GRAB: A Dataset of Whole-Body Human Grasping of Objects

Omid Taheri, Nima Ghorbani, Michael J. Black, and Dimitrios Tzionas

Max Planck Institute for Intelligent Systems, Tübingen, Germany
{otaheri,nghorbani,black,dtzionas}@tuebingen.mpg.de

Abstract. Training computers to understand, model, and synthesize human grasping requires a rich dataset containing complex 3D object shapes, detailed contact information, hand pose and shape, and the 3D body motion over time. While "grasping" is commonly thought of as a single hand stably lifting an object, we capture the motion of the entire body and adopt the generalized notion of "whole-body grasps". Thus, we collect a new dataset, called *GRAB* (GRasping Actions with Bodies), of whole-body grasps, containing full 3D shape and pose sequences of 10 subjects interacting with 51 everyday objects of varying shape and size. Given MoCap markers, we fit the full 3D body shape and pose, including the articulated face and hands, as well as the 3D object pose. This gives detailed 3D meshes over time, from which we compute contact between the body and object. This is a unique dataset, that goes well beyond existing ones for modeling and understanding how humans grasp and manipulate objects, how their full body is involved, and how interaction varies with the task. We illustrate the practical value of GRAB with an example application; we train GrabNet, a conditional generative network, to predict 3D hand grasps for unseen 3D object shapes. The dataset and code are available for research purposes at https://grab.is.tue.mpg.de.

1 Introduction

A key goal of computer vision is to estimate human-object interactions from video to help understand human behavior. Doing so requires a strong model of such interactions and learning this model requires data. However, capturing such data is not simple. Grasping involves both gross and subtle motions, as humans involve their whole body and dexterous finger motion to manipulate objects. Therefore, objects contact multiple body parts and not just the hands. This is difficult to capture with images because the regions of contact are occluded. Pressure sensors or other physical instrumentation, however, are also not a full solution as they can impair natural human-object interaction and do not capture full-body motion. Consequently, there are no existing datasets of complex human-object interaction that contain full-body motion, 3D body shape, and detailed body-object contact. To fill this gap, we capture a novel dataset of full-body 3D humans dynamically interacting with 3D objects as illustrated in Fig. 1. By accurately tracking 3D body and object shape, we reason about

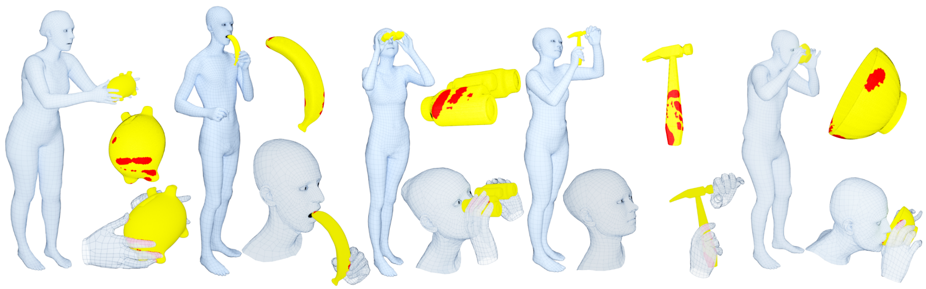

Fig. 1: Example "whole-body grasps" from the GRAB dataset. A "grasp" is usually thought of as a single hand interacting with an object. Using objects, however may involve more than just a single hand. From left to right: (i) passing a piggy bank, (ii) eating a banana, (iii) looking through binoculars, (iv) using a hammer, (v) drinking from a bowl. Contact between the object and the body is shown in red on the object; here contact areas are spatially extended to aid visualization. See the video on our website for a wide range of sequences with various objects and intents.

contact resulting in a dataset with detail and richness beyond existing grasping datasets.

Most previous work focuses on prehensile "grasps" [43]; i.e. a single human hand stably lifting or using an object. The hands, however, are only part of the story. For example, as infants, our earliest grasps involve bringing objects to the mouth [58]. Consider the example of drinking from a bowl in Fig. 1 (right). To do so, we must pose our body so that we can reach it, we orient our head to see it, we move our arm and hand to stably lift it, and then we bring it to our mouth, making contact with the lips, and finally we tilt the head to drink. As this and other examples in the figure illustrate, human grasping and using of everyday objects involves the *whole body*. Such interactions are fundamentally *three-dimensional*, and contact occurs between objects and multiple body parts.

Dataset. Such whole-body grasping [25] has received much less attention [4, 37] than single hand-object grasping [3, 10, 14, 27, 43]. To model such grasping we need a dataset of humans interacting with varied objects, capturing the full 3D surface of both the body and objects. To solve this problem we adapt recent motion capture techniques, to construct a new rich dataset called **GRAB** for "*GRasping Actions with Bodies.*" Specifically, we adapt MoSh++ [38] in two ways. First, MoSh++ estimates the 3D shape and motion of the body and hands from MoCap markers; here we extend this to include facial motion. For increased accuracy we first capture a 3D scan of each subject and fit the SMPL-X body model [46] to it. Then MoSh++ is used to recover the pose of the body, hands and face. Note that the face is important because it is involved in many interactions; see in Fig. 1 (second from left) how the mouth opens to eat a banana. Second, we also accurately capture the motion of 3D objects as the are manipulated by the subjects. To this end, we use small hemispherical markers on the objects and show that these do not impact grasping behavior. As a result, we obtain detailed 3D meshes for both the object and the human (with a full body, articulated fingers and face) moving over time while in interaction, as shown in Fig. 1. Using these meshes we then infer the body-object contact (red regions in Fig. 1). Unlike [5] this gives both the contact and the full body/hand pose over time. Interaction is dynamic, including in-hand manipulation and re-grasping. GRAB captures 10 different people (5 male and 5 female) interacting with 51 objects from [5]. Interaction takes place in 4 different contexts: lifting, handing over, passing from one hand to the other, and using, depending on the affordances and functionality of the object.

Applications. GRAB supports multiple uses of interest to the community. First, we show how GRAB can be used to gain insights into hand-object contact in everyday scenarios. Second, there is significant interest in training models to grasp 3D objects [59]. Thus, we use GRAB to train a conditional variational autoencoder (cVAE) to generate plausible grasps for unseen 3D objects. Given a randomly posed 3D object, we predict plausible hand parameters (wrist pose and finger articulation) appropriate for grasping the object. To encode arbitrary 3D object shapes, we employ the recent basis point set (BPS) representation [52], whose fixed size is appropriate for neural networks. Then, by conditioning on a new 3D object shape, we sample from the learned latent space, and generate hand grasps for this object. We evaluate both quantitatively and qualitatively the resulting grasps and show that they look natural.

