Understanding Everyday Hands in Action from RGB-D Images

Gregory Rogez, James Steven Supancic, Deva Ramanan

Understanding Everyday Hands in Action from RGB-D Images

Grégory Rogez
Inria Rhône-Alpes
gregory.rogez@inria.fr

James S. Supančič III
University of California, Irvine
jsupanci@ics.uci.edu

Deva Ramanan
Carnegie Mellon University
deva@cs.cmu.edu

Abstract

We analyze functional manipulations of handheld objects, formalizing the problem as one of fine-grained grasp classification. To do so, we make use of a recently developed fine-grained taxonomy of human-object grasps. We introduce a large dataset of 12000 RGB-D images covering 71 everyday grasps in natural interactions. Our dataset is different from past work (typically addressed from a robotics perspective) in terms of its scale, diversity, and combination of RGB and depth data. From a computer-vision perspective, our dataset allows for exploration of contact and force prediction (crucial concepts in functional grasp analysis) from perceptual cues. We present extensive experimental results with state-of-the-art baselines, illustrating the role of segmentation, object context, and 3D-understanding in functional grasp analysis. We demonstrate a near 2X improvement over prior work and a naive deep baseline, while pointing out important directions for improvement.

1. Introduction

Humans can interact with objects in complex ways, including grasping, pushing, or bending them. In this work, we address the perceptual problem of parsing such interactions, with a focus on handheld, manipulatable objects. Much previous work on hand analysis tends to focus on kinematic pose estimation [17, 12]. Interestingly, the same kinematic pose can be used for dramatically different *functional manipulations* (Fig. 1), where differences are manifested in terms of distinct contact points and force vectors. Thus, contact points and forces play a crucial role when parsing such interactions from a functional perspective.

Problem setup: Importantly, we wish to analyze human-object interactions *in situ*. To do so, we make use of wearable depth cameras to ensure that recordings are mobile (allowing one to capture diverse scenes [33, 7]) and passive (avoiding the need for specialized pressure sensors/gloves [6, 24]). We make no explicit assumption about the environment, such as known geometry [32]. However, we do make explicit use of depth cues, motivated by the fact that humans make use of depth for near-field analysis [15]. Our problem formulation is thus: given a first-person RGB-D image of a hand-object interaction, predict the 3D kinematic hand pose, contact points, and force vectors.

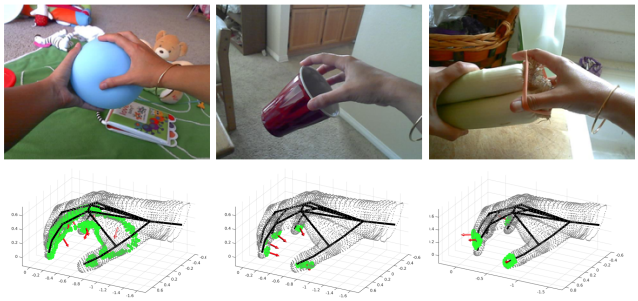


Figure 1. **Same kinematic pose, but different functions:** We show 3 images of near-identical kinematic hand pose, but very different functional manipulations, including a wide-object grasp (**a**), a precision grasp (**b**), and a finger extension (**c**). Contact regions (green) and force vectors (red), visualized below each image, appear to define such manipulations. This work (1) introduces a large-scale dataset for predicting pose+contacts+forces from images and (2) proposes an initial method based on fine-grained grasp classification.

Motivation: We see several motivating scenarios and applications. Our long-term goal is to produce a truly functional description of a scene that is useful for an autonomous robot. When faced with a novel object, it will be useful to know how if it can be pushed or grasped, and what forces and contacts are necessary to do so [40]. A practical application of our work is imitation learning or learning by demonstration for robotics [3, 16], where a robot can be taught a task by observing humans performing it. Finally, our problem formulation has direct implications for assistive technology. Clinicians watch and evaluate patients performing everyday hand-object interactions for diagnosis and evaluation [2]. A patient-wearable camera that enabled automated parsing of object manipulations would allow for long-term monitoring.

Why is this hard? Estimating forces from visual signals typically requires knowledge of object mass and velocity, which is difficult to reliably infer from a single image or

even a video sequence. Isometric forces are even more difficult to estimate because no motion may be observed. Finally, even traditional tasks such as kinematic hand pose estimation are now difficult because manipulated objects tend to generate significant occlusions. Indeed, much previous work on kinematic hand analysis considers isolated hands in free-space [43], which is a considerably easier problem.

Approach: We address the continuous problem of pose+contact+force prediction as a discrete fine-grained classification task, making use of a recent 73-class taxonomy of fine-grained hand-object interactions developed from the robotics community [28]. Our approach is inspired by prototype-based approaches for continuous shape estimation that treat the problem as a discrete categorical prediction tasks, such as shapemes [34] or poselets [5]. However, rather than learning prototypes, we make use of expert domain knowledge to quantize the space of manipulations, which allows us to treat the problem as one of (fine-grained) classification. A vital property of our classification engine is that it is data-driven rather than model-based. We put forth considerable effort toward assembling a large collection of diverse images that span the taxonomy of classes. We experiment with both parametric and exemplar-based classification architectures trained on our collection.

Our contributions: Our primary contribution is (1) a new "in-the-wild", large-scale dataset of fine-grained grasps, annotated with contact points and forces. Importantly, the data is RGB-D and collected from a wearable perspective. (2) We develop a pipeline for fine-grained grasp classification exploiting depth and RGB data, training on combinations of both real and synthetic training data and making use of state-of-the-art deep features. Overall, our results indicate that grasp classification is challenging, with accuracy approaching 20% for a 71-way classification problem. (3) We describe a simple post-processing exemplar framework that predicts contacts and forces associated with hand manipulations, providing an initial proof-of-concept system that addresses this rather novel visual prediction task.

