

3D Hand Pose Detection in Egocentric RGB-D Images

Grégory Rogez · J. S. Supančič III · Maryam Khademi ·
J. M. M. Montiel · Deva Ramanan

arXiv:1412.0065v1 [cs.CV] 29 Nov 2014

Received: date / Accepted: date

Abstract We focus on the task of everyday hand pose estimation from egocentric viewpoints. For this task, we show that depth sensors are particularly informative for extracting near-field interactions of the camera wearer with his/her environment. Despite the recent advances in full-body pose estimation using Kinect-like sensors, reliable monocular hand pose estimation in RGB-D images is still an unsolved problem. The problem is considerably exacerbated when analyzing hands performing daily activities from a first-person viewpoint, due to severe occlusions arising from object manipulations and a limited field-of-view. Our system addresses these difficulties by exploiting strong priors over viewpoint and pose in a discriminative tracking-by-detection framework. Our priors are operationalized through a photorealistic synthetic model of egocentric scenes, which is used to generate training data for learning depth-based pose classifiers. We evaluate our approach on an annotated dataset of real egocentric object manipulation scenes and compare to both commercial and academic approaches. Our method provides state-of-the-art performance for both hand detection and pose estimation in egocentric RGB-D images.

Keywords egocentric vision, hand pose, object manipulation, RGB-D sensor

This work was supported by EU grant Egovision4Health.

Grégory Rogez, J. M. M. Montiel
Aragon Institute of Engineering Research (i3A), Universidad de Zaragoza, Spain
E-mail: {groguez,josemari}@unizar.es

Grégory Rogez, J. S. Supančič III, Maryam Khademi, Deva Ramanan
Dept. of Computer Science, University of California, Irvine, USA
E-mail: {groguez,jsupanci,mkhademi,dramanan}@unizar.es

1 Introduction

Much recent work has explored various applications of egocentric RGB cameras, spurred on in part by the availability of low-cost mobile sensors such as Google Glass, Microsoft SenseCam, and the GoPro camera. Many of these applications, such as life-logging [14], medical rehabilitation [59], and augmented reality [3], require inferring the interactions of the first-person observer with his/her environment while recognizing his/her activities. Whereas third-person-view activity analysis is often driven by human full-body pose, egocentric activities are often defined by hand pose and the objects that the camera wearer interacts with. Towards that end, we specifically focus on the tasks of hand detection and hand pose estimation from egocentric viewpoints of daily activities. We show that depth-based cues, extracted from an *egocentric depth camera*, provides an extraordinarily helpful cue for egocentric hand-pose estimation.

One may hope that depth simply “solves” the problem, based on successful systems for real-time human pose estimation based on Kinect sensor [47] and prior work on articulated hand pose estimation for RGB-D sensors [30, 55, 17, 36, 55]. Recent approaches have also tried to exploit the 2.5D data from Kinect-like devices to understand complex scenarios such as object manipulation [21] or two interacting hands [32]. We show that various assumptions about visibility/occlusion and manual tracker initialization may not hold in an egocentric setting, making the problem still quite challenging.

Challenges: Three primary challenges arise for hand pose estimation in everyday egocentric views, compared to 3rd person views. First, *tracking is less reliable*. Even assuming that manual initialization is possible, a limited field-of-view from an egocentric viewpoint causes hands to frequently move outside the camera view frustum. This makes it dif-



Fig. 1 Challenges. We contrast third person depth (a) and RGB images (b) (overlaid with pose estimates from [36]) with depth image and RGB images (c,d) from egocentric views of daily activities. **Hands leaving the field-of-view**, **self-occlusions**, **occlusions due to objects** and **malsegmentability due to interactions with the environment** are common hard cases in egocentric settings.

ficult to apply tracking models that rely on accurate estimates from previous frames, since the hand may not even be visible. Second, *active hands are difficult to segment*. Many previous systems for 3rd-person views make use of simple depth-based heuristics to both detect and segment the hand. These are difficult to apply during frames where users interact with objects and surfaces in their environment. Finally, *fingers are often occluded* by the hand (and other objects being manipulated) in egocentric views, considerably complicating articulated pose estimation. See examples in Fig. 1.

Our approach: We describe a successful approach to hand-pose estimation that makes use of the following key observations. First, **depth cues** provide an extraordinarily helpful signal for pose estimation in the near-field, first-person viewpoints. Though this observation may seem obvious, state-of-the-art methods for egocentric hand detection do not make use of depth [25, 24]. Moreover, in our scenario, depth cues are not “cheating” as humans themselves make use of stereopsis for near-field analysis [44, 11]. Second, the egocentric setting provides strong priors over **viewpoint**, **grasps**, and **interacting-objects**. We operationalize these priors by generating synthetic training data with a rendered 3D hand model. In contrast to previous work that uses a “floating hand”, we mount a synthetic egocentric camera to a virtual full-body character interacting with a library of everyday objects. This allows us to make use of contextual cues for both data generation and recognition (see Fig. 3). Third, we treat

pose estimation (and detection) as a discriminative multi-class classification problem. To efficiently evaluate a large number of pose-specific classifiers, we make use of **hierarchical cascade** architectures. Unlike much past work, we classify global poses rather than local parts, which allows us to better reason about self-occlusions. Our classifiers process single frames, using a tracking-by-detection framework that avoids the need for manual initialization (see Fig. 2c-e).

Evaluation: Unlike human pose estimation, there exists no standard benchmarks for hand pose estimation, especially in egocentric videos. We believe that quantifiable performance is important for many broader applications such as health-care rehabilitation, for example. Thus, for the evaluation of our approach, we have collected and annotated (full 3D hand poses) our own benchmark dataset of real egocentric object manipulation scenes, which we will release to spur further research. It is surprisingly difficult to collect annotated datasets of hands performing real-world interactions; indeed, many prior work on hand pose estimation evaluate results on synthetically-generated data. We developed a semi-automatic labelling tool which allows to accurately annotate partially occluded hands and fingers in 3D, given real-world RGB-D data. We compare to both commercial and academic approaches to hand pose estimation, and demonstrate that our method provides state-of-the-art performance for both hand detection and pose estimation in egocentric RGB-D images.

Overview: This work is an extension of [39]. This manuscript explains our approach with considerably more detail, reviews a broader collection of related work, and provide new and extensive comparisons with state-of-the-art methods. We review related work in Sec. 2 and present our approach in Sec. 3, focusing on our synthetic training data generation procedure (Sec. 3.1) and our hierarchical multi-class architecture (Sec. 3.2). We conclude with experimental results in Sec. 4.

2 Related work

Egocentric hands: Previous work examined the problem of recognizing objects [10, 34] and interpreting American Sign Language poses [51] from wearable cameras. Much work has also focused on hand detection [25, 24], hand tracking [20, 19, 18, 29], finger tracking [7], and hand-eye tracking [43] from wearable cameras. Often, hand pose estimation is examined during active object manipulations [28, 38, 37, 9]. Most such previous work makes use of RGB sensors. Our approach demonstrates that an egocentric depth camera makes things considerably easier.

Egocentric depth: Depth-based wearable cameras are attractive because depth cues can be used to better reason about occlusions arising from egocentric viewpoints. There

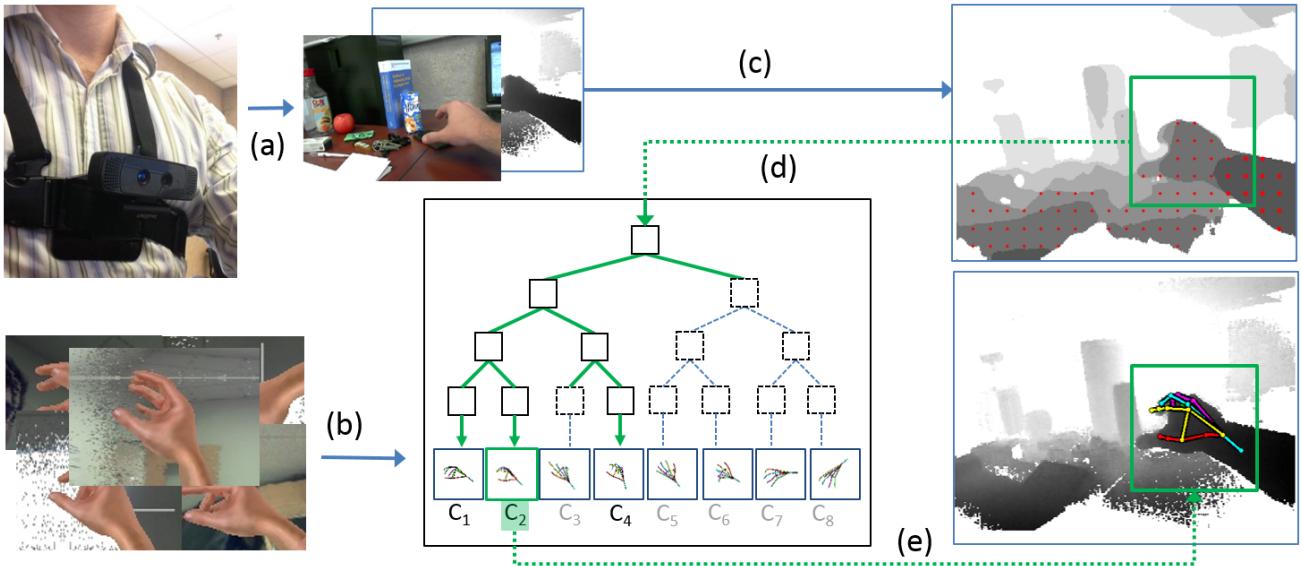


Fig. 2 System overview. (a) Chest-mounted RGB-D camera. (b) Synthetic egocentric hand exemplars are used to define a set of hand pose classes and train a multi-class hand classifier. The depth map is processed to select a sparse set of image locations (c) which are classified obtaining a list of probable hand poses (d). Our system produces a final estimate by reporting one or more top-scoring pose classes (e).

has been surprisingly little prior work in this vein, with notable exceptions focusing on targeted applications such as navigation for the blind [27] and recent work on egocentric object understanding [5,26]. We posit that one limitation may be the need for small form-factors for wearable technology, while structured light sensors such as the Kinect often make use of large baselines. We show that time-of-flight depth cameras are an attractive alternative for wearable depth-sensing, since they do not require large baselines and so require smaller form-factors.