In summary, this work makes the following contributions: (1) we introduce a unique dataset capturing real "whole-body grasps" of 3D objects, including full-body human motion, object motion, in-hand manipulation and re-grasps; (2) to capture this, we adapt MoSh++ to solve for the body, face and hands of SMPL-X to obtain detailed moving 3D meshes; (3) using these meshes and tracked 3D objects we compute plausible contact on the object and the human and provide an analysis of observed patterns; (4) we show the value of our dataset for machine learning, by training a novel conditional neural network to generate 3D hand grasps for unseen 3D objects. The dataset, models, and code are available for research purposes at `https://grab.is.tue.mpg.de`.

2 Related Work

Hand Grasps: Hands are crucial for grasping and manipulating objects. For this reason, many studies focus on understanding grasps and defining taxonomies [3, 10, 14, 27, 43, 51]. This work has explored the object shape and purpose of grasps [10], contact areas on the hand captured by sinking objects in ink [27], pose and contact areas [3] captured with an integrated data-glove [11] and tactile-glove [51], or number of fingers in contact with the object and thumb position [14]. A key element for these studies is capturing accurate hand poses, relative hand-object configurations and contact areas.

Whole-Body Grasps: Often people use more than a single hand to interact with objects. However, there are not many works in the literature on this topic [4, 25]. Borras et al. [4] use MoCap data [39] of people interacting with a scene with multi-contact, and present a body pose taxonomy for such whole-body grasps. Hsiao et al. [25] focus on imitation learning with a database of whole-body grasp demonstrations with a human teleoperating a simulated robot. Although these works go into the right direction, they use unrealistic humanoid models and simple objects [4, 25] or synthetic ones [25]. Instead, we use the SMPL-X model [46] to capture "whole-body", face and dexterous in-hand interactions.

Capturing Interactions with MoCap: MoCap is often used to capture, synthesize or evaluate humans interacting with scenes. Lee et al. [36] capture a 3D body skeleton interacting with a 3D scene and show how to synthesize new motions in new scenes. Wang et al. [73] capture a 3D body skeleton interacting with a large geometric objects. Han et al. [20] present a method for automatic labeling of hand markers, to speed up hand tracking for VR. Le et al. [35] capture a hand interacting with a phone to study the "comfortable areas", while Feit et al. [13] capture two hands interacting with a keyboard to study typing patterns. Other works [34, 49] focus on graphics applications. Kry et al. [34] capture a hand interacting with a 3D shape primitive, instrumented with a force sensor. Pollard et al. [49] capture the motion of a hand to learn a controller for physically based grasping. Mandery et al. [39] sit between the above works, capturing humans interacting with both big and handheld objects, but without articulated faces and fingers. None of the previous work captures full 3D bodies, hands and faces together with 3D object manipulation and contact.

Capturing Contact: Capturing human-object contact is hard, because the human and object heavily occlude each other. One approach is instrumentation with touch and pressure sensors, but this might bias natural grasps. Pham et al. [47] predefine contact points on objects to place force transducers. More recent advances in tactile sensors allow accurate recognition of tactile patterns and handheld objects [65]. Some approaches [3] use a data glove [11] with an embedded tactile glove [51, 67] but this combination is complicated and the two modalities can be hard to synchronize. A microscopic-domain tactile sensor [16] is introduced in [26], but is not easy to use on human hands. Mascaro et al. [40] attach a minimally invasive camera to detect changes in the coloration of fingernails. Brahmbhatt et al. [5] use a thermal camera to directly observe the "thermal print" of a hand on the grasped object. However, for this they only capture static grasps that last long enough for heat transfer. Consequently, even recent datasets that capture realistic hand-object [15, 19] or body-scene [23, 61] interaction avoid directly measuring contact.

3D Interaction Models: Learning a model of human-object interactions is useful for graphics and robotics to help avatars [12, 63] or robots [21] interact with their surroundings, and for vision [18, 42] to help reconstruct interactions from ambiguous data. However, there is a chicken-and-egg problem; to capture or synthesize data to learn a model, one needs such a model in the first place. For this reason, the community has long used hand-crafted approaches that exploit contact and physics, for body-scene [6, 22, 23, 57, 76], body-object [30, 32, 37], or hand-object [24, 45, 54, 55, 62, 68, 69, 72, 77, 78] scenarios. These approaches compute contact approximately; this may be rough for humans modeled as 3D skeletons [30, 37] or shape primitives [6, 32, 62], or relatively accurate when using 3D meshes, whether generic [54, 68], personalized [22, 57, 69, 72], or based on 3D statistical models [23, 24].

To collect training data, several works [54, 55] use synthetic Poser [50] hand models, manually articulated to grasp 3D shape primitives. Contact points and forces are also annotated [54] through proximity and inter-penetration of 3D meshes. In contrast, Hasson et al. [24] use the robotics method GraspIt [41] to automatically generate 3D MANO [56] grasps for ShapeNet [7] objects and render synthetic images of the hand-object interaction. However, GraspIt optimizes for hand-crafted grasp metrics that do not necessarily reflect the distribution of human grasps (see Sup. Mat. Sec. C.2 of [24], and [17]). Alternatively, Garcia-Hernando et al. [15] use magnetic sensors to reconstruct a 3D hand skeleton and rigid object poses; they capture 6 subjects interacting with 4 objects. This dataset is used by [33, 66] to learn to estimate 3D hand and object poses, but suffers from noisy poses and significant inter-penetrations (see Sec. 5.2 of [24]).

For bodies, Kim et al. [30] use synthetic data to learn to detect contact points on a 3D object, and then fit an interacting 3D body skeleton to them. Savva et al. [61] use RGB-D to capture 3D body skeletons of 5 subjects interacting in 30 3D scenes, to learn to synthesize interactions [61], affordance detection [60], or to reconstruct interaction from videos [42]. Mandery et al. [39] use optical MoCap to capture 43 subjects interacting with 41 tracked objects, both large and small. This is similar to our effort but they do not capture fingers or 3D body shape, so cannot reason about contact. Corona et al. [9] use this dataset to learn context-aware body motion prediction. Starke et al. [63] use Xsens IMU sensors [75] to capture the main body of a subject interacting with large objects, and learn to synthesize avatar motion in virtual worlds. Hassan et al. [23] use RGB-D and 3D scene constraints to capture 20 humans as SMPL-X [46] meshes interacting with 12 static 3D scenes, but do not capture object manipulation. Zhang et al. [79] use this data to learn to generate 3D scene-aware humans.

We see that only parts of our problem have been studied. We draw inspiration from prior work, in particular [5, 23, 25, 39]. We go beyond these by introducing a new dataset of real "whole-body" grasps, as described in the next section.