2. Related Work

Hand pose with RGB(D): Hand pose estimation is a well-studied task, using both RGB and RGB-D sensors as input. Much work formulates the task as articulated tracking over time [25, 23, 22, 4, 31, 42, 44], but we focus on single-image hand pose estimation during object manipulations. Relatively few papers deal with object manipulations, with the important exceptions of [39, 38, 27, 26]. Most similar to us is [32], who estimate contact forces during hand-object interactions, but do so in a "in-the-lab" scenario where objects of known geometry are used. We focus on single-frame "in-the-wild" footage where the observer is instrumented, but the environment (and its constituent objects) are not.

Egocentric hand analysis: Spurred by the availability of cheap wearable sensors, there has been a considerable amount of recent work on object manipulation and grasp analysis from egocentric viewpoints [11, 8, 18, 7, 13]. The detection and pose estimation of human hands from wearable cameras was explored in [36]. [8] propose a fully automatic vision-based approach for grasp analysis from a wearable RGB camera, while [18] explores unsupervised clustering techniques for automatically discovering common modes of human hand use. Our work is very much inspired by such lines of thought, but we take a data-driven perspective, focusing on large-scale dataset collection guided by a functional taxonomy.

Grasp taxonomies: Numerous taxonomies of grasps have been proposed, predominantly from the robotics community. Early work by Cutkosky [9] introduced 16 grasps, which were later extended to 33 by Felix et al [14], following a definition of a grasp as a "static hand postures with which an object can be held with one hand". Though this excluded two-handed, dynamic, and gravity-dependent grasps, this taxonomy has been widely used [37, 8, 7]. Our work is based on a recent fine-grained taxonomy proposed in [28], that significantly broadens the scope of manipulations to include *non-prehensile* object interactions (that are technically not grasps, such as pushing or pressing) as well as other gravity-dependent interactions (such as lifting). The final taxonomy includes 73 grasps that are annotated with various qualities (including hand shape, force type, direction of movement and effort).

Datasets. Because grasp understanding is usually addressed from a robotics perspective, the resulting methods and datasets developed for the problem tend to be tailored for that domain. For example, robotics platforms often require an unavoidable real-time constraint, limiting the choice of algorithms, which also (perhaps implicitly) limited the difficulty of the data in terms of diversity (few subjects, few objects, few scenes). We overview the existing grasp datasets in Table 1 and tailor our new dataset to "fill the gap" in terms of overall scale, diversity, and annotation detail.

Dataset	View	Cam.	Sub.	Scn	Frms	Label	Tax.
YALE [7]	Ego	RGB	4	4	9100	Gr.	33
UTG [8]	Ego	RGB	4	1	?	Gr.	17
GTEA [13]	Ego	RGB	4	4	00	Act.	7
UCI-EGO [36]	Ego	RGB-D	2	4	400	Pose	?
Ours	Ego	RGB-D	8	> 5	$12,000$	Gr.	71

Table 1. **Object manipulation datasets.** [7] captured 27.7 hours but labelled only 9100 frames with grasp annotations. While our dataset is balanced and contains the same amount of data for each grasp, [7] is imbalanced in that common grasps appear much more often than rare grasps (10 grasps suffice to explain 80% of the data). [8] uses the same set of objects for the 4 subjects.

Figure 2. **GUN-71: Grasp Understanding dataset.** We have captured (from a chest-mounted RGB-D camera) and annotated our own dataset of fine-grained grasps, following the recent taxonomy of [28]. In the **top** row, the "writing tripod" grasp class exhibits low variability in object and pose/view across 6 different subjects and environments. In the **second** row, "flat hand cupping" exhibits high variability in objects and low variability in pose due to being gravity-dependent. In the **third** row, "trigger press" exhibits high variability in objects and pose/view. Finally, in **bottom**, we show 6 views of the same grasp captured for a particular object and a particular subject in our dataset.

3. GUN-71: Grasp UNderstanding Dataset

We begin by describing our dataset, visualized in Fig. 2. We start with the 73-class taxonomy of [28], but omit grasps 50 and 61 because of their overly-specialized nature (holding a ping-pong racket and playing saxophone, respectively), resulting in 71 classes.

3.1. Data capture

To capture truly in-the-wild data, we might follow the approach of [7] and monitor unprompted subjects behaving naturally throughout the course of a day. However, this results in a highly imbalanced distribution of observed object manipulations. [7] shows that 10 grasps suffice to explain 80% of the object interactions of everyday users. Balanced class distributions arguably allow for more straightforward analysis, which is useful when addressing a relatively unexplored problem. Collecting a balanced distribution in such a unprompted manner would be prohibitively expensive, both in terms of raw data collection and manual annotation. Instead, we prompt users using the scheme below.

Capture sessions: We ask subjects to perform the 71 grasps on personal objects (typical for the specific grasp), mimicking real object manipulation scenarios in their home environment. Capture sessions were fairly intensive and laborious as a result. We mount Intel's Senz3D, a wearable time-of-flight sensor [20, 10, 29], on the subjects's chest using a GoPro harness (as in [36]). We tried to vary the types of objects as much as possible and considered between 3 and 4 different objects per subject for each of the 71 grasps. For each hand-object configuration, we took between 5 and 6 views of the manipulation scene. These views correspond to several steps of a particular action (opening a lid, pouring water) as well as different 3D locations and orientation of the hand holding the object (with respect to the camera).