Depth-based pose: Our approach is closely inspired by the Kinect system and its variants [47], which makes use of synthetically generated depth maps for articulated pose estimation. Notably, Kinect follows in the tradition of local part models [4,8], which are attractive in that they require less training data to model a large collection of target poses. However, it is unclear if local methods can deal with large occlusions (such as those encountered during egocentric object manipulations) where local information can be ambiguous. Our approach differs in that our classifiers classify global poses rather than local parts. Finally, much previous work assumes that hands are easily segmented or detected. Such assumptions simply do not hold for everyday egocentric interactions.

Interacting objects: Estimating the pose of a hand manipulating an object is challenging [13] due to occlusions and ambiguities in segmenting the object versus the hand. It is attractive to exploit contextual cues through simultaneously tracking hands [21,22] and the object. [2] use multi-cameras to reduce the number of full occlusions. We jointly

model hand and objects with synthetic hand-object exemplars as in [42]. However, instead of modeling floating hands, we model them in a realistic egocentric context that is constrained by the full human body.

Tracking vs detection: Temporal reasoning is also particularly attractive because one can use dynamics to resolve ambiguities arising from self and object occlusions. Much prior work on hand-pose estimation takes this route [30,55,17,36]. Our approach differs in that we focus on single-image hand pose estimation, which is required to avoid manual (re)initialization. Exceptions include [58,54,56], who also process single images but focus on third-person views.

Generative vs discriminative: Generative model-based approaches have historically been more popular for hand pose estimation [52]. A detailed 3D model of the hand pose is usually employed for articulated pose tracking [31,32] and detailed 3D pose estimation [23]. Discriminative approaches [55,17] for hand pose estimation tend to require large datasets of training examples, synthetic, realistic or combined [55]. Learning formalisms include boosted classifier trees [33] and randomized decision forests [17], and regression forests [55,54]. Sridhar et al. [50] propose a hybrid approach that combines discriminative part-based pose retrieval with a generative model-based tracker. Our approach uses a computer graphics model to generate training data, which is then used to learn discriminative pose-specific classifiers.

Hierarchical cascades: We approach pose estimation as a hierarchical multi-class classification task, a strategy that dates back at least to Gavrila *et al* [12]. Our framework fol-



Fig. 3 Training data. We show on the left hand side our avatar that we mount with a virtual egocentric camera. In the middle we show the EveryDayHands animation library [6] used to generate realistic hand-object configurations. On the right, we present some examples of resulting training images rendered using Poser.

lows a line of work that focuses on efficient implementation through coarse-to-fine hierarchical cascades [61, 53, 4, 40]. Our work differs in its discriminative training and large-scale ensemble averaging over an exponentially-large set of cascades, both of which considerably improve accuracy and speed.

3 Our method

Our method works by using a computer graphics model to generate synthetic training data. We then use this data to train a classifier for pose estimation. We describe each stage in turn.

3.1 Synthesizing training data

We represent a hand pose as a vector of joint angles of a kinematic skeleton θ . We use a hand-specific forward kinematic model to generate a 3D hand mesh given a particular θ . In addition to hand pose parameters θ , we also need to specify a camera vector ϕ that specifies both a viewpoint and position. We experimented with various priors and various rendering packages.

Floating hands vs full-body characters: Much work on hand pose estimation makes use of an isolated “floating” hand mesh model to generate synthetic training data. Popular software packages include the open-source `libhand` [57] and commercial Poser [48, 46]. We posit that modeling a full character body, and specifically, the full arm, will provide important contextual cues for hand pose estimation. To generate egocentric data, we mount a synthetic camera on the chest of a virtual full-body character, naturally mimicking our physical data collection process. To generate data corresponding to different body and hand shapes, we make use of Poser’s character library.

Pose prior: Our hand model consists of 26 joint angles, $\theta \in [0, 2\pi]^{26}$. It is difficult to specify priors over such high-dimensional spaces. We take a non-parametric data-driven approach. We first obtain a training set of joint angles $\{\theta_n\}$ from a collection of grasping motion capture data [41]. We then augment this core set of poses with synthetic perturbations, making use of rejection sampling to remove invalid poses. Specifically, we first generate proposals by perturbing the i^{th} joint angle of training sample n with Gaussian noise

$$\theta_n[i] \rightarrow \theta_n[i] + \epsilon \quad \text{where} \quad \epsilon \sim N(0, \sigma_i) \quad (1)$$

The noise variance σ_i is obtained by manual tuning on validation data. Notably, we also perturb the *entire arm* of the full character-body, which generates natural (egocentric) viewpoint variations of hand configurations. Note that we consider smaller perturbations for fingers to keep grasping poses reasonable. We remove those samples that result in poses that are self-intersecting or lie outside the field-of-view.

Viewpoint prior: The above pose perturbation procedure for θ naturally generates realistic egocentric camera viewpoints ϕ for our full character models. We also performed some diagnostic experiments with a floating hand model. To specify a viewpoint prior in such cases, we limited the azimuth ϕ_{az} to lie between 180 ± 30 (corresponding to rear viewpoints), elevation ϕ_{el} to lie between -30 and 10 (since hands tend to lie below the chest mount), and bank ϕ_b to lie between ± 30 . We obtained these ranges by looking at a variety of collected data (not used for testing).

Interacting objects: We wish to explore egocentric hand pose estimation in the context of natural, functional hand movement. This often involves interactions with the surrounding environment and manipulations of nearby objects. We posit that generating such contextual training data will be important for good test-time accuracy. However, modeling the space of hand grasps and the world of manipulable objects is itself a formidable challenge. We make use of the

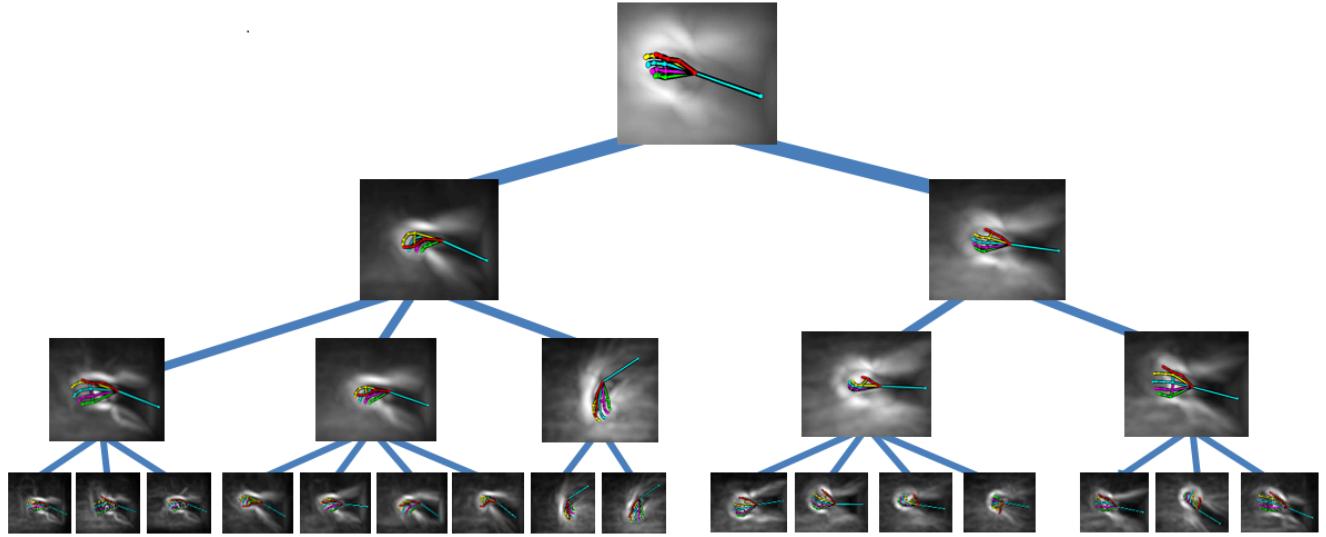


Fig. 4 Hierarchy of hand poses. We visualize a hierarchical graph $G = (V, E)$ of quantized poses with $K = 16$ leaves. The i^{th} node in this tree represents a coarse pose class, visualized with the average hand pose and the average gradient map over all the exemplars in that coarse pose.