3 Dataset

To manipulate an object, the human needs to approach its 3D surface, and bring their skin to come in *physical contact* to apply forces. Realistically capturing such human-object interactions, especially with "whole-body grasps", is a challenging problem. First, the object may occlude the body and vice-versa, resulting in *ambiguous* observations. Second, for physical interactions it is crucial to reconstruct an accurate and detailed 3D *surface* for both the human and the object. Additionally, the capture has to work across multiple scales (body, figners and

Fig. 2: MoCap markers used to capture humans and objects. **Left:** We attach 99 reflective markers per subject; 49 for the body, 14 for the face and 36 for the fingers. We use spherical 4.5 mm radius markers for the body and hemi-spherical 1.5 mm radius ones for the hands and face. **Right:** Example 3D printed objects from [5]. We glue 1.5 mm radius hemi-spherical markers (the gray dots) on the objects. These makers are small enough to be unobtrusive. The 6 objects on the right are mostly used by one or more hands, while the 6 on the left involve "whole-body grasps".

face) and for objects of varying complexity. We address these challenges with a unique combination of state-of-the-art solutions that we adapt to the problem.

There is a fundamental trade-off with current technology; one has to choose between (a) accurate motion with instrumentation and without natural RGB images, or (b) less accurate motion but with RGB images. Here we take the former approach; for an extensive discussion we refer the reader to Sup. Mat.

3.1 Motion Capture (MoCap)

We use a Vicon system with 54 infrared "Vantage 16" [71] cameras that capture 16 MP at 120 fps. The large number of cameras minimized occlusions and the high frame rate captures temporal details of contact. The high resolution allows small (1.5 mm radius) hemi-spherical markers. This minimizes their influence on finger and face motion and does not alter how people grasp objects. Details of the marker setup are shown in Fig. 2. Even with many cameras, motion capture of the body, face, and hands, together with objects, is uncommon because it is so challenging. MoCap, markers become occluded, labels are swapped, and ghost makers appear. MoCap cleaning was done by four trained technicians using Vicon's Shōgun-Post software.

Capturing Human MoCap: To capture human motion, we use the marker set of Fig. 2 (left). The body markers are attached on a tight body suit with a velcro-based at a distance of roughly $d_b = 9.5$ mm from the body surface. The hand and face markers are attached directly to the skin with special removable glue, therefore the distance to it is roughly $d_h = d_f \approx 0$ mm. Importantly, no

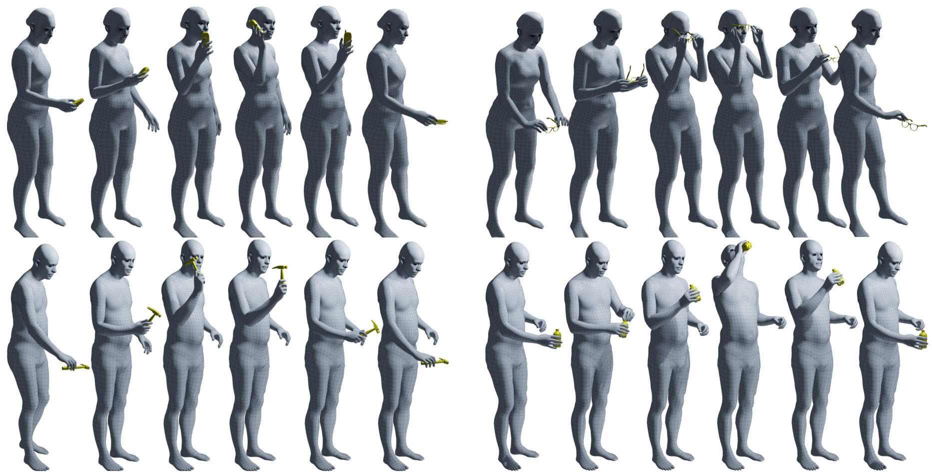

Fig. 3: We capture humans interacting with objects over time and reconstruct sequences of 3D meshes for both, as described in Sec. 3.1 and Sec. 3.2. Note the realistic and plausible placement of objects in the hands, and the "whole-body" involvement. The video on our website shows more examples.

hand glove is used and hand markers are placed only on the dorsal side, leaving the palmar side completely uninstrumented, for natural interactions.

Capturing Objects: To reconstruct interactions accurately, it is important to know the precise 3D object surface geometry. We therefore use the CAD object models of [5], and 3D print them with a Stratasys Fortus 360mc [64] printer; see Fig. 2 (right). Each object o is then represented by a known 3D mesh with vertices V_o. To capture object motion, we attach on the 1.5 mm hemi-spherical markers with strong glue directly to the object surface. We use at least 8 markers per object, empirically distributing them on the object so that at least 3 of them are always observed. The size and placement of the markers makes them unobtrusive. In Sup. Mat. we show empirical evidence that makers have minimal influence on grasping.

3.2 From MoCap Markers to 3D Surfaces

Human Model: We model the human with the SMPL-X [46] 3D body model. SMPL-X jointly models the body with an articulated face and fingers; this expressive body model is critical to capture physical interactions. More formally, SMPL-X is a differentiable function $M_b(\beta, \theta, \psi, \gamma)$ that is parameterized by body shape β, pose θ, facial expression ψ and translation γ. The output is a 3D mesh $M_b = (V_b, F_b)$ with $N_b = 10475$ vertices $V_b \in \mathbb{R}^{(N_b \times 3)}$ and triangles F_b. The shape parameters $\beta \in \mathbb{R}^{10}$ are coefficients in a learned low-dimensional linear shape space. This lets SMPL-X represent different subject identities with the same mesh topology. The 3D joints, $J(\beta)$, of a kinematic skeleton are regressed from the body shaped defined by β. The skeleton has 55 joints in total; 22 for the body, 15 joints per hand for finger articulation, and 3 for the neck and eyes. Corrective blend shapes are added to the body shape and then posed body is defined by linear blend skinning with this underlying skeleton. The overall pose parameters $\theta = (\theta_b, \theta_f, \theta_h)$ are comprised of $\theta_b \in \mathbb{R}^{66}$ and $\theta_f \in \mathbb{R}^{9}$ parameters in axis-angle representation for the main body and face joints correspondingly, with 3 degrees of freedom (DoF) per joint, and $\theta_h \in \mathbb{R}^{60}$ parameters in a lower-dimensional pose space for both hands, i.e. 30 DoF per hand following [24]. For more details, please see [46].

Model-Marker Correspondences: For the human body we define, a priori, the rough marker placement on the body as shown in Fig. 2 (left). Exact marker locations on individual subjects are then computed automatically using MoSh++ [38]. In contrast to the body, the objects have different shapes and mesh topologies. Markers are placed according to the object shape, affordances and expected occlusions during interaction; Fig. 2 (right), Therefore, we annotate object-specific vertex-marker correspondences, and do this once per object.