Diversity: This process led to the capture of roughly 12,000 RGB-D images labeled with one of the 71 grasps. We captured 28 objects per grasp, resulting in $28 \times 71 = 1988$ different hand-object configurations with 5-6 views for each. We consider 8 different subjects (4 males and 4 females) in 5 different houses, ensuring that "house mates" avoid using the same objects to allow leave-one-out experiments (we can leave out one subject for testing and ensure that the objects will be novel as well). Six of our eight subjects were right handed. To ensure consistency, we asked the two left-handed subjects to perform grasps with their right hand. We posited that body shape characteristics might effect accuracy/generalizability, particularly in terms of hand size, shape, and movement. To facilitate such analysis, we also measured arm and finger lengths for each subject.

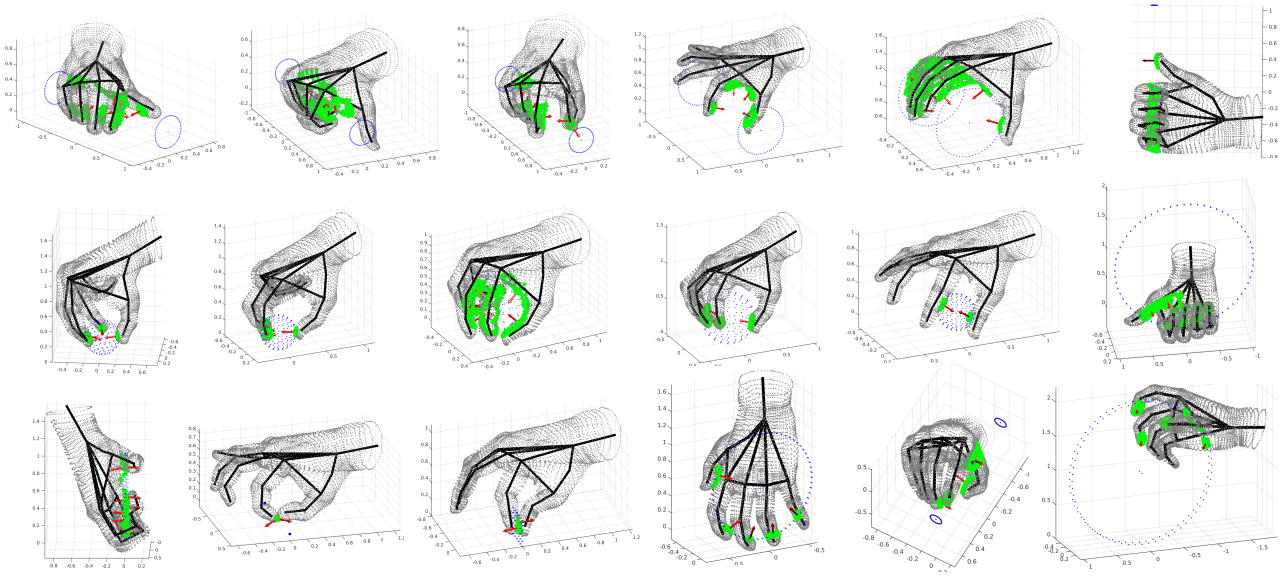


Figure 3. **Contact point and forces.** We show the 3D hand model for 18 grasps of the considered taxonomy. We also show the contact points (in green) and forces (in red) corresponding to each grasp. The blue points help visualize the shape of the typical object associated with each of these 18 grasps. We can observe that power grasps have wider contact areas on finger and palm regions, while precision grasps exhibit more localized contact points on finger tips and finger pads.

4. Synthetic (training) data generation

3D hand-object models: In addition to GUN-71, we construct a synthetic training dataset that will be used during our grasp-recognition pipeline. To construct this synthetic dataset, we make use of synthetic 3D hand models. We obtain a set of 3D models by extending the publicly-available Poser models from [39] to cover the selected grasps from [28]'s taxonomy (by manually articulating the models to match the visual description of grasp).

Contact and force annotations: We compute contact points and applied forces on our 3D models with the following heuristic procedure. First, we look for physical points of contact between the hand and object mesh. We do this by intersecting the triangulated hand and object meshes with the efficient method of [30]. We produce a final set of contact regions by connected-component clustering the set of 3D vertices lying within an intersection boundary. To estimate a force vector, we assume that contact points are locally stable and will not slide along the surface of the object (implying the force vector is normal to the surface of the object). We estimate this normal direction by simply reporting the average normal of vertices within each contact region. Note this only produces an estimate of the force direction, and not magnitude. Nevertheless, we find our heuristic procedure to produce surprisingly plausible estimates of contact points and force directions for each 3D model (Fig. 3).

Synthetic training data: We use our 3D models to generate an auxiliary dataset of synthetic depth data, annotated with 3D poses, grasp class label, contacts, and force direction vectors. We additionally annotate each rendered depth map with a segmentation mask denoting background, hand, and object pixels. We render over $200,000$ training instances ($3,000$ per grasp). We will release our models, rendering images, as well as GUN-71 (our dataset of real-world RGB-D images) to spun further research in the area.

5. Recognition pipeline

We now describe a fairly straightforward recognition system for recognizing grasps given real-world RGB-D images. Our pipeline consists of two stages; hand segmentation and fine-grained grasp classification.

5.1. Segmentation

The first stage of our pipeline is segmenting the hand from background clutter, both in the RGB and depth data. Many state-of-the-art approaches [8, 38, 39] employ user-specific skin models to localize and segment out the hand. We want a system that does not require such a user-specific learning stage and could be applied to any new user and environment, and so instead make use of depth cues to segment out the hand.