EveryDayHands animation library [6], which contains 40 canonical hand grasps. This package was originally designed as a computer animation tool, but we find the library to cover a reasonable taxonomy of grasps for egocentric recognition. A surprising empirical fact is that humans tend to use a small number of grasps for everyday activities - by some counts, 9 grasps are enough to account for 80% of human interactions [60]. Following this observation, we manually amassed a collection of everyday common objects from model repositories [1]. Our objects include spheres and cylinders (of varying sizes), utensils, phones, cups, etc. We paired each object a viable grasp (determined through visual inspection), yielding a final set of 52 hand-object combinations. We apply our rejection-sampling technique to generate a large number of grasp-pose and viewpoint perturbations, yielding a final dataset of 10,000 synthetic egocentric hand-object examples. Some examples are shown in Fig. 3.

3.2 Hierarchical cascades

We use our training set to learn a model that simultaneously detects hands and estimates their pose. Both tasks are addressed with a scanning window classifier, that outputs one of K discrete pose classes or a background label. One may need a large K to model lots of poses, increasing training/testing times and memory footprints. We address such difficulties through coarse-to-fine sharing and scanning-window cascades. Such architectures have been previously explored in [61, 61, 4, 40]. We contribute methods for efficient discriminative training, efficient run-time evaluation, and ensemble averaging. To describe our contributions, we first re-

cast previous approaches in a mathematical framework that is amenable to our proposed modifications.

Hierarchical quantization: Firstly, we represent each training example as a depth image x and a label vector y of joint positions in a canonical coordinate frame with normalized position and scale. We quantize this space of poses $\{y\}$ into K discrete values with K -mean clustering. We then agglomeratively merge these quantized poses into a hierarchical tree $G = (V, E)$ with K -leaves, following the procedure of [40]. Each node $i \in V$ represents a coarse pose class. We visualize the tree in Fig. 4.

Coarse-to-fine sharing: Given a test image x , a binary classifier tuned for coarse pose-class $i \in V$ is evaluated as:

$$f_i(x) = \prod_{j \in A_i} h_j(x) \quad \text{where} \quad h_j(x) = \mathbf{1}_{[w_j^T x > 0]} \quad (2)$$

where $\mathbf{1}$ is the indicator function which evaluates to 1 or 0. Here, f_i is a “strong” classifier for class i obtained by ANDing together “weak” binary predictions h_j from a set $j \in A_i$. A_i is the set of “ancestor” nodes encountered on the path from i to the root of the tree (i , its parents, grandparents, etc.). Each weak classifier h_j is a thresholded linear function that is defined on appearance features extracted from window x . We use HOG appearance features extracted from subregions within the window, meaning that w_j can be interpreted as a zero-padded “part” template tuned for coarse-pose j . Parts higher in the tree tend to be generic and capture appearance features common to many pose classes. Parts lower in the tree, toward the leaves, tend to capture pose-specific details.

Breadth-first search (BFS): The prediction for pose-class i will be 1 if and only if all classifiers in the ancestor

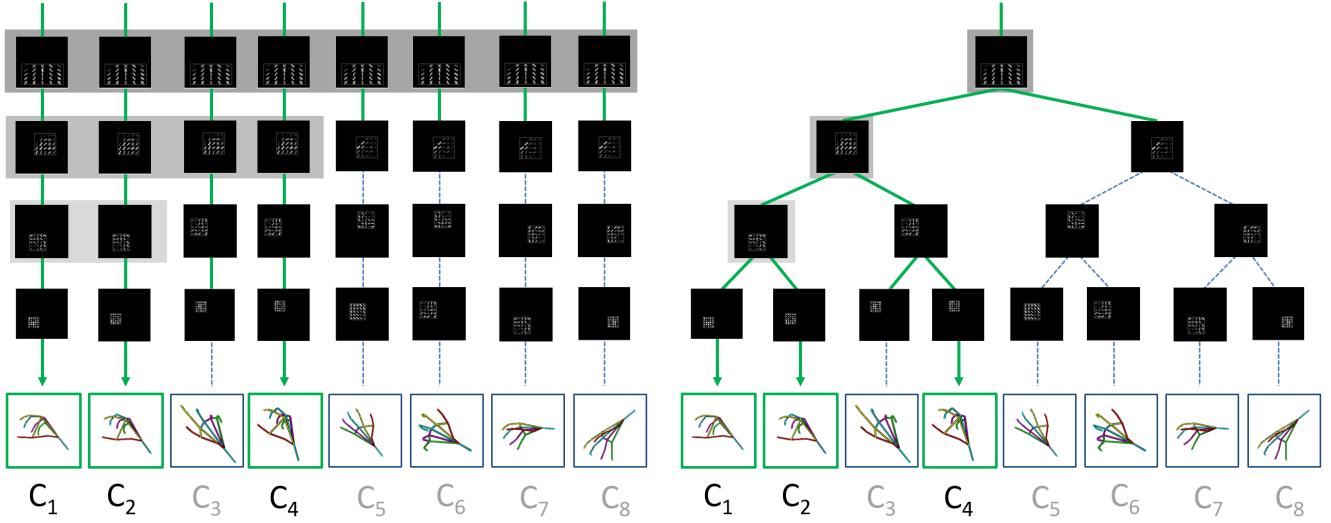


Fig. 5 Hierarchical cascades. We approach pose-estimation as a K -way classification problem. We define a *linear-chain* cascade of rejectors for each of K pose class (*left*). By sharing “weak” classifiers across these cascades, we can efficiently organize the collection into a coarse-to-fine tree $G = (V, E)$ with K leaves (*right*). Weak classifiers near the root of tree are tuned to fire on a large collection of pose classes, while those near the leaves are specific to particular pose classes. Each linear cascade can be recovered from the tree by enumerating the ancestors of a leaf node.

set predict 1. If node i fails to fire, its children (and their descendants) cannot be detected and so can be immediately rejected. We can efficiently prune away large portions of the pose-space when evaluating region x with a truncated breadth-first search (BFS) through graph G (see Alg. 1), making scanning-window evaluation at test-time quite efficient. Though a queue-based BFS is a natural implementation, we have not seen this explicitly described in previous work on hierarchical cascades [61, 61, 4, 40]. As we will show, such an “algorithmic” perspective immediately suggests straightforward improvements for training and modeling averaging.

```

input : Image window  $x$ , classifiers  $\{w_i, i \in V\}$ 
output:  $vote(i)$  for each leaf class  $i$ .
1 create a queue Q;
2 enqueue 1 onto Q;
3 while  $Q \neq empty$  do
4   | i=Q.dequeue();
5   | if  $w_i^T x > 0$  then
6   |   | for  $\forall k \in child(i)$  do
7   |   |   | enqueue k onto Q;
8   |   | end
9   |   | if  $k = \emptyset$  then
10  |   |   | vote(i)=1;
11  |   | end
12  | end
13 end
```

Algorithm 1: Classification with a single cascade. We perform a truncated breadth-first search (BFS) using a first-in, first-out queue. We insert the children of the node i into the queue only if the weak classifier w_i successfully fires.

Multi-class detection: When Eq (2) is evaluated on leaf classes, it is mathematically equivalent to a collection of K *linear chain* cascades tuned for particular poses. These linear chains are visualized in Fig. 5. From this perspective, multiple poses (or leaf classes) may fire on a single window x . This is in contrast to other hierarchical classifiers such as decision trees, where only a single leaf can be reached. We generally report the highest-scoring pose as the final result, but alternate high-scoring hypotheses may still be useful (since they can be later refined using say, a tracker). We define the score for a particular pose class by aggregating binary predictions over a large *ensemble* of cascades, as described below.

Ensembles of cascades: To increase robustness, we aggregate predictions across an ensemble of classifiers. [40] describes an approach that makes use of a pool of weak part classifiers at node i :

$$h_i(x) \in H_i \quad \text{where} \quad |H_i| = M \quad (3)$$

One can instantiate a tree by selecting a weak classifier (from its candidate pool H_i) for each node i in graph $G = (V, E)$. This defines a potentially exponential-large set of instantiations $M^{|V|}$, where M is the size of each candidate pool. In practice, [40] found that averaging predictions from a small random subset of trees significantly improved results.

3.3 Joint training of exponential ensembles

In this section, we present several improvements that apply in our problem domain. Because of local ambiguities due to self-occlusions, we expect individual part templates

to be rather weak. This in turn may cause premature cascade rejections. We describe modifications that reduce premature rejections through joint training of weak classifiers and exponentially-large ensembles of cascades.

