exner. In-hand manipulation skills. *Development of hand skills in the child*, pages 35–45, 1992. 3
- [5] Guillermo Garcia-Hernando, Shanxin Yuan, Seungryul Baek, and Tae-Kyun Kim. First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 409–419, 2018. 1, 2
- [6] Yana Hasson, Gül Varol, Dimitrios Tzionas, Igor Kalevatykh, Michael J. Black, Ivan Laptev, and Cordelia Schmid. Learning joint reconstruction of hands and manipulated objects. In *CVPR*, 2019. 2
- [7] Guido Heumer, Heni Ben Amor, Matthias Weber, and Bernhard Jung. Grasp recognition with uncalibrated data gloves—a comparison of classification methods. In *2007 IEEE Virtual Reality Conference*, pages 19–26. IEEE, 2007. 1
- [8] Nikolaos Kyriazis and Antonis Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9–16, 2013. 2
- [9] Yun Lin and Yu Sun. Grasp planning based on strategy extracted from demonstration. In *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 4458–4463. IEEE, 2014. 1
- [10] Andrew T Miller and Peter K Allen. Grasplit! a versatile simulator for robotic grasping. *IEEE Robotics & Automation Magazine*, 11(4):110–122, 2004. 3
- [11] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2018. 2
- [12] Grégory Rogez, James S Supancic, and Deva Ramanan. Understanding everyday hands in action from rgb-d images. In *Proceedings of the IEEE international conference on computer vision*, pages 3889–3897, 2015. 3
- [13] Javier Romero, Dimitrios Tzionas, and Michael J. Black. Embodied hands: Modeling and capturing hands and bodies together. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)*, 36(6), Nov. 2017. 3
- [14] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. Hand keypoint detection in single images using multiview bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1145–1153, 2017. 2
- [15] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013. 1
- [16] Bugra Tekin, Federica Bogo, and Marc Pollefeys. H+ o: Unified egocentric recognition of 3d hand-object poses and interactions. *arXiv preprint arXiv:1904.05349*, 2019. 2
- [17] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, 33, August 2014. 1
- [18] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016. 2
- [19] Shinji Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (4):376–380, 1991. 3
- [20] Aaron Wetzler, Ron Slossberg, and Ron Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. *arXiv preprint arXiv:1507.05726*, 2015. 1
- [21] Qi Ye, Shanxin Yuan, and Tae-Kyun Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *European conference on computer vision*, pages 346–361. Springer, 2016. 2
- [22] Shanxin Yuan, Qi Ye, Bjorn Stenger, Siddhant Jain, and Tae-Kyun Kim. Bighand2.2m benchmark: Hand pose dataset and state of the art analysis. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017. 1