Human and Object Tracking: To ensure accurate human shape, we capture a 3D scan of each subject and fit SMPL-X to it following [56]. We fit these personalized SMPL-X models to our cleaned 3D marker observations using MoSh++ [38]. Specifically we optimize over pose, θ, expressions, ψ, and translation, γ, while keeping the known shape, β, fixed. The weights of MoSh++ for the finger and face data terms are tuned on a synthetic dataset, as described in Sup. Mat. An analysis of MoSh++ fitting accuracy is also provided in Sup. Mat.

Objects are simpler because they are rigid and we know their 3D shape. Given three or more detected markers, we solve for the rigid object pose $\theta_o \in \mathbb{R}^6$. IHere we track the human and object separately and on a per-frame basis. Figure 3 shows that our approach captures realistic interactions and reconstructs detailed 3D meshes for both the human and the object, over time. The video on our website shows a wide range of reconstructed sequences.

3.3 Contact Annotation

Since contact cannot be directly observed, we estimate it using 3D proximity between the 3D human and object meshes. In theory, they come in contact when the distance between them is zero. In practice, however, we relax this and define contact when the distance, $d \leq \epsilon_{contact}$, for a threshold $\epsilon_{contact}$. This helps address: (1) measurement and fitting errors, (2) limited mesh resolution, (3) the fact that human soft tissue deforms when grasping an object, while the SMPL-X model cannot model this.

Given these issues, accurately estimating contact is challenging. Consider the hand grasping a wine glass in Fig. 4 (right), where the color rings indicate intersections. Ideally, the glass should be in contact with the thumb, index and middle fingers. "Contact under-shooting" results in fingers hovering close to the object surface, but not on it, like the thumb. "Contact over-shooting", results in fingers penetrating the object surface around the contact area, like the index (purple intersections) and middle finger (red intersections). The latter case is especially problematic for thin objects where a penetrating finger can pass through

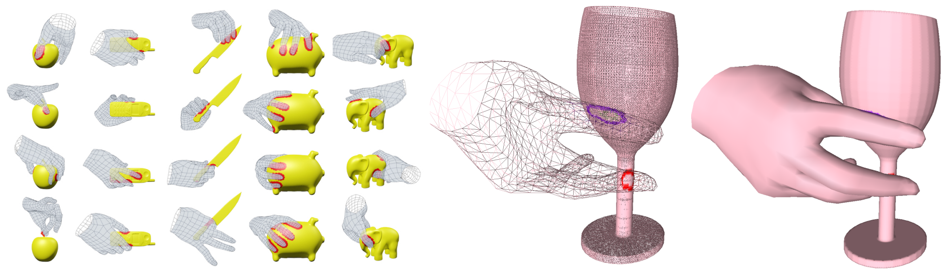

Fig. 4: Left: Accurate tracking lets us compute realistic contact areas (red) for each frame (Sec. 3.3). For illustration, we render only the hand of SMPL-X and spatially extend the red contact areas for visibility. **Right:** Detection of "intersection ring" triangles during contact annotation (Sec. 3.3).

the object, intersecting it on two sides. In this example, we want to annotate contact only with the outer surface of the object and not the inner one.

We account for "contact over-shooting" cases with an efficient heuristic. We use a fast method [29, 46] to detect intersections, cluster them in connected "intersection rings", $\mathcal{R}_b \subsetneq V_b$ and $\mathcal{R}_o \subsetneq V_o$, and label them with the intersecting body part, seen as purple and red rings in Fig. 4 (right). The "intersection ring", $\mathcal{R}_b$, segments the body mesh M_b to give the "penetrating sub-mesh" $\mathcal{M}_b \subsetneq M_b$. (1) When a body part gives only one intersection, we annotate the points $V_o^{\mathcal{C}} \subset V_o$ on the object all vertices enclosed by the ring $\mathcal{R}_o$ as being in contact. We then annotate as contact points, $V_b^{\mathcal{C}} \subset V_b$, on the body all vertices that lie close to $V_o^{\mathcal{C}}$ with a distance $d_{o \rightarrow b} \leq \epsilon_{contact}$. (2) In case of multiple intersections i we take into account only the ring $\mathcal{R}_b^i$ corresponding to the largest intersection subset, $\mathcal{M}_b^i$.

For body parts that are not found in contact above, there is the possibility of "contact under-shooting". To address this, we compute the distance from each object vertex V_o, to each non-intersecting body vertex V_b. We then annotate as contact vertices, $V_o^{\mathcal{C}}$ and $V_b^{\mathcal{C}}$, the ones with $d_{o \rightarrow b} \leq \epsilon_{contact}$. We empirically find that $\epsilon_{contact} = 4.5$ mm works well for our purposes.

3.4 Dataset Protocol

Human-object interaction depends on various factors including the human body shape, object shape and affordances, object functionality, or interaction intent, to name a few. We therefore capture 10 people (5 men and 5 women), of various sizes and nationalities, interacting with the objects of [5]; see example objects in Fig. 2 (right). All subjects gave informed written consent to share their data for research purposes.

For each object we capture interactions with 4 different intents, namely "use" and "pass" (to someone), borrowed from [5], as well as "lift" and "off-hand pass" (from one hand to the other). Figure 3 shows some example 3D capture sequences for the "use" intent. For each sequence we: (i) we randomize initial

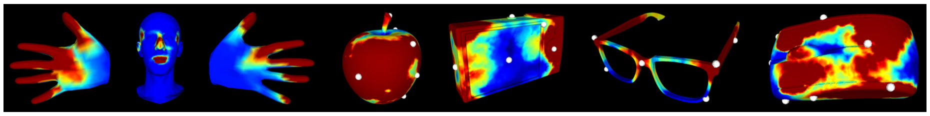

Fig. 5: Contact "heatmaps". **Left:** For the body we focus on "use" sequences to show "whole-body grasps". **Right:** For objects we include all intents. Object markers (light gray) are unobtrusive and can lie on "hot" (red) contact areas.

Table 1: Size of the GRAB dataset. GRAB is sufficiently large to enable training of data-driven models of grasping as shown in Sec. 4.

Intent	"Use"	"Pass"	"Lift"	"Off-hand"	Total
# Sequences	579	414	274	67	1334
# Frames	605.796	335.733	603.381	77.549	1.622.459

object placement to increase motion variance, (ii) we instruct the subject to follow an intent, (iii) the subject starts from a T-pose and approaches the object, (iv) they perform the instructed task, and (v) they leave the object and returns to a T-pose. The video on our website shows the richness of our protocol with a wide range of captured sequences.