Sequential training: Much previous work assumes pose-specific classifiers are given [53, 4] or independently learned [40]. For example, [40] trains weak classifiers w_i by treating all *all* training examples from pose-class/node i as positives, and examples from *all* other poses as negatives. Instead, we use only the training examples that pass through the rejection cascade up to node i . This better reflects the scenario at test-time. This requires classifiers to be trained in a sequential fashion, in a similar coarse-to-fine BFS over nodes from the root to the leaves (Alg. 2).

```

input : Training data  $(x_n, y_n)$  and tree  $G = (V, E)$ .
output: Weak part classifiers  $\{w_i, i \in V\}$ .
1 create a queue Q;
2 enqueue  $(1, \{x_n\}, \{y_n\})$  onto Q;
3 while  $Q \neq \text{empty}$  do
4    $(i, x, y) = Q.\text{dequeue}()$ ;
5    $w_i = \text{Train}(x, y \in \text{class}_i)$ ;
6    $(x, y) := \{(x_n, y_n) : w_i^T x_n > 0\}$ ;
7   for  $\forall k \in \text{child}(i)$  do
8     | enqueue  $(k, x, y)$  onto Q;
9   end
10 end
```

Algorithm 2: Cascade sequential training. We perform a BFS through pose classes, training weak classifiers w_i by enqueueing node indices i and the training data (x, y) that reaches that node. *Train* returns a (linear SVM) model given training examples with binary labels. With a slight abuse of notation, $y \in \text{class}_i$ denotes a set of binary indicators that specify which examples belong to class_i , where class_i is the set of leaf classes reachable through a BFS from node i .

Exponentially-large ensembles: Rogez et al. [40] average votes across a small number (around hundred) of explicitly constructed trees. By averaging over a larger set, we reduce the chance of a premature cascade rejection. We describe a simple procedure for exactly computing the average over the exponentially-large set of $M^{|V|}$ in Alg. 3. Our insight is that one can compute an *implicit* summation (of votes) over the set by caching partial summations t during the BFS. We refer the reader to the algorithm and caption for a detailed description. To train the pool of weak classifiers, we can leverage our sequential training procedure. Simply replace Lines 5 and 6 of Alg. 2 with the following:

$$\{w_{ij} : j = 1 \dots M\} = \text{TrainEnsemble}(x, y \in \text{class}_i) \quad (4)$$

$$(x, y) := \{(x_n, y_n) : \sum_j w_{ij}^T x_n > 0\} \quad (5)$$

where *TrainEnsemble* is a learning algorithm that returns an ensemble of M models by randomly selecting subsets of training data or subsets of features. We select random subsets of features corresponding to local regions from window x . This allows the returned models w_{ij} to be visualized as "parts" (Fig. 5).

input : Image window x , weak classifier pools $\{H_i, i \in V\}$
output: $\text{vote}(i)$ for each leaf class i .

```

1 create a queue Q;
2 enqueue  $(1, 1)$  onto Q;
3 while  $Q \neq \text{empty}$  do
4    $(i, t) = Q.\text{dequeue}()$ ;
5    $t := t + \sum_{h_i \in H_i} h_i(x)$ ;
6   if  $t > 0$  then
7     for  $\forall k \in \text{child}(i)$  do
8       | enqueue  $(k, t)$  onto Q;
9     end
10    if  $k = \emptyset$  then
11      |  $\text{vote}(i) = t$ ;
12    end
13  end
14 end
```

Algorithm 3: Classification with exponentially large number of cascades. When processing a node i , all its associated weak classifiers $h_i \in H_i$ are evaluated. We keep track of the (exponentially-large) number of successful ensemble components by enqueueing a running estimate t and node index i . Once the queue is empty, $\text{vote}[i]$ is populated with the number of ensemble components that fired on leaf class i (which is upper bounded by $M^{|V|}$).

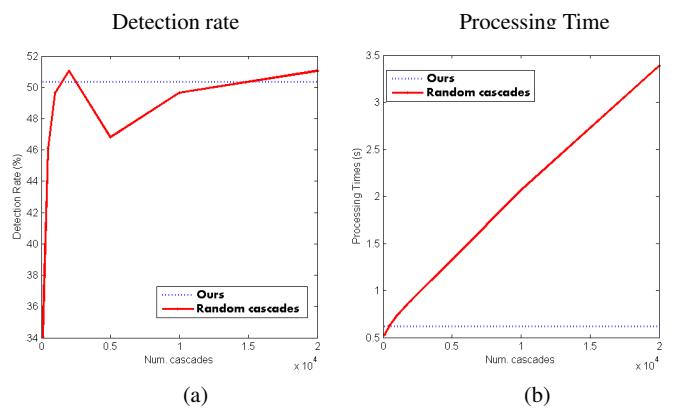


Fig. 6 Comparison with random cascades. We show in (a) that our new detector is equivalent to an exponentially-large number of Random Cascades (RC) from [40]. In (b), we show that the RC computational cost increases linearly with the number of cascades and that, when considering a very large number of cascades, our model is more efficient.



Fig. 7 Test data. We show several examples of training RGB-D images captured with the chest-mounted Intel Creative camera from Fig. 2a. Real egocentric object manipulation scenes have been collected and annotated (full 3D hand poses) for evaluation.

3.4 Implementation issues

Sparse search: We leverage two additional assumptions to speed up our scanning-window cascades: 1) hands must lie in a valid range of depths, i.e., hands can not appear further away from the chest-mounted camera than physically possible and 2) hands tend to be of a canonical size s . These assumptions allow for a much sparser search compared to a classic scanning window, as only “valid windows” need be classified. A median filter is first applied to the depth map $d(x, y)$. Locations greater than arms length (75 cm) away are then pruned. Assuming a standard pinhole camera with focal length f , the expected image height of a hand at valid location (x, y) is given by $S_{map}(x, y) = \frac{s}{f}d(x, y)$. We apply our hierarchical cascades to valid positions on a search grid (16-pixel strides in x-y direction) and quantized scales given by S_{map} , visualized as red dots in Fig. 2c.

Features: We experiment with two additional sets of features x . Following much past work, we make use of HOG descriptors computed from RGB signals. We also evaluated oriented gradient histograms on depth images (HOG-D). While not as common, such a gradient-based depth descriptor can be shown to capture histograms of normal directions (since normals can be computed from the cross product of depth gradients) [49]. For depth, we use 5x5 HOG blocks and 16 signed orientation bins.

4 Experiments

Depth sensor: Much recent work on depth-processing has been driven by the consumer-grade PrimeSense sensor [45], which is based on structured light technology. At its core, this approach relies on two-view stereopsis (where correspondence estimation is made easier by active illumination). This may require large baselines between two views, which is undesirable for our egocentric application for two reasons; first, this requires larger form-factors, making the camera less mobile. Second, this produces occlusions for points in

the scene that are not visible in both views. Time-of-flight depth sensing, while less popular, is based on a pulsed light emitter that can be placed arbitrarily close to the main camera, as no baseline is required. This produces smaller form factors and reduces occlusions in that camera view. Specifically, we make use of the consumer-grade TOF sensor from Creative [15] (see Fig. 2a).

Dataset: We have collected and annotated (full 3D hand poses) our own benchmark dataset of real egocentric object manipulation scenes, which we will release to spur further research¹. We developed a semi-automatic labelling tool which allows to accurately annotate partially occluded hands and fingers in 3D. A few 2D joints are first manually labelled in the image and used to select the closest synthetic exemplars in the training set. A full hand pose is then created combining the manual labelling and the selected 3D exemplar. This pose is manually refined, leading to the selection of a new exemplar, and the creation of a new pose. This iterative process is followed until an acceptable labelling is achieved. We captured 4 sequences of 1000 frames each, which were annotated every 10 frames in both RGB and Depth. We use 2 different subjects (male/female) and 4 different indoor scenes. Some examples are presented in Fig. 7.

Parameters: We train a cascade model trained with $K = 100$ classes, a hierarchy of 6 levels and $M = 3$ weak classifiers per node. We synthesize 100 training images per class. We experimented with larger numbers of classes (up to $K = 1000$), but did not observe significant improvement. We suspect this is due to the restricted set of viewpoints and grasp poses present in egocentric interactions (which we see as a contribution of our work). As a point of contrast, we trained a non-egocentric hand baseline in Fig. 8 that operated best at $K = 800$ classes.