Depth-based hand detection: We train a P-way classifier designed to report one of $P = 1500$ quantized hand poses, using the approach of [35]. This classifier is trained on the synthetic training data, which is off-line clustered into P pose classes. Note that the set of pose classes P

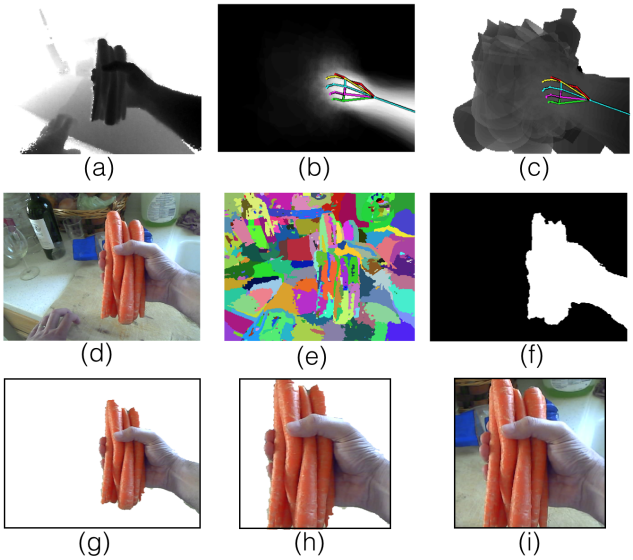


Figure 4. **Segmentation.** We show the different steps of our segmentation stage: the depth map (**a**) is processed using a K-way pose classifier [35], which reports a quantized pose detection k and associated foreground prior b_{ik} (**b**) and mean depth μ_{ik}(**c**) (used to compute a posterior following Eq. 1). To incorporate bottom-up RGB cues, we first extract superpixels (**e**) and then label superpixels instead of pixels to produce a segmentation mask (**f**). This produces a segmented RGB image in (**g**) , which can then be cropped (**h**) and/or unsegmented (**i**). We concatenate (deep) features extracted from (d), (g), (h), and (i) to span a variety of resolutions and local/global contexts.

is significantly larger than the set of fine-grained grasps $K = 71$. We use the segmentation mask associated with this coarse quantized pose detection to segment out the hand (and object) from the test image, described further below.

Pixel model: We would like to use hand detections to generate binary segmentation masks. To do so, we use a simple probabilistic model where x_i denotes the depth value of pixel i and $y_i \in \{0, 1\}$ is its binary foreground/background label. We write the posterior probability of label y_i given observation x_i, all conditioned on pose class k as:

$$p(y_i|x_i, k) \propto p(y_i|k)p(x_i|y_i, k) \quad (1)$$

which can easily derived from Bayes rule . The first term on the right-hand-side is the "prior" probability of pixel i being fg/bg, and the second term is a "likelihood" of observing a depth value given a pose class k and label:

$$p(y_i = 1|k) = b_{ik} \quad Bernoulli \quad (2)$$

$$p(x_i|y_i = 1, k) = N(x_i; \mu_{ik}, \sigma^2_{ik}) \quad Normal \quad (3)$$

$$p(x_i|y_i = 0, k) \propto \text{constant} \quad Uniform \quad (4)$$

We use a pixel-specific Bernoulli distribution for the prior, and an univariate Normal and Uniform (uninformative) distribution for the likelihood. Intuitively, foreground depths tend to be constrained by the pose, while the background will not be. Given training data of depth images x with foreground masks y and pose class labels k, it is straightforward to estimate model parameters $\{b_{ik}, \mu_{ik}, \sigma_{ik}\}$ with maximum likelihood estimation (frequency counts, sample means, and sample variances). We visualize the pixel-wise Bernoulli prior b_{ik} and mean depth μ_{ik} for a particular class k in Fig. 4-b and Fig. 4-c.

RGB-cues: Thus far, our segmentation model does not make use of RGB-based grouping cues such as color changes across object boundaries. To do so, we first compute RGB-based superpixels [1] on a test image and reason about the binary labels of superpixels rather than pixels:

$$label_j = I\Big(\frac{1}{|S_j|}\sum_{i \in S_j} p(y_i|x_i, k) > .5\Big) \quad (5)$$

where S_j denotes the set of pixels from superpixel j. We show a sample segmentation in Fig. 4. Our probabilistic approach tends to produce more reliable segmentations than existing approaches based on connected-component heuristics [19].

5.2. Fine-grained classification

We use the previous segmentation stage to produce features that will be fed into a $K = 71$-way classifier. We use state-of-the-art deep networks – specifically, Deep19 [41] – to extract a 3096 dimensional feature. We extract off-the-shelf deep features extracted for (1) the entire RGB image, (2) a cropped window around the detected hand, and (3) a segmented RGB image (Fig. 4 (d,g,h,i)). We resize each window to a canonical size (of 224 x 224 pixels) before processing. The intuition behind this choice is to mix high and low resolution features, as well as global (contextual) and local features. The final concatenated descriptors are fed into a linear multi-class SVM for processing.

Exemplar matching: The above stages return an estimate for the employed grasp and a fairly accurate quantized pose class, but it is still quantized nonetheless. One can refine this quantization by returning the closest synthetic training example belonging to the recognized grasp and the corresponding pose cluster. We do this by returning the training example n from quantized class k with the closest foreground depth:

$$NN(x) = \min_{n \in Class_k} \sum_i y_i^n (x_i^n - \mu_{ik})^2 \quad (6)$$

We match only foreground depths in the n^{th} synthetic training image x^n, as specified by its binary label y^n. Because each synthetic exemplar is annotated with hand-object contact points and forces from its parent 3D hand model, we can predict forces and contact points by simply transferring them from the selected grasp model to the exemplar location in the 3D space.