3.5 Analysis

Dataset Analysis. The dataset contains 1334 sequences and over 1.6M frames of MoCap; Table 1 provides a detailed breakdown. Here we analyze those frames where we have detected contact between the human and the object. We assume that the object is static on a table and can move only due to grasping. Consequently, consider contact frames to be those in which the object's position deviates in the vertical direction by at least 5 mm from its initial position and in which at least 50 body vertices are in contact with the object. This results in $952,514$ contact frames that we analyze below. The exact thresholds of these contact heuristics have little influence on our analysis, see Sup. Mat.

By uniquely capturing the whole body, and not just the hand, interesting interaction patterns arise. highlightCR By focusing on "use" sequences that highlight the object functionality, we observe that 92% of contact frames involve the right hand, 39% the left hand, 31% both hands, and 8% involve the head. For the first category the per-finger contact likelihood, from thumb to pinky, is 100%, 96%, 92%, 79%, 39% and for the palm 24%. For more results see Sup. Mat.

To visualize the fine-grained contact information, we integrate over time the binary per-frame contact maps, and generate "heatmaps" encoding the contact likelihood of contact across the whole body surface. Figure 5 (left) shows such "heatmaps" for "use" sequences. "Hot" areas (red) denote high likelihood of contact, while "cold" areas (blue) denote low likelihood. We see that both the hands and face are important for using everyday objects, highlighting the importance

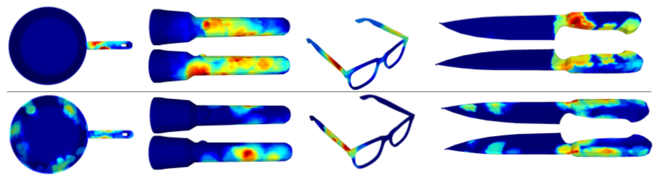

Fig. 6: Effect of interaction intent on contact during grasping. We show the "use" (top) and "pass" (bottom) intents for 4 different objects.

of capturing the whole interacting body. For the face, the "hot" areas are the lips, the nose, the temporal head area, and the ear. For hands, the fingers are more frequently in contact than the palm, with more contact on the right than left. The palm seems more important for right-hand grasps than for left-hand ones, possibly because all our subjects are right-handed. Contact patterns are also influenced by the size of the object and the size of the hand; see Sup. Mat. for a visualization.

Figure 6 shows the effect of the intent. Contact for "use" sequences complies with the functionality of the object; e.g. people do not touch the knife blade or the hot area of the pan, but they do contact the on/off button of the flashlight. For "pass" sequences subjects tend to contact one side of the object irrespective of affordances, leaving the other one free to be grasped by the receiving person.

For natural interactions it is important to have a minimally intrusive setup. While our MoCap markers are small and unobtrusive (Figure 2 (right)), we ask whether subjects may be biased in their grasps by these markers? Figure 5 (right) shows contact "heatmaps" for some objects across all intents. These clearly show that markers are often located in "hot" areas, suggesting that subjects do not avoid grasping these locations. Further analysis based on K-means clustering of grasps can be found in Sup. Mat.

4 GrabNet: Learning to Grab an object

We show the value of the GRAB dataset with a challenging example application; we use it to train a model that generates plausible 3D MANO [56] grasps for a previously unseen 3D object. Our model, called GrabNet, is comprised of two main modules. First, we employ a conditional variational autoencoder (cVAE) [31], called CoarseNet, that generates an initial grasp. For this it learns a grasping embedding space Z conditioned on the object shape, that is encoded using the Basis Point Set (BPS) [52] representation as a set of distances from the basis points to the nearest object points. CoarseNet's grasps are reasonable, but realism can improve by refining contacts based on the distances D between the hand and the object. We do this with a second network, called RefineNet. The architecture of GrabNet is shown in Fig. 7, for more details see Sup. Mat.

Pre-processing. For training, we gather all frames with right-hand grasps that involve some minimal contact, for details see Sup. Mat. We then center each training sample, i.e. hand-object grasp, at the centroid of the object and compute the $BPS_o \in R^{4096}$ representation for the object, used for conditioning.

CoarseNet. We pass the object shape BPS_o along with initial MANO wrist rotation θ_{wrist} and translation γ to the encoder $Q(Z|\theta_{wrist}, \gamma, BPS_o)$ that produces a latent grasp code $Z \in R^{16}$. The decoder $P(\bar{\theta}, \bar{\gamma}|Z, BPS_o)$ maps Z and BPS_o to MANO parameters with full finger articulation $\bar{\theta}$, to generate a 3D grasping hand. For the training loss, we use standard cVAE loss terms (KL divergence, weight regularizer), a data term on MANO mesh edges (L1), as well as a penetration and a contact loss. For the latter, we learn candidate contact point weights from GRAB, in contrast to handcrafted ones [23] or weights learned from artificial data [24]. At inference time, given an unseen object shape BPS_o, we sample the latent space Z and decode our sample to generate a MANO grasp.

RefineNet. The grasps estimated by CoarseNet are plausible, but can be refined for improved contacts. For this, RefineNet takes as input the initial grasp ($\bar{\theta}$, $\bar{\gamma}$) and the distances D from MANO vertices to the object mesh. The distances are weighted according to the vertex contact likelihood learned from GRAB. Then, RefineNet estimates refined MANO parameters ($\hat{\theta}$, $\hat{\gamma}$) in 3 iterative steps as in [28], to give the final grasp. To train RefineNet, we generate a synthetic dataset; we sample CoarseNet grasps as ground truth and we perturb their hand pose parameters to simulate noisy input estimates. We use the same training losses as for CoarseNet.

GrabNet. Given an unseen 3D object, we first obtain an initial grasp estimate with CoarseNet, and pass this to RefineNet to obtain the final grasp estimate. For simplicity, the two networks are trained separately, but we expect end-to-end refinement to be beneficial, as in [24]. Figure 8 (right) shows some generated examples; our generations look realistic, as explained later in the evaluation section. For more qualitative results, see the video on our website and images in Sup. Mat.

Contact. As a free by-product of our 3D grasp predictions, we can compute contact between the 3D hand and object meshes, following Sec. 3.3. Contacts for GrabNet estimates are shown with red in Figure 8 (right). Other methods for contact prediction, like [5], are pure bottom-up approaches that label a vertex as in contact or not, without explicit reasoning about the hand structure. In contrast, we follow a top-down approach; we first generate a 3D grasping hand, and then compute contact with explicit anthropomorphic reasoning.

Evaluation - CoarseNet/RefineNet. We first quantitatively evaluate the two main components, by computing the reconstruction vertex-to-vertex error. For CoarseNet the errors are 12.1 mm, 14.1 mm and 18.4 mm for the training, validation and test set respectively. For RefineNet the errors are 3.7 mm, 4.1 mm and 4.4 mm. The results show that the components, that are trained separately, work reasonably well before plugging them together.