¹ Please visit www.gregrogez.net/

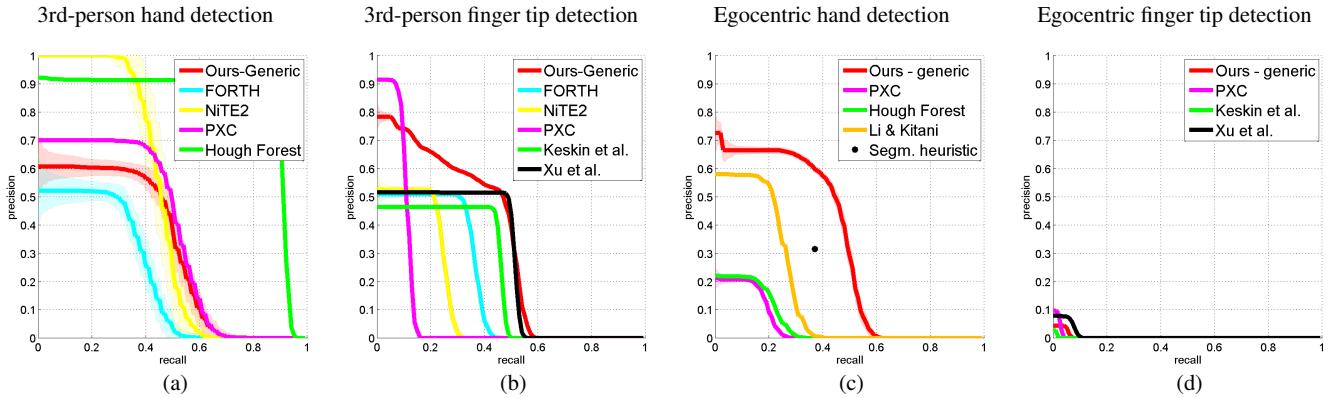


Fig. 8 Third-person vs 1st person . Numerical results for 3rd-person (a-b) and egocentric (c-d) sequences. We compare our method (tuned for generic priors) to state-of-the-art techniques from industry (NITE2 [35] and PXC [15]) and academia (FORTH [30], Keskin et al [16] and Xu et al. [58]) in terms of (a) hand detection and (b) finger tips detection. We refer the reader to the main text for additional description, but emphasize that (1) our method is competitive (or out-performs) prior art for detection and pose estimation and (2) pose estimation is considerably harder in egocentric views.

4.1 Benchmark performance

In this subsection, we validate our proposed architecture on a in-house dataset of 3rd-person hands. Our goal is to compare performance with standard baselines, verifying that our architecture is competitive. In this section, we make use of a generic set of 3rd-person views for training and use $K = 800$ pose classes. We then use this system as a starting point for egocentric analysis, exploring various configurations and priors further in the next section.

Evaluation: We evaluate both hand detections and pose estimation. A candidate *detection* is deemed correct if it sufficiently overlaps the ground-truth bounding-box (in terms of area of intersection over union) by at least 50%. As some baseline systems report the pose of only confident fingers, we measure *finger-tip detection* accuracy as a proxy for pose estimation. To make this comparison fare, we only score visible finger tips and ignore occluded ones.

Baselines: We compare our method to state-of-the-art techniques from industry [35, 15] and academia [30, 58, 16, 24]. Because public code is not available, we re-implemented Xu et al' [58] and Keskin et al' [16], verifying that our performance matched published results in [54]. Xu proposes a three stage pipeline: (1) detect position and in-plane rotation with a Hough forest (2) estimate joint angles with a second stage Hough forest (3) apply an articulated model to validate global consistency of joint estimates. Keskin's model also has three stages: (1) estimate global hand shape from local votes (2) given the estimated shape, apply a shape-specific decision forest to predict a part label for each pixel and (3) apply mean-shift to regress joint positions. Because Keskin's model assumes detection is solved, we experimented with several different first-stage detectors before settling on using Xu's first stage Hough forest, due to its superior per-

formance. Thus, both models share the same hand detector in our evaluation.

Third-person vs egocentric: Following Fig. 8, our hierarchical cascades are competitive for 3rd-person hand detection (a) and state-of-the-art for finger detection (b). When evaluating the same models (without retraining) on egocentric test data (c) and (d), most methods (including ours) perform significantly worse. FORTH and NITE2 trackers catastrophically fail since hands frequently leave the view, and so are omitted from (c) and (d). Random Forest baselines [58, 16] drop in performance, even for hand detection. We posit this drop comes difficulties in segmenting egocentric hands (as shown in Fig. 9). To test this hypothesis, we develop a custom segmentation heuristic that looks for arm pixels near the image border, followed by connected-component segmentation. We also experiment with the (RGB) pixel-level hand detection algorithm from [24]. These segmentation algorithms outperform many baselines for hand-detection, but still underperforms our hierarchical cascade. We conclude that (1) hand pose estimation is considerably harder in the egocentric setting and (2) our (generic) pose estimation system is a state-of-the-art starting point for our subsequent analysis. Finally, we posit that our strong performance (at least for hand detection) arises from our global hand classifiers, while most baselines tend to classify local parts.

4.2 Diagnostic analysis

In this section, we further explore various configurations and priors for our approach, tuned for the egocentric setting.

Evaluation: Since our algorithm always returns a full articulated hand pose, we evaluate pose estimation with 2D-RMS re-projection error of keypoints. This time, we score the 20 keypoints defining the hand, including those which

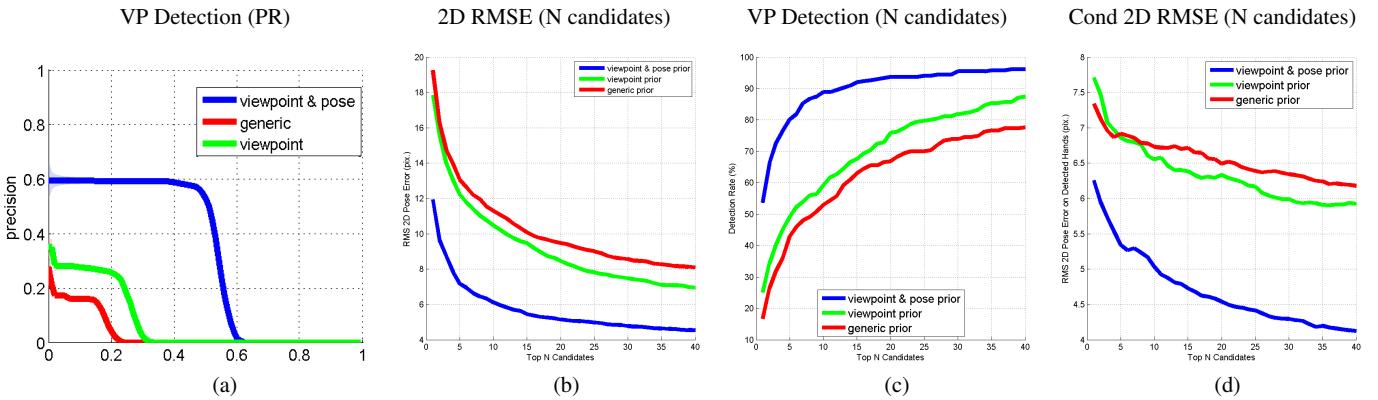


Fig. 10 Quantitative results varying our prior. We evaluate three different priors with respect to (a) viewpoint-consistent hand detection (precision-recall curve), (b) 2D RMS error, (c) viewpoint-consistent detections and (d) 2D RMS error conditioned on viewpoint-consistent detections. Please see text for detailed description of our evaluation criteria and analysis. In general, egocentric-pose priors considerably improve performance, validating our egocentric-synthesis engine from Sec. 3.1. When tuned for $N = 10$ candidates per image, our system produces pose hypotheses that appear accurate enough to initialize a tracker.

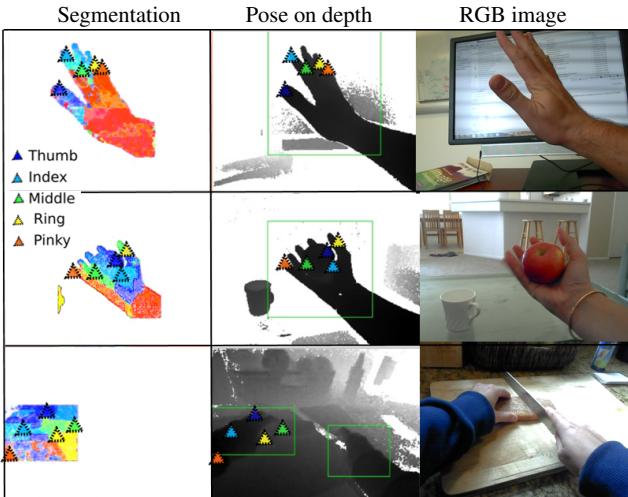


Fig. 9 Qualitative results obtained by state-of-the-art [16]: (top) Shows a test sample which is akin to those addressed by prior work. The hand is easily segmentable from the background and no object interaction is present. In this case, the method from [16] correctly identifies 4 of the fingers and has a small localization error for the 5th. (middle) Here the hand is holding a ball. Now, only the pinky is correctly localized, despite the algorithm being provided with object interaction in the training data. This demonstrates the combinatorial increase in difficulty posed by introducing objects. (bottom) Finally, when we are able to correctly detect the hand, but cannot easily segment it from the background, methods based on per-pixel classification fail because they produce strong “garbage” classifications for background.

are occluded. We believe this is important, because numerous occlusions arise from egocentric viewpoints and object manipulation tasks. This evaluation criteria will give a better sense of how well our method actually recognizes global hand poses, even in case of partially occluded hands. For additional diagnosis, we categorize errors into detection failures, correct detections but incorrect viewpoint, and correct detection and viewpoint but incorrect articulated pose. Specif-

ically, *viewpoint-consistent detections* are detections for which the RMS error of all 2D joint positions falls below a coarse threshold (10 pixels). *Conditional 2D RMS error* is the re-projection error for well-detected (viewpoint-consistent) hands. Finally, we also plot accuracy as a function of the number of *N candidate detections* per image. With enough hypotheses, accuracy must max out at 100%, but we demonstrate that good accuracy is often achievable with a small number of candidates (which may later be re-ranked, by say, a tracker).