6. Experiments

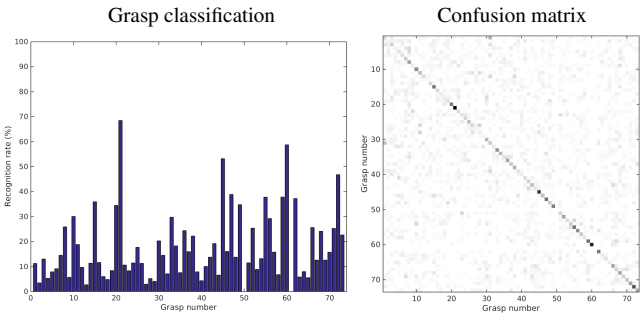
For all the experiments of this section, we use a leave-one-out approach where we train our 1-vs-all SVM classifiers on 7 subjects and test on the last 8^{th} subject. We repeat that operation with the 8 subjects and average the results. When analyzing our results, we refer to grasps by their id#. In the supplementary material, we include a visualization of all grasps in our taxonomy.

Baselines: We first run some "standard" baselines: HOG-RGB, HOG-Depth, and an off-the-shelf deep RGB feature [41]. We obtained the following average classification rate: HOG-RGB (3.30%), HOG-Depth (6.55%), concatenated HOG-RGB and HOG-Depth (6.55%) and Deep-RGB (11.31%). Consistent with recent evidence, deep features considerably outperform their hand-designed counterparts, though overall performance is still rather low (Tab. 2).

Segmented/cropped data: Next, we evaluate the role of context and clutter. Using segmented RGB images marginally decreases accuracy of deep features from 11.31% to 11.10%, but recognition rates appear are more homogeneous. Looking at the individual grasp classification rates, segmentation helps a little for most grasps but hurts the accuracy of "easy" grasps where context or object shape are important (but removed in the segmentation). This includes non-prehensile "pressing" grasps (interacting with a keyboard) and grasps associated with unique objects (chopsticks). Deep features extracted from a cropped segmentation and cropped detection increase accuracy to 12.55% and 13.67%, respectively, suggesting that some amount of local context around the hand and object helps.

Competing methods: [38, 8] make use of HOG templates defined on segmented RGB images obtained with skin detection. Because skin detectors did not work well on our (in-the-wild) dataset, we re-implemented [8] using HOG templates defined on our depth-based segmentations and obtained 7.69% accuracy. To evaluate recent non-parametric methods [38], we experimented with a naive nearest neighbor (NN) search using the different features extracted for the above experiments and obtained 6.10%, 6.97%, 6.31% grasp recognition accuracy using Deep-RGB, cropped-RGB and cropped+segmented-RGB. For clarity, these replace the K-way SVM classifier with a NN search. The significant drop in performance suggests that the learning is important, implying that our dataset is still not big enough to cover all possible variation in pose, objects and scenes.

Cue-combination: To take advantage of detection and segmentation without hurting classes where context is important, we trained our SVM grasp classifier on the concatenation of all the deep features. Our final overall classification rate of 17.97% is a considerable improvement over a naive deep model 11.31% as well as (our reimplementation of) prior work 7.69%. The corresponding recognition rates per grasp and confusion matrices corresponding to this classifier are given in Fig. 5.



(a) (b)

Figure 5. **RGB Deep feature + SVM.** We show the individual classification rates for the 71 grasps in our dataset (**a**) and the corresponding confusion matrix in (**b**).

Features	Acc.	top 20	top 10	min	max
HOG-RGB	3.30	7.20	9.59	0.00	28.54
HOG-Depth	6.55	12.96	15.74	0.66	26.18
HOG-RGBD	6.54	13.76	19.24	0.00	45.62
Deep-RGB [41]	11.31	25.92	35.28	0.69	61.39
Deep-RGB(segm.)	11.10	21.56	26.51	0.69	29.46
HOG-RGB (cropped)	5.84	11.22	14.03	0.00	27.85
Deep-RGB (cropped)	13.67	27.32	36.95	1.22	55.35
HOG-RGB (crop.+segm.) [8]	7.69	15.23	18.65	0.69	30.77
HOG-Depth (crop.+segm.)	10.68	22.04	27.99	0.52	42.40
Deep-RGB (crop.+segm.)	12.55	22.89	27.85	0.69	37.49
Deep-RGB (All)	**17.97**	**36.20**	**44.97**	**2.71**	**68.48**

Table 2. **Grasp classification results.** We present the result obtained when training a K-way linear SVM (K=71) with different types of features: HOG-RGB, HOG-Depth and Deep-RGB features, on the whole workspace, i.e. entire image, on a cropped detection window or on cropped and segmented image.

View	Acc.	top 20	top 10	min	max
All (All)	17.97	36.20	44.97	**2.71**	68.48
Best scoring view	**22.67**	**47.53**	**59.04**	0	**79.37**

Table 3. **View selection.** We present grasp recognition results obtained when training a K-way linear SVM on a concatenation of Deep features. We present the results obtained when computing the average classification rate over 1) the entire dataset and 2) over the top scoring view of each hand-object configuration.

	71 Gr. [28]	33 Gr. [14]	17 Gr. [9]
All views	17.97	20.50	20.53
Best scoring view	22.67	21.90	23.44

Table 4. **Grasp classification for different sized taxonomies.** We present the results obtained for $K = 71$ [28], $K = 33$ [14] and $K = 17$ [9], smaller taxonomies being obtained by selecting the corresponding subsets of grasps.