Evaluation - GrabNet. To evaluate GrabNet generated grasps, we perform a user study through AMT [1]. We take 6 test objects from the dataset and, for each object, we generate 20 grasps, mix them with 20 ground-truth grasps, and show them with a rotating 3D viewpoint to subjects. Then we ask

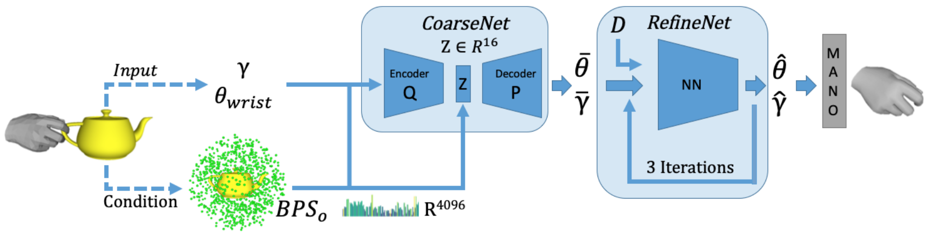

Fig. 7: GrabNet architecture. GrabNet generates MANO [56] grasps for unseen object shapes, encoded with a BPS [52] representation. It is comprised of two main modules. First, with CoarseNet we predict an initial plausible grasp. Second, we refine this with RefineNet to produce better contacts with the object.

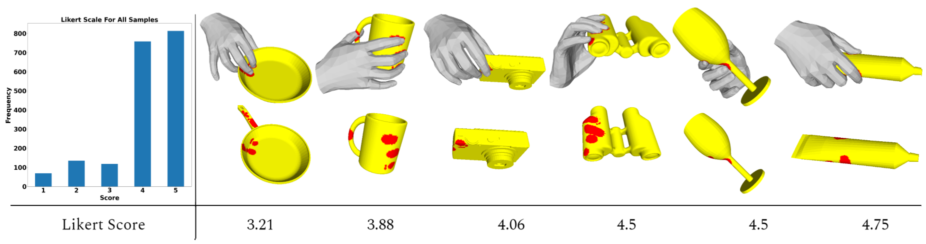

Fig. 8: Grasps generated by GrabNet for unseen objects; grasps look natural. As free by-product of 3D mesh generation, we get the red contact areas. For each grasp we show the average Likert score from all annotators. On the left the average Likert score is shown for all generated grasps.

participants how they agree with the statement "Humans can grasp this object as the video shows" on a 5-level Likert scale (5 is "strongly agree" and 1 is "strongly disagree"). To filter out the noisy subjects, namely the ones who do not understand the task or give random answers, we use catch trials that show implausible grasps. We remove subjects who rate these catch trials as realistic; see Sup. Mat. for details. Table 2 (left) shows the user scores for both ground-truth and generated grasps.

Evaluation - Contact. Figure 9 shows examples of contact areas (red) generated by [5] (left) and our approach (right). The method of [5] gives only 10 predictions per object, some with zero contact. Also, a hand is supposed to touch the whole red area; this is often not anthropomorphically plausible. Our contact is a product of MANO-based inference and is, by construction, anthropomorphically valid. Also, one can draw infinite samples from our learned grasping latent space. For further evaluation, we follow a protocol similar to [5] for our data. For every unseen test object we generate 20 grasps, and for each one we find both the closest ground-truth contact map and the closest ground-truth hand vertices, for comparison. Table 2 (right) reports the average error over all 20 predictions, in % for the former and cm for the latter case.

Table 2: GrabNet evaluation for 6 test objects. The "AMT" column shows user study results; grasp quality is rated from 1 (worst) to 5 (best). The "vertices" and "contact" columns evaluate grasps against the closest ground-truth one.

	AMT				Vertices	Contact
	Generation		Ground Truth		mean(cm)	%
Test Object	mean	std	mean	std	N=100	N=20
binoculars	4.09	0.93	4.27	0.80	2.56	4.00
camera	4.40	0.79	4.34	0.76	2.90	3.75
frying pan	3.19	1.30	4.49	0.67	3.58	4.16
mug	4.13	1.00	4.36	0.78	1.96	3.25
toothpaste	4.56	0.67	4.42	0.77	1.78	5.39
wineglass	4.32	0.88	4.43	0.79	1.92	4.56
Average	4.12	1.04	4.38	0.77	2.45	4.18

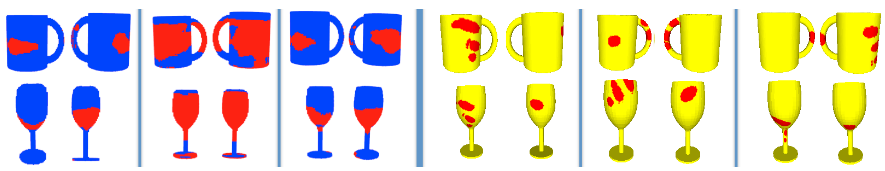

Fig. 9: Comparison of GrabNet (right) to ContactDB [5] (left). For each estimation we render two views, following the presentation style of [5].

5 Discussion

We provide a new dataset to the community that goes beyond previous motion capture or grasping datasets. We believe that GRAB will be useful for a wide range of problems. Here we show that it provides enough data and variability to train a novel network to predict object grasping, as we demonstrate with GrabNet. But there is much more that can be done. Importantly, GRAB includes the full body motion, enabling a much richer modeling than GrabNet.

Limitations: By focusing on accurate MoCap, we do not have synced image data. However, GRAB can support image-based inference [8, 28] by enabling rendering of synthetic human-object interaction [24, 53, 70] or learning priors to regularizing ill-posed inference of human-object interaction from 2D images [42].

Future Work: GRAB can support learning human-object interaction models [42, 61], robotic grasping from imitation [74], learning mappings from MoCap markers to meshes [20], rendering synthetic images [24, 53, 70], inferring object shape/pose from interaction [2, 44], or analysis of temporal patterns [48].

Acknowledgements: We thank S. Polikovsky, M. Höschle (MH) and M. Landry (ML) for the MoCap facility. We thank F. Mattioni, D. Hieber, and A. Valis for MoCap cleaning. We thank ML and T. Alexiadis for trial coordination, MH and F. Grimminger for 3D printing, V. Callaghan for voice recordings and J. Tesch for renderings.
Disclosure: In the last five years, MJB has received research gift funds from Intel, Nvidia, Facebook, and Amazon. He is a co-founder and investor in Meshcapade GmbH, which commercializes 3D body shape technology. While MJB is a part-time employee of Amazon, his research was performed solely at, and funded solely by, MPI.