Pose+viewpoint prior: We explore 3 different priors: 1) a *generic prior* obtained using a floating “libhand” hand with all possible random camera viewpoints and pose configurations, 2) a *viewpoint prior* obtained limiting a floating hand to valid egocentric viewpoints and 3) a *viewpoint & pose prior* obtained using our full synthesis engine described in Sec. 3.1, i.e. using a virtual egocentric camera mounted on a full body avatar manipulating objects. Note that we respectively consider 800, 140 and 100 classes to train these models. In Fig. 10, we show that, in general, a viewpoint prior produces a marginal improvement, while our full egocentric-specific pose and viewpoint prior considerably improves accuracy in all cases. This suggests that our synthesis algorithm correctly operationalizes egocentric viewpoint and pose priors, which in turn leads us to make better hypothesis for daily activities/grasp poses. With a modest number of candidates ($N = 10$), our final system produces viewpoint-consistent detections in 90% of the test frames with an average 2D RMS error of 5 pixels. From a qualitative perspective, this performance appears accurate enough to initialize a tracker.

Ablative analysis: To further analyze our system, we perform an ablative analysis that turns “off” different aspects of our system: sequential training, ensemble of cascades, depth feature, sparse search and additional object prior. Hand detection and conditional 2D hand RMS error are given

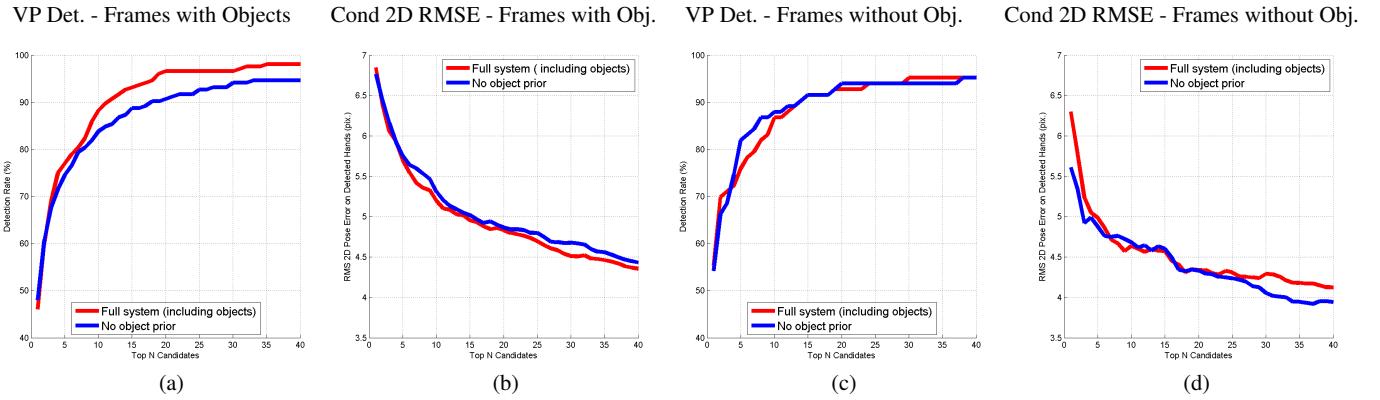


Fig. 12 Object prior. Effect of modeling a hand with and without an object on hand detection (a and c) and hand pose recognition (b and d). Results are given for test frames with (a and b) and without (c and d) object manipulation. Again, we measure viewpoint-consistent detections (a and c) and 2D RMS error conditioned on well-detected hands (b and d). Both hand detection and pose recognition are considerably more challenging for frames with object interactions, likely due to additional occlusions from the manipulated objects. The use of an object prior provides a small but noticeable improvement for frames with objects, but does not affect the performance of the system for the frames without objects.

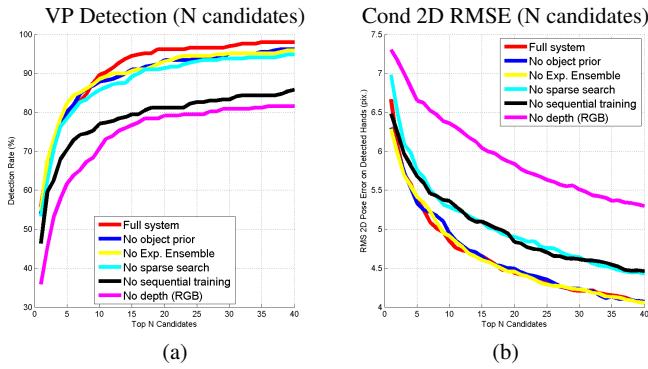


Fig. 11 Ablative analysis. We evaluate performance when turning off particular aspects of our system, considering both (a) viewpoint-consistent detections (b) 2D RMS error conditioned on well-detected hands. When turning off our exponentially large ensemble or synthetic training, we use the default of 100-component ensemble as in [40]. When turning off the depth feature, we use a classifier trained on aligned RGB images. Please see the text for further discussion of these results.

in Fig. 11. Depth HOG features and sequential training of parts are by far the crucial components of our system. Turning these parameters off decreases the detection rate by a substantial amount (between 10 and 30%). Our exponentially-large ensemble of cascades and sparse search marginally improve accuracy but are much more efficient: in average, the exponentially-large ensemble is 2.5 times faster than an explicit search over a 100-element ensemble (as in [40]), while the sparse search is 3.15 times faster than a dense grid. Modeling objects produces better detections, particularly for larger numbers of candidates. In general, we find this additional prior helps more for those test frames with object manipulations as detailed below.

Modeling objects: In Fig. 12, we analyze the effect of object interactions on egocentric hand detection and pose es-

timation when employing an object prior (or not). We plot the accuracy (for both viewpoint-consistent detections and conditional 2D RMS error) on those test frames with (a,b) and without (c,d) object interactions. The corresponding plots computed on the whole dataset (using both types of frames) is already shown in Fig. 11a and Fig. 11b. We see that additional modeling of interacting hands and objects (with an object prior) somewhat improves performance for frames with object manipulation without affecting the performance of the system for the frames without objects.

Number of parts: In Fig. 13, we show the effect of varying the number of parts M at each branch of our cascade model. We analyze both hand detection rate and hand pose average accuracy. These plots clearly validate our choice of using $M = 3$ parts per branch (Fig. 13a). The performance decreases when considering more parts because the classifier is more likely to produce a larger amount of false positives. Additionally, we can see in Fig. 13b that using more than 3 parts does not improve the accuracy in terms of hand pose.

Pixel threshold: In Fig. 14, we show the effect of varying the pixel-overlap threshold for computing correct detection and average pose accuracy. A lower pixel threshold decreases detection rate but increases pose accuracy on detected hands and vice-versa. Our 10-pixel threshold is a good trade-off between these 2 criteria. In (c) and (d), we show the percentage of detection and pose correctness for 2 different thresholds: 10 pixels (c) used throughout the main paper and 5 pixels (d). The measure we used for detection (blue area) is much more strict than a simple bounding boxes overlap criteria (green area) as it only considers valid detections when the hand pose is also correctly estimated.

Qualitative results: We invite the reader to view our supplementary videos. We illustrate successes in difficult

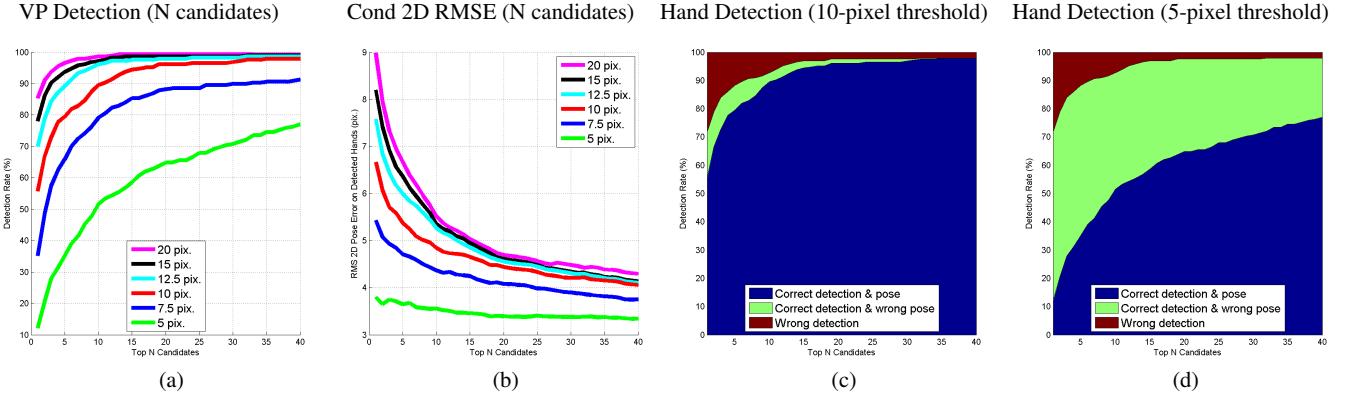


Fig. 14 Pixel threshold. Effect of the pixel threshold on viewpoint-consistent detections (a) and 2D RMS error conditioned on well-detected hands(b). A lower pixel threshold decreases detection rate but increases pose accuracy on detected hands, while a higher threshold increases detection rate and decreases pose accuracy. We chose to use a threshold of 10 pixels as trade-off between these 2 criteria. In c and d, we show the percentage of detection and pose correctness for 2 different thresholds: 10 pixels (c) used all over the main paper and 5 pixels (d). The green area corresponds to correct detections (in term of bounding boxes overlap between estimated and ground-truth poses) but incorrect poses. The measure we use for detection (blue area) is much more strict as it only considers valid a detection when the hand pose is also correctly estimated.