Easy cases: High-performing grasp classes (Fig. 5) tend to be characterized by limited variability in terms of viewpoint (i.e. position and orientation of the hand w.r.t camera)

and/or object: eg. opening a lid (#10), writing (#20), holding chopsticks (#21), measuring with a tape (#33), grabbing a large sphere such as a basketball (#45), using screwdriver (#47), trigger press (#49), using a keyboard (#60), thumb press (#62), holding a wine cup (#72). Other high-performing classes tend to exhibit limited occlusions of the hand: hooking a small object(#15) and palm press (#55).

Common confusions: Common confusions in Fig. 6 suggest that finger kinematics are a strong cue captured by deep features. Many confusions correspond to genuinely similar grasps that differ by small details that might be easily occluded by the hand or the manipulated object: "Large diameter" (#1) and "Ring" (#31) are both used to grasp cylindrical objects, except that "Ring" only uses thumb and index finger. When the last three fingers are fully occluded by the object, it is visually impossible to differentiate them (see Fig. 6-c). "Adduction-Grip" (#23) and "Middle-over-Index"(#51) both involve grasping an object using the index and middle finger. Abduction-Grip holds the object between the two fingers, while Middle-over-Index holds the object using the pad of the middle finger and nail of the index finger (see Fig. 6-f).

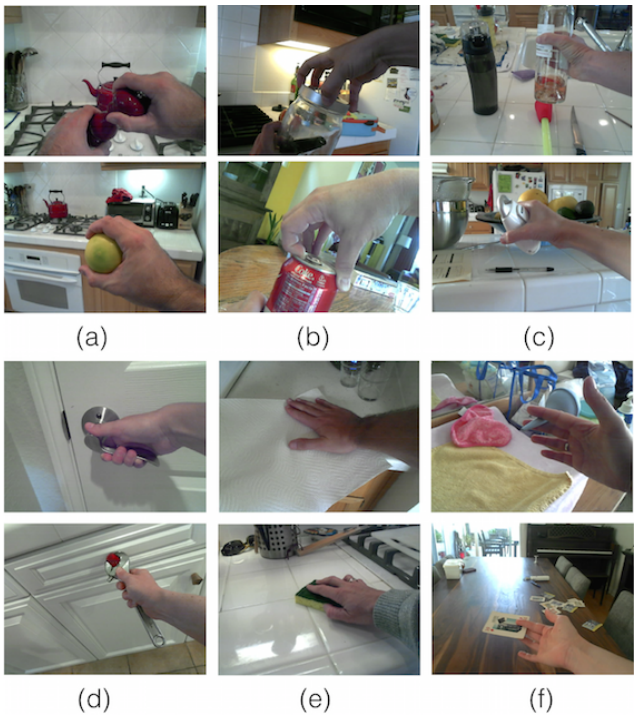


Figure 6. **Common confusions.** The confusions occur when some fingers are occluded (**a** and **c**) or when the poses are very similar but the functionality (associated forces and contact points) is different (**b**, **d**, **e** and **f**).

Best view: To examine the effect of viewpoint, we select the top-scoring view for each grasp class, increasing accuracy from 17.97% to 22.67% (Tab. 3). Comparing the two sets of recognition rates, best-view generally increases the performance of easy grasps significantly more than difficult ones - e.g., the average recognition rate of the top 20 grasps grow from 36.20% to 47.53%, while the top 10 grasps grows from 44.97% to 59.04%. This suggests that some views may be considerably more ambiguous than others.

Comparison to state-of-the-art. We now compare our results to those systems evaluated on previous grasp datasets. Particularly relevant is [8], which presents visual grasp recognition results in similar settings, i.e. egocentric perspective and daily activities. In their case, they consider a reduced 17-grasp taxonomy from Cutkosky [9], obtaining 31% with HOG features overall and 42% on a specific "machinist sequence" from [7]. Though these results appear more accurate than ours, its important to note that their dataset contains less variability in the background and scenes, and, crucially, their system appears to require training a skin detector on a subset of the test set. Additionally, it is not clear if they (or indeed, other past work) allow for the same subject/scene to be included across the train and testset. If we allow for this, recognition rate dramatically increases to 85%. This highly suggestive of overfitting, and can be seen a compelling motivation for the distinctly large number of subjects and scenes that we capture in our dataset.

Evaluations on limited taxonomies: If we limit our taxonomy to the 17 grasps from [8], i.e. by evaluating only the subset of 17 classes, we obtain 20.53% and 23.44% (best view). See Tab. 4. These numbers are comparable to those reported in [8]. Best-view may be a fair comparison because [7] used data that was manually labelled, where annotators were explicitly instructed to only annotate those frames that were visually unambiguous. In our case, subjects were asked to naturally perform object manipulations, and the data was collected "as-is". Finally, if we limit our taxonomy to the 33 grasps from Feix et al. [14], we obtained 20.50% and 21.90% (best view). The marginal improvement when evaluating grasps from smaller taxonomies suggests that the new classes are not much harder to recognize. Rather, we believe that overall performance is somewhat low because our dataset is genuinely challenging due to diverse subjects, scenes, and objects.