References

1. Amazon Mechanical Turk. `https://www.mturk.com`
2. Behbahani, F.M., Singla-Buxarrais, G., Faisal, A.A.: Haptic SLAM: An ideal observer model for bayesian inference of object shape and hand pose from contact dynamics. In: Haptics: Perception, Devices, Control, and Applications (2016)
3. Bernardin, K., Ogawara, K., Ikeuchi, K., Dillmann, R.: A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. IEEE Transactions on Robotics (T-RO) **21**(1), 47–57 (2005)
4. Borras, J., Asfour, T.: A whole-body pose taxonomy for loco-manipulation tasks. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2015)
5. Brahmbhatt, S., Ham, C., Kemp, C.C., Hays, J.: ContactDB: Analyzing and predicting grasp contact via thermal imaging. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
6. Brubaker, M.A., Fleet, D.J., Hertzmann, A.: Physics-based person tracking using the anthropomorphic walker. International Journal of Computer Vision (IJCV) **87**(1), 140 (2009)
7. Chang, A.X., homas A. Funkhouser, eonidas J. Guibas, Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., Xiao, J., Yi, L., Yu, F.: ShapeNet: An information-rich 3D model repository. arXiv:1512.03012 (2015)
8. Choutas, V., Pavlakos, G., Bolkart, T., Tzionas, D., Black, M.J.: Monocular expressive body regression through body-driven attention. In: Proceedings of the European Conference on Computer Vision (ECCV) (2020)
9. Corona, E., Pumarola, A., Alenyà, G., Moreno-Noguer, F.: Context-aware human motion prediction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)
10. Cutkosky, M.R.: On grasp choice, grasp models, and the design of hands for manufacturing tasks. IEEE Transactions on Robotics and Automation **5**(3), 269–279 (1989)
11. Cyberglove III data glove. `http://www.cyberglovesystems.com/cyberglove-iii`
12. ElKoura, G., Singh, K.: Handrix: animating the human hand. In: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (2003)
13. Feit, A.M., Weir, D., Oulasvirta, A.: How we type: Movement strategies and performance in everyday typing. In: Proceedings of the CHI Conference on Human Factors in Computing Systems (2016)
14. Feix, T., Romero, J., Schmiedmayer, H.B., Dollar, A.M., Kragic, D.: The GRASP taxonomy of human grasp types. IEEE Transactions on Human-Machine Systems **46**(1), 66–77 (2016)
15. Garcia-Hernando, G., Yuan, S., Baek, S., Kim, T.K.: First-person hand action benchmark with rgb-d videos and 3d hand pose annotations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
16. GelSight tactile sensor. `http://www.gelsight.com`
17. Goldfeder, C., Ciocarlie, M.T., Dang, H., Allen, P.K.: The Columbia grasp database. In: IEEE International Conference on Robotics and Automation (ICRA) (2009)
18. Hamer, H., Gall, J., Weise, T., Van Gool, L.: An object-dependent hand pose prior from sparse training data. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2010)

19. Hampali, S., Oberweger, M., Rad, M., Lepetit, V.: HO-3D: A multi-user, multi-object dataset for joint 3D hand-object pose estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
20. Han, S., Liu, B., Wang, R., Ye, Y., Twigg, C.D., Kin, K.: Online optical marker-based hand tracking with deep labels. ACM Transactions on Graphics (TOG) **37**(4), 166:1166:10 (2018)
21. Handa, A., Wyk, K.V., Yang, W., Liang, J., Chao, Y.W., Wan, Q., Birchfield, S., Ratliff, N., Fox, D.: DexPilot: Vision based teleoperation of dexterous robotic hand-arm system. In: IEEE International Conference on Robotics and Automation (ICRA) (2019)
22. Hasler, N., Rosenhahn, B., Thormahlen, T., Wand, M., Gall, J., Seidel, H.: Markerless motion capture with unsynchronized moving cameras. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
23. Hassan, M., Choutas, V., Tzionas, D., Black, M.J.: Resolving 3D human pose ambiguities with 3D scene constrains. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
24. Hasson, Y., Varol, G., Tzionas, D., Kalevatykh, I., Black, M.J., Laptev, I., Schmid, C.: Learning joint reconstruction of hands and manipulated objects. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
25. Hsiao, K., Lozano-Perez, T.: Imitation learning of whole-body grasps. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (2006)
26. Johnson, M.K., Cole, F., Raj, A., Adelson, E.H.: Microgeometry capture using an elastomeric sensor. ACM Transactions on Graphics (TOG) **30**(4), 46:1–46:8 (2011)
27. Kamakura, N., Matsuo, M., Ishii, H., Mitsuboshi, F., Miura, Y.: Patterns of static prehension in normal hands. American Journal of Occupational Therapy **34**(7), 437–445 (1980)
28. Kanazawa, A., Black, M.J., Jacobs, D.W., Malik, J.: End-to-end recovery of human shape and pose. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
29. Karras, T.: Maximizing parallelism in the construction of bvhs, octrees, and k-d trees. In: Proceedings of the ACM SIGGRAPH/Eurographics Conference on High-Performance Graphics (2012)
30. Kim, V.G., Chaudhuri, S., Guibas, L., Funkhouser, T.: Shape2pose: Human-centric shape analysis. ACM Transactions on Graphics (TOG) **33**(4), 120:1–120:12 (2014)
31. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: International Conference on Learning Representations (ICLR) (2014)
32. Kjellstrom, H., Kragic, D., Black, M.J.: Tracking people interacting with objects. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
33. Kokic, M., Kragic, D., Bohg, J.: Learning task-oriented grasping from human activity datasets. IEEE Robotics and Automation Letters (RA-L) **5**(2), 3352–3359 (2020)
34. Kry, P.G., Pai, D.K.: Interaction capture and synthesis. ACM Transactions on Graphics (TOG) **25**(3), 872–880 (2006)
35. Le, H.V., Mayer, S., Bader, P., Henze, N.: Fingers' range and comfortable area for one-handed smartphone interaction beyond the touchscreen. In: Proceedings of the CHI Conference on Human Factors in Computing Systems (2018)
36. Lee, K.H., Choi, M.G., Lee, J.: Motion patches: building blocks for virtual environments annotated with motion data. ACM Transactions on Graphics (TOG) **25**(3), 898–906 (2006)