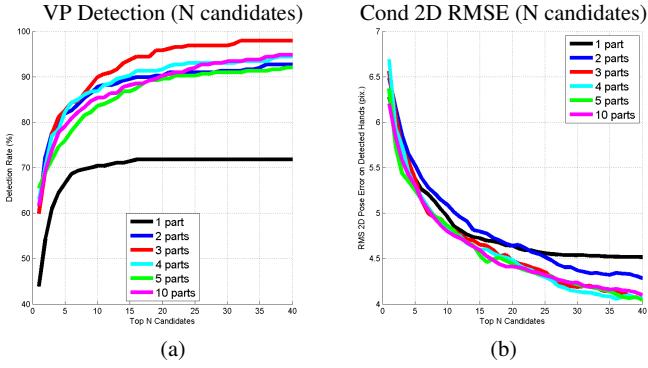


Fig. 13 Choice of the number of parts. We evaluate the importance of choosing the right number of parts in our cascade model by analyzing the performance achieved for different numbers. For each case, we compute the viewpoint-consistent detections (a) and 2D RMS error conditioned on well-detected hands (b) when varying the number of possible candidates.

scenarios in Fig. 15 and analyze common failure modes in Fig. 16. Please see the figures for additional discussion.

5 Conclusion

We have focused on the task of hand pose estimation from egocentric viewpoints. For this problem specification, we have shown that TOF depth sensors are particularly informative for extracting near-field interactions of the camera wearer with his/her environment. We describe a detailed, computer graphics model for generating egocentric training data with realistic full-body and object interactions. We use this data to train discriminative K-way classifiers for quantized pose estimation. To deal with a large number of classes, we advance previous methods for hierarchical cascades of multi-class rejectors, both in terms of accuracy and

speed. Finally, we have provided an insightful analysis of the performance of our algorithm on a new real-world annotated dataset of egocentric scenes. Our method provides state-of-the-art performance for both hand detection and pose estimation in egocentric RGB-D images.

References

1. <https://3dwarehouse.sketchup.com/> 5
2. Ballan, L., Taneja, A., Gall, J., Gool, L.J.V., Pollefeys, M.: Motion capture of hands in action using discriminative salient points. In: ECCV (6), pp. 640–653 (2012) 3
3. den Bergh, M.V., Gool, L.J.V.: Combining rgb and tof cameras for real-time 3d hand gesture interaction. In: WACV, pp. 66–72 (2011) 1
4. Chen, B., Perona, P., Bourdev, L.: Hierarchical cascade of classifiers for efficient poselet evaluation. In: BMVC, pp. 1–10 (2014) 3, 5, 6, 7
5. Damen, D., Gee, A.P., Mayol-Cuevas, W.W., Calway, A.: Egocentric real-time workspace monitoring using an rgb-d camera. In: IROS (2012) 3
6. Daz3D: Every-hands pose library. <http://www.daz3d.com/everyday-hands-poses-for-v4-and-m4> (2013) 4
7. Dominguez, S., Keaton, T., Sayed, A.: A robust finger tracking method for multimodal wearable computer interfacing. Multimedia, IEEE Transactions on **8**(5), 956–972 (2006). DOI 10.1109/TMM.2006.879872 2
8. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. CVIU **108**(1-2), 52–73 (2007) 3
9. Fathi, A., Farhadi, A., Rehg, J.: Understanding egocentric activities. In: ICCV (2011) 2
10. Fathi, A., Ren, X., Rehg, J.: Learning to recognize objects in egocentric activities. In: CVPR (2011) 2
11. Fielder, A.R., Moseley, M.J.: Does stereopsis matter in humans? Eye **10**(2), 233–238 (1996) 2
12. Gavrila, D.M., Philomin, V.: Real-time object detection for smart vehicles. In: Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on, vol. 1, pp. 87–93. IEEE (1999) 3

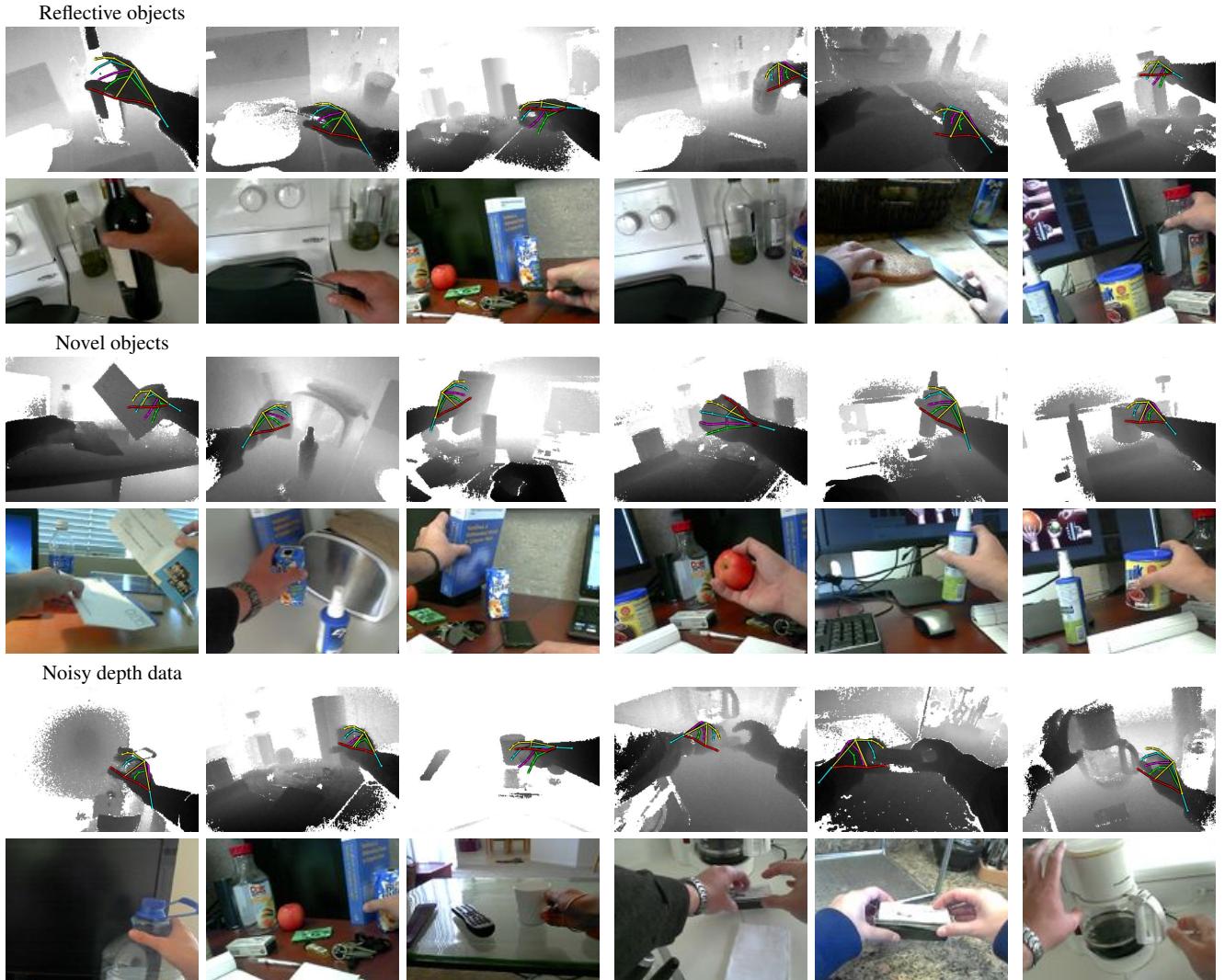


Fig. 15 Good detections. We show a sample of challenging frames where the hand is correctly detected by our system. Reflective objects (**top row**: wine bottle, pan, phone, knife and plastic bottle) produce incorrect depth maps due to interactions with our sensor’s infrared illuminant. Novel objects (**middle row**: envelope, juice box, book, apple, spray and chocolate powder box) require generalization to objects not synthesized at train-time, while noisy depth data (**bottom row**) showcases the robustness of our system.