Force and contact point prediction: Finally, we present preliminary results for force and contact prediction. We do so by showing the best-matching synthetic 3D exemplar from the detected pose class, along with its contact and force annotations. Fig. 7 shows frames for which the entire pipeline detection + grasp recognition + exemplar matching led to an acceptable prediction. Unfortunately, we are not able to provide a numerical evaluation as obtaining ground-truth annotation of contact and forces is challenging. One attractive option is to use active force sensors,

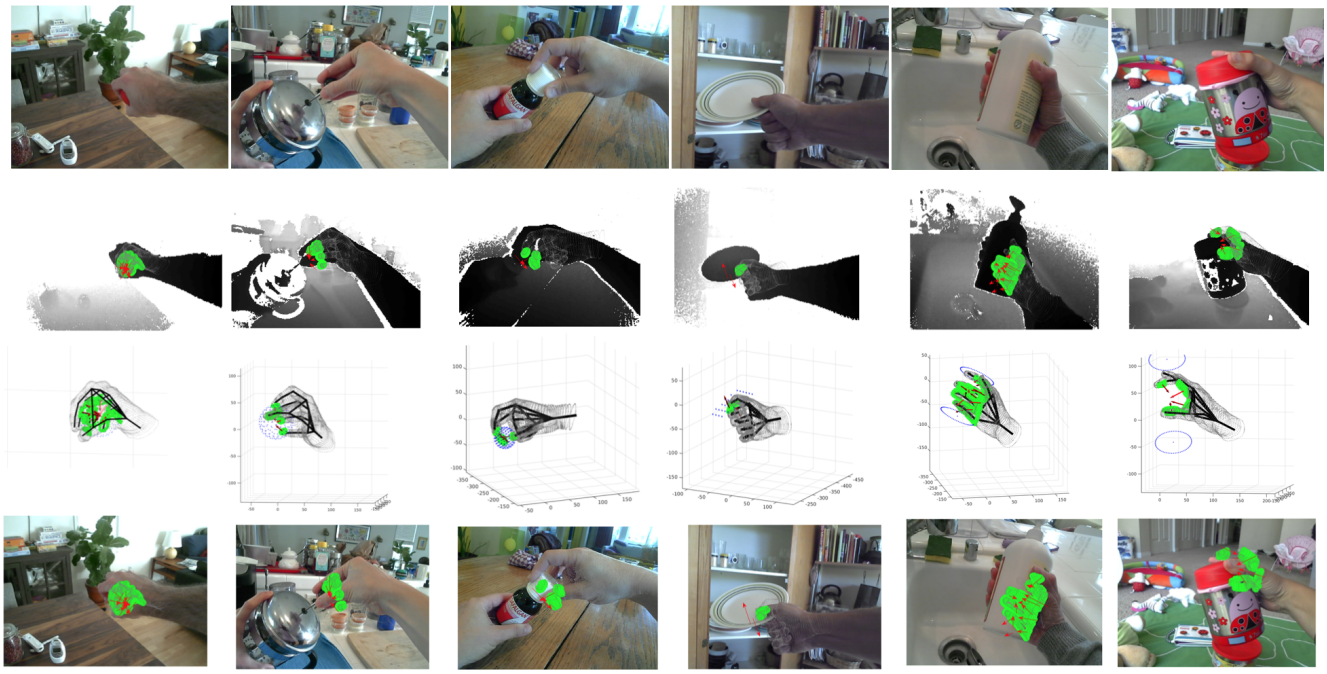


Figure 7. **Force and contact points prediction.** We show frames for which the entire pipeline detection + grasp recognition + exemplar matching led to an acceptable prediction of forces and contact points. For each selected frame, we show from top to bottom: the RGB image, the depth image with contact points and forces (respectively represented by green points and red arrows, the top scoring 3D exemplar with associated forces and contact points, and finally the RGB image with overlaid forces and contact points.

either embedded into pressure-sensitive gloves worn by the user or through objects equipped with force sensors at predefined grasp points (as done for a simplified cuboid object in [32]). While certainly attractive, active sensing somewhat violates the integrity of a truly in-the-wild, everyday dataset.

7. Conclusions

We have introduced the challenging problem of understanding hands in action, including force and contact point prediction, during scenes of in-the-wild, everyday object manipulations. We have proposed an initial solution that reformulates this high-dimensional, continuous prediction task as a discrete fine-grained (functional grasp) classification task. To spur further research, we have captured a new large scale dataset of fine-grained grasps that we will release together with 3D models and rendering engine. Importantly, we have captured this dataset from an egocentric perspective, using RGB-D sensors to record multiple scenes and subjects. We have also proposed a pipeline which exploits depth and RGB data, producing state-of-the-art grasp recognition results. Our first analysis show that depth information is crucial for detection and segmentation, while the richer RGB feature allows for a better grasp recognition. Overall, our results indicate that grasp classification is challenging, with accuracy approaching 20% for a 71-way classification problem.

We have used a single 3D model per grasp. In future work, it would be interesting to (1) model within-grasp variability, capturing the dependence of hand kinematics on object shape and size and (2) consider subject-specific 3D hand shape models [21], which could lead to more accurate set of synthetic exemplars (and associated forces and contacts).

Acknowledgement. G. Rogez was supported by the European Commission under FP7 Marie Curie IOF grant "Egovision4Health" (PIOF-GA-2012-328288). J. Supancic and D. Ramanan were supported by NSF Grant 0954083, ONR-MURI Grant N00014-10-1-0933, and the Intel Science and Technology Center - Visual Computing. We thank our hand models Allon H., Elisabeth R., Marga C., Nico L., Odile H., Santi M. and Sego L. for participating in the data collection.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. Slic superpixels compared to state-of-the-art superpixel methods. *PAMI*, 34(11):2274–2282, 2012.

[2] S. Allin and D. Ramanan. Assessment of Post-Stroke Functioning using Machine Vision. In *MVA*, 2007.

[3] B. D. Argall, S. Chernova, M. Veloso, and B. Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5):469–483, 2009.

[4] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *CVPR (2)*, pages 432–442, 2003.

[5] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.

[6] M. Bouzit, G. Burdea, G. Popescu, and R. Boian. The rutgers master ii-new design force-feedback glove. *Mechatronics, IEEE/ASME Transactions on*, 7(2):256–263, 2002.

[7] I. M. Bullock, T. Feix, and A. M. Dollar. The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. *I. J. Robotic Res.*, 34(3):251–255, 2015.