37. Li, Z., Sedlar, J., Carpentier, J., Laptev, I., Mansard, N., Sivic, J.: Estimating 3D motion and forces of person-object interactions from monocular video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
38. Mahmood, N., Ghorbani, N., F. Troje, N., Pons-Moll, G., Black, M.J.: AMASS: Archive of motion capture as surface shapes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2019)
39. Mandery, C., Terlemez, Ö., Do, M., Vahrenkamp, N., Asfour, T.: The KIT whole-body human motion database. In: International Conference on Advanced Robotics (ICAR) (2015)
40. Mascaro, S.A., Asada, H.H.: Photoplethysmograph fingernail sensors for measuring finger forces without haptic obstruction. IEEE Transactions on Robotics and Automation (TRA) **17**(5), 698–708 (2001)
41. Miller, A.T., Allen, P.K.: Graspit! A versatile simulator for robotic grasping. IEEE Robotics Automation Magazine (RAM) **11**(4), 110–122 (2004)
42. Monszpart, A., Guerrero, P., Ceylan, D., Yumer, E., Mitra, N.J.: iMapper: Interaction-guided scene mapping from monocular videos. ACM Transactions on Graphics (TOG) **38**(4), 92:1–92:15 (2019)
43. Napier, J.R.: The prehensile movements of the human hand. The Journal of bone and joint surgery **38**(4), 902–913 (1956)
44. Oberweger, M., Wohlhart, P., Lepetit, V.: Generalized feedback loop for joint hand-object pose estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **42**(8), 1898–1912 (2020)
45. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2011)
46. Pavlakos, G., Choutas, V., Ghorbani, N., Bolkart, T., Osman, A.A.A., Tzionas, D., Black, M.J.: Expressive body capture: 3D hands, face, and body from a single image. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
47. Pham, T., Kyriazis, N., Argyros, A.A., Kheddar, A.: Hand-object contact force estimation from markerless visual tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) **40**(12), 2883–2896 (2018)
48. Pirk, S., Krs, V., Hu, K., Rajasekaran, S.D., Kang, H., Yoshiyasu, Y., Benes, B., Guibas, L.J.: Understanding and exploiting object interaction landscapes. ACM Transactions on Graphics (TOG) **36**(3), 31:1–31:14 (2017)
49. Pollard, N.S., Zordan, V.B.: Physically based grasping control from example. In: Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation (2005)
50. POSER: 3D rendering and animation software. `https://www.posersoftware.com`
51. Pressure Profile Systems Inc. (PPS). `https://pressureprofile.com`
52. Prokudin, S., Lassner, C., Romero, J.: Efficient learning on point clouds with basis point sets. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
53. Ranjan, A., Hoffmann, D.T., Tzionas, D., Tang, S., Romero, J., Black, M.J.: Learning multi-human optical flow. International Journal of Computer Vision (IJCV) (jan 2020)
54. Rogez, G., Supančič III, J.S., Ramanan, D.: Understanding everyday hands in action from RGB-D images. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (2015)

55. Romero, J., Kjellström, H., Kragic, D.: Hands in action: real-time 3D reconstruction of hands in interaction with objects. In: IEEE International Conference on Robotics and Automation (ICRA) (2010)
56. Romero, J., Tzionas, D., Black, M.J.: Embodied hands: Modeling and capturing hands and bodies together. ACM Transactions on Graphics (TOG) **36**(6), 245:1–245:17 (2017)
57. Rosenhahn, B., Schmaltz, C., Brox, T., Weickert, J., Cremers, D., Seidel, H.: Markerless motion capture of man-machine interaction. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2008)
58. Ruff, H.A.: Infants' manipulative exploration of objects: Effects of age and object characteristics. Developmental Psychology **20**(1), 9 (1984)
59. Sahbani, A., El-Khoury, S., Bidaud, P.: An overview of 3D object grasp synthesis algorithms. Robotics and Autonomous Systems (RAS) **60**(3), 326336 (2012)
60. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: SceneGrok: Inferring action maps in 3D environments. ACM Transactions on Graphics (TOG) **33**(6), 212:1–212:10 (2014)
61. Savva, M., Chang, A.X., Hanrahan, P., Fisher, M., Nießner, M.: PiGraphs: Learning interaction snapshots from observations. ACM Transactions on Graphics (TOG) **35**(4), 139:1139:12 (2016)
62. Sridhar, S., Mueller, F., Zollhoefer, M., Casas, D., Oulasvirta, A., Theobalt, C.: Real-time joint tracking of a hand manipulating an object from RGB-D input. In: Proceedings of the European Conference on Computer Vision (ECCV) (2016)
63. Starke, S., Zhang, H., Komura, T., Saito, J.: Neural state machine for character-scene interactions. ACM Transactions on Graphics (TOG) **38**(6), 209:1–209:14 (2019)
64. Stratasys Fortus 360mc: 3D printing. `https://www.stratasys.com/resources/search/white-papers/fortus-360mc-400mc`
65. Sundaram, S., Kellnhofer, P., Li, Y., Zhu, J.Y., Torralba, A., Matusik, W.: Learning the signatures of the human grasp using a scalable tactile glove. Nature **569**(7758), 698–702 (2019)
66. Tekin, B., Bogo, F., Pollefeys, M.: H+O: Unified egocentric recognition of 3D hand-object poses and interactions. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2019)
67. Tekscan grip system: Tactile grip force and pressure sensing. `https://www.tekscan.com/products-solutions/systems/grip-system`
68. Tsoli, A., Argyros, A.A.: Joint 3d tracking of a deformable object in interaction with a hand. In: Proceedings of the European Conference on Computer Vision (ECCV) (2018)
69. Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., Gall, J.: Capturing hands in action using discriminative salient points and physics simulation. International Journal of Computer Vision (IJCV) **118**(2), 172–193 (2016)
70. Varol, G., Romero, J., Martin, X., Mahmood, N., Black, M.J., Laptev, I., Schmid, C.: Learning from synthetic humans. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2017)
71. Vicon Vantage: Cutting edge, flagship camera with intelligent feedback and resolution. `https://www.vicon.com/hardware/cameras/vantage`
72. Wang, Y., Min, J., Zhang, J., Liu, Y., Xu, F., Dai, Q., Chai, J.: Video-based hand manipulation capture through composite motion control. ACM Transactions on Graphics (TOG) **32**(4), 43:1–43:14 (2013)
73. Wang, Z., Chen, L., Rathore, S., Shin, D., Fowlkes, C.: Geometric pose affordance: 3D human pose with scene constraints. arXiv:1905.07718 (2019)

74. Welschehold, T., Dornhege, C., Burgard, W.: Learning manipulation actions from human demonstrations. In: IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (2016)
75. XSENS: Inertial motion capture. `https://www.xsens.com/motion-capture`
76. Yamamoto, M., Yagishita, K.: Scene constraints-aided tracking of human body. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2000)
77. Ye, Y., Liu, C.K.: Synthesis of detailed hand manipulations using contact sampling. ACM Transactions on Graphics (TOG) **31**(4), 41:141:10 (2012)
78. Zhang, H., Bo, Z.H., Yong, J.H., Xu, F.: InteractionFusion: real-time reconstruction of hand poses and deformable objects in hand-object interactions. ACM Transactions on Graphics (TOG) **38**(4), 48:1–48:11 (2019)
79. Zhang, Y., Hassan, M., Neumann, H., Black, M.J., Tang, S.: Generating 3D people in scenes without people. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2020)