13. Hamer, H., Schindler, K., Koller-Meier, E., Gool, L.J.V.: Tracking a hand manipulating an object. In: ICCV (2009) **3**
14. Hodges, S., Williams, L., Berry, E., Izadi, S., Srinivasan, J., Butler, A., Smyth, G., Kapur, N., Wood, K.: SenseCam: A retrospective memory aid. UbiComp (2006) **1**
15. Intel: Perceptual computing sdk (2013). URL <http://software.intel.com/en-us/vcsource/tools/perceptual-computing-sdk> **8, 9**
16. Keskin, C., Kiraç, F., Kara, Y.E., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: ECCV 2012, pp. 852–863 (2012) **9, 10**
17. Keskin, C., Kiraç, F., Kara, Y.E., Akarun, L.: Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In: ECCV (6), pp. 852–863 (2012) **1, 3**
18. Kölsch, M.: An appearance-based prior for hand tracking. In: ACIVS (2), pp. 292–303 (2010) **2**
19. Kölsch, M., Turk, M.: Hand tracking with flocks of features. In: CVPR (2), p. 1187 (2005) **2**
20. Kurata, T., Kato, T., Kourogi, M., Jung, K., Endo, K.: A functionally-distributed hand tracking method for wearable visual interfaces and its applications. In: MVA, pp. 84–89 (2002) **2**
21. Kyriazis, N., Argyros, A.A.: Physically plausible 3d scene tracking: The single actor hypothesis. In: CVPR (2013) **1, 3**
22. Kyriazis, N., Argyros, A.A.: Scalable 3d tracking of multiple interacting objects. In: CVPR (2014) **3**
23. de La Gorce, M., Fleet, D.J., Paragios, N.: Model-based 3d hand pose estimation from monocular video. IEEE PAMI **33**(9), 1793–1805 (2011) **3**
24. Li, C., M. Kitani, K.: Model recommendation with virtual probes for egocentric hand detection. In: ICCV (2013) **2, 9**
25. Li, C., M. Kitani, K.: Pixel-level hand detection in ego-centric videos. In: CVPR (2013) **2**
26. Lin, Y., Hua, G., Mordohai, P.: Egocentric object recognition leveraging the 3d shape of the grasping hand. In: ECCV Workshop on Consumer Depth Camera for Vision (CDC4V), pp. 1–11 (2014) **3**
27. Mann, S., Huang, J., Janzen, R., Lo, R., Rampersad, V., Chen, A., Doha, T.: Blind navigation with a wearable range camera and

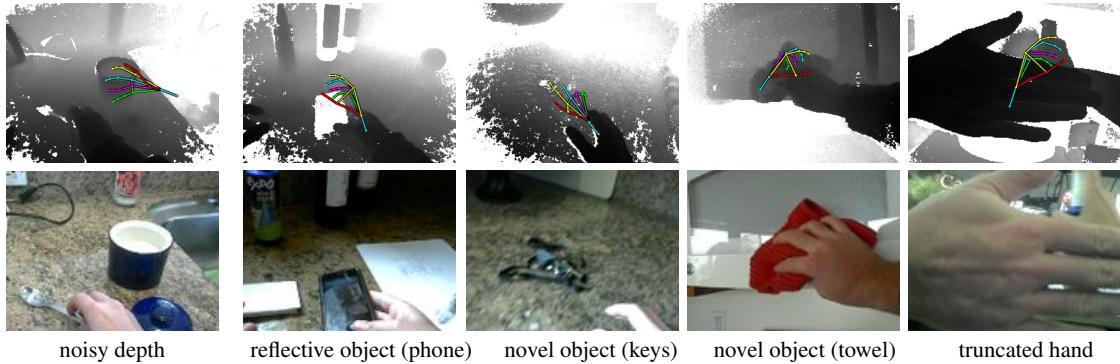


Fig. 16 Hard cases. We show frames where the hand is not correctly detected by our system, even with 40 candidates. These hard cases include excessively-noisy depth data, hands manipulating reflective material (phone) or unseen/deformable objects that look considerably different from those in our training set (e.g. keys, towels), and truncated hands.

- vibrotactile helmet. In: ACM International Conf. on Multimedia, MM '11 (2011) 2
28. Mayol, W., Davison, A., Tordoff, B., Molton, N., Murray, D.: Interaction between hand and wearable camera in 2d and 3d environments. In: BMVC (2004) 2
 29. Morerio, P., Marcenaro, L., Regazzoni, C.S.: Hand detection in first person vision. In: FUSION (2013) 2
 30. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3d tracking of hand articulations using kinect. In: BMVC (2011) 1, 3, 9
 31. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In: ICCV (2011) 3
 32. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the Articulated Motion of Two Strongly Interacting Hands. In: CVPR (2012) 1, 3
 33. Ong, E.J., Bowden, R.: A boosted classifier tree for hand shape detection. In: FGR (2004) 3
 34. Pirsavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR (2012) 2
 35. PrimeSense: Nite2 middleware (2013). URL <http://www.openni.org/files/nite/> 9
 36. Qian, C., Sun, X., Wei, Y., Tang, X., Sun, J.: Realtime and robust hand tracking from depth. In: CVPR (2014) 1, 2, 3
 37. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: CVPR, pp. 3137–3144. IEEE (2010) 2
 38. Ren, X., Philipose, M.: Egocentric recognition of handled objects: Benchmark and analysis. In: IEEE Workshop on Egocentric Vision (2009) 2
 39. Rogez, G., Khademi, M., Supancic, J., Montiel, J., Ramanan, D.: 3d hand pose detection in egocentric rgbd images. In: ECCV Workshop on Consumer Depth Camera for Vision (CDC4V), pp. 1–11 (2014) 2
 40. Rogez, G., Rihan, J., Orrite, C., Torr, P.H.S.: Fast human pose detection using randomized hierarchical cascades of rejectors. IJCV **99**(1), 25–52 (2012) 3, 5, 6, 7, 11
 41. Romero, J., Feix, T., Kjellstrom, H., Kragic, D.: Spatio-temporal modeling of grasping actions. In: IROS (2010) 4
 42. Romero, J., Kjellstrom, H., Ek, C.H., Kragic, D.: Non-parametric hand pose estimation with object context. Im. and Vision Comp. **31**(8), 555 – 564 (2013) 3
 43. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: CVPR (2013) 2
 44. Sakata, H., Taira, M., Kusunoki, M., Murata, A., Tsutsui, K.i., Tanaka, Y., Shein, W.N., Miyashita, Y.: Neural representation of three-dimensional features of manipulation objects with stereopsis. Experimental Brain Research **128**(1-2), 160–169 (1999) 2
 45. Sense, P.: The primesensortmreference design 1.08. Prime Sense (2011) 8
 46. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, pp. 750–757. IEEE (2003) 4
 47. Shotton, J., Fitzgibbon, A.W., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: CVPR (2011) 1, 3
 48. SmithMicro: Poser10. <http://poser.smithmicro.com/> (2010) 4
 49. Spinello, L., Arras, K.O.: People detection in rgb-d data. In: IROS (2011) 8
 50. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive markerless articulated hand motion tracking using rgb and depth data. In: ICCV (2013) 3
 51. Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computing. In: International Symposium on Wearable Computing (1998) 2
 52. Stenger, B., Thayananthan, A., Torr, P., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. PAMI **28**(9), 1372–1384, (2006) 3
 53. Stenger, B., Thayananthan, A., Torr, P.H., Cipolla, R.: Estimating 3d hand pose using hierarchical multi-label classification. Image and Vision Computing **25**(12), 1885–1894 (2007) 3, 7
 54. Tang, D., Chang, H.J., Tejani, A., Kim, T.: Latent regression forest: Structured estimation of 3d articulated hand posture. In: CVPR (2014) 3, 9
 55. Tang, D., Kim, T.H.Y.T.K.: Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In: ICCV (2013) 1, 3
 56. Tompson, J., Stein, M., LeCun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM Trans. Graph. **33**(5), 169 (2014) 3
 57. Šarić, M.: Libhand: A library for hand articulation (2011). URL <http://www.libhand.org/>. Version 0.9 4
 58. Xu, C., Cheng, L.: Efficient hand pose estimation from a single depth image. In: ICCV (2013) 3, 9
 59. Yang, R., Sarkar, S., Loeding, B.L.: Handling movement epenthesis and hand segmentation ambiguities in continuous sign language recognition using nested dynamic programming. PAMI **32**(3), 462–477 (2010) 1
 60. Z. Zheng, J., De La Rosa, S., M. Dollar, A.: An investigation of grasp type and frequency in daily household and machine shop tasks. In: ICRA, pp. 4169–4175 (2011) 5
 61. Zehnder, P., Koller-Meier, E., Van Gool, L.J.: An efficient shared multi-class detection cascade. In: BMVC, pp. 1–10 (2008) 3, 5, 6