[8] M. Cai, K. M. Kitani, and Y. Sato. A scalable approach for understanding the visual structures of hand grasps. In *ICRA*, 2015.

[9] M. R. Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *IEEE T. Robotics and Automation*, 5(3):269–279, 1989.

[10] D. Damen, A. P. Gee, W. W. Mayol-Cuevas, and A. Calway. Egocentric real-time workspace monitoring using an rgb-d camera. In *IROS*, 2012.

[11] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. W. Mayol-Cuevas. You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014.

[12] A. Erol, G. Bebis, M. Nicolescu, R. D. Boyle, and X. Twombly. Vision-based hand pose estimation: A review. *CVIU*, 108(1-2):52–73, 2007.

[13] A. Fathi, X. Ren, and J. Rehg. Learning to recognize objects in egocentric activities. In *CVPR*, 2011.

[14] T. Feix, R. Pawlik, H. Schmiedmayer, J. Romero, and D. Kragic. A comprehensive grasp taxonomy. In *RSS Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, 2009.

[15] A. R. Fielder and M. J. Moseley. Does stereopsis matter in humans? *Eye*, 10(2):233–238, 1996.

[16] M. A. Goodrich and A. C. Schultz. Human-robot interaction: a survey. *Foundations and trends in human-computer interaction*, 1(3):203–275, 2007.

[17] R. P. Harrison. Nonverbal communication. *Human Communication As a Field of Study: Selected Contemporary Views*, 113, 1989.

[18] D.-A. Huang, W.-C. Ma, M. Ma, and K. M. Kitani. How do we use our hands? discovering a diverse set of common grasps. In *CVPR*, 2015.

[19] Intel. Perceptual computing sdk, 2013.

[20] Y. Jang, S. Noh, H. J. Chang, T. Kim, and W. Woo. 3d finger CAPE: clicking action and position estimation under self-occlusions in egocentric viewpoint. *IEEE Trans. Vis. Comput. Graph.*, 21(4):501–510, 2015.

[21] S. Khamis, T. J., S. J., K. C., I. S., and A. Fitzgibbon. Learning an efficient model of hand shape variation from depth images. In *CVPR*, 2015.

[22] M. Kölsch. An appearance-based prior for hand tracking. In *ACIVS (2)*, pages 292–303, 2010.

[23] M. Kölsch and M. Turk. Hand tracking with flocks of features. In *CVPR (2)*, page 1187, 2005.

[24] P. G. Kry and D. K. Pai. Interaction capture and synthesis. In *ACM Transactions on Graphics (TOG)*, volume 25, pages 872–880. ACM, 2006.

[25] T. Kurata, T. Kato, M. Kourogi, K. Jung, and K. Endo. A functionally-distributed hand tracking method for wearable visual interfaces and its applications. In *MVA*, 2002.

[26] N. Kyriazis and A. A. Argyros. Physically plausible 3d scene tracking: The single actor hypothesis. In *CVPR*, 2013.

[27] N. Kyriazis and A. A. Argyros. Scalable 3d tracking of multiple interacting objects. In *CVPR*, 2014.

[28] J. Liu, F. Feng, Y. C. Nakamura, and N. S. Pollard. A taxonomy of everyday grasps in action. In *14th IEEE-RAS International Conf. on Humanoid Robots, Humanoids 2014*.

[29] S. Mann, J. Huang, R. Janzen, R. Lo, V. Rampersad, A. Chen, and T. Doha. Blind navigation with a wearable range camera and vibrotactile helmet. In *ACM International Conf. on Multimedia*, MM '11, 2011.

[30] T. Moller. A fast triangle-triangle intersection test. *Journal of Graphics Tools*, 2:25–30, 1997.

[31] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Tracking the Articulated Motion of Two Strongly Interacting Hands. In *CVPR*, 2012.

[32] T.-H. Pham, A. Kheddar, A. Qammaz, and A. A. Argyros. Towards force sensing from vision: Observing hand-object interactions to infer manipulation forces. In *CVPR*, 2015.

[33] H. Pirsiavash and D. Ramanan. Detecting activities of daily living in first-person camera views. In *CVPR*, 2012.

[34] X. Ren, C. C. Fowlkes, and J. Malik. Figure/ground assignment in natural images. In *Computer Vision–ECCV 2006*, pages 614–627. Springer, 2006.

[35] G. Rogez, J. S. S. III, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR*, 2015.

[36] G. Rogez, M. Khademi, J. Supancic, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric RGB-D images. In *ECCV Workshop on Consumer Depth Camera For Computer Vision*, 2014.

[37] J. Romero, T. Feix, H. Kjellstrom, and D. Kragic. Spatio-temporal modeling of grasping actions. In *IROS*, 2010.

[38] J. Romero, H. Kjellström, C. H. Ek, and D. Kragic. Non-parametric hand pose estimation with object context. *Image Vision Comput.*, 31(8):555–564, 2013.

[39] J. Romero, H. Kjellstrom, and D. Kragic. Hands in action: real-time 3D reconstruction of hands in interaction with objects. In *ICRA*, pages 458–463.

[40] A. Saxena, J. Driemeyer, J. Kearns, and A. Y. Ng. Robotic grasping of novel objects. In *NIPS*, 2006.

[41] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[42] B. Stenger, A. Thayananthan, P. Torr, and R. Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *PAMI*, 28(9):1372–1384,, 2006.

[43] J. Supancic, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: data, methods, and challenges. In *ICCV*, 2015.

[44] R. Wang and J. Popovic. Real-time hand-tracking with a color glove. *ACM Trans on Graphics*, 28(3), 2009.