3D Hand Pose Detection in Egocentric RGB-D Images

Grégory Rogez[1,2], Maryam Khademi[1], J. S. Supančič III[1], J. M. M. Montiel[2] and Deva Ramanan[1]

[1]Dept. of Computer Science, University of California, Irvine, USA
{grogez,mkhademi,supanci,dramanan}@ics.uci.edu
[2]Aragon Institute of Engineering Research (i3A), Universidad de Zaragoza, Spain
{grogez,josemari}@unizar.es

Abstract. We focus on the task of hand pose estimation from egocentric viewpoints. For this problem specification, 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 exacerbated when considering a wearable sensor and a first-person camera viewpoint: the occlusions inherent to the particular camera view and the limitations in terms of field of view make the problem even more difficult. We propose to use task and viewpoint specific synthetic training exemplars in a discriminative detection framework. We also exploit the depth features for a sparser and faster detection. We evaluate our approach on a real-world annotated dataset and propose a novel annotation technique for accurate 3D hand labelling even in case of partial occlusions.

Keywords: egocentric vision, hand pose, multi-class classifier, RGB-D sensor

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 [1], medical rehabilitation [2], and augmented reality [3], require inferring the interactions of the first-person observer with his/her environment. Towards that end, we specifically focus on the task of hand pose estimation from egocentric viewpoints. 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 [4] and prior work on articulated hand pose estimation for RGB-D sensors [5–7]. Recent approaches have also tried to exploit the 2.5D data from Kinect-like devices to understand complex scenarios such as object manipulation [6] or two interacting hands [5]. 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: Most previous work has formulated the hand pose recognition task as a tracking problem given RGB or RGBD sequences with manual initialization. We

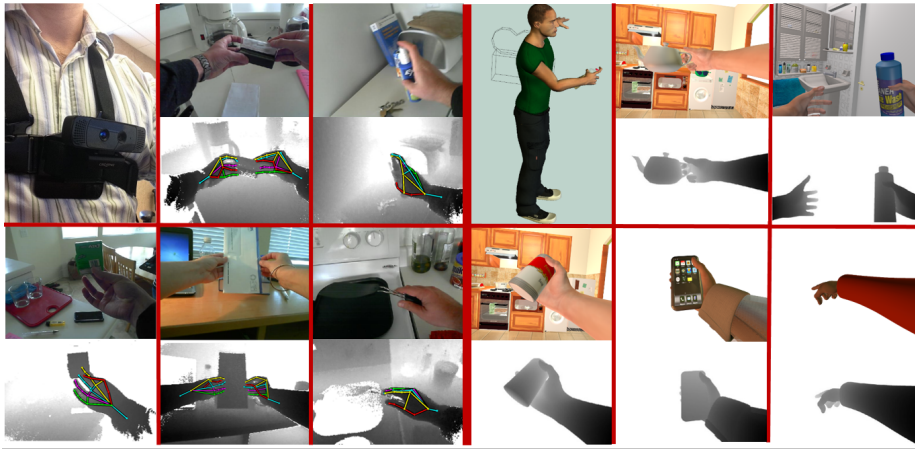

Fig. 1. Testing (left) and training data (right). We show on the left hand side several examples of annotated training RGBD images captured with a chest-mounted Intel Creative camera. On the right, we present some examples of training images rendered using Poser.

would like a fully-automatic method that processes egocentric videos of daily activities, which is even more challenging for the following reasons. First, a limited field-of-view from an egocentric viewpoint causes hands to frequently move outside the camera view frustum, making it difficult to apply tracking models that rely on accurate estimates from previous frames. Second, fingers are often occluded by the hand (and possible other objects being manipulated) in first-person viewpoints, making hand detection and articulated pose estimation more challenging than the typically third-person viewpoint (see Fig. 2).

Our approach: We describe a successful approach to hand-pose estimation that makes use of the following key observations. First, motivated by biological evidence [8], we show that **depth cues** provide an extraordinarily helpful signal for pose estimation in the near-field, first-person viewpoints. We find that time-of-flight depth cameras provide good depth estimates over a near-field workspace (0-70cm from the camera) while being easily mobile. Second, the egocentric setting provides a strong **viewpoint, shape, and interacting-object prior** over hand poses. We operationalize this prior by building parametric models over viewpoints of poses of a 3D, mesh-based hand model while interacting with common household objects. We then sample from this model (with an egocentric prior over viewpoint and hand shape) to generate large, synthetic depth data for training hand classifiers (see Fig. 2b). Third, **sparse, discriminative** classifiers allow us efficiently evaluate a large family of pose-specific classifiers. 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

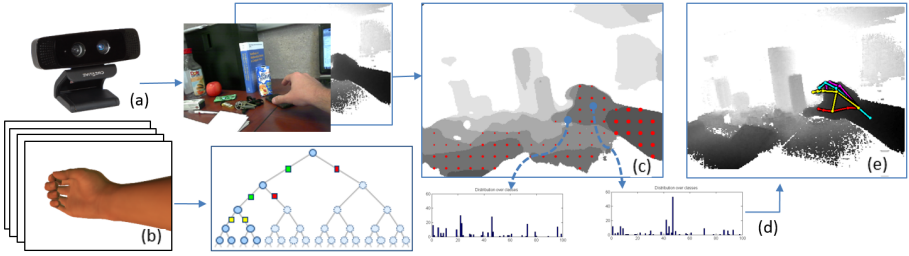

Fig. 2. System. (a) Chest-mounted RGB-D camera. (b) Synthetic egocentric hand exemplars are used to train a multi-class cascade classifier. The depth map is processed to select a sparse set of image locations (c) which are classified obtaining distributions over probable hand poses (d). An estimate is made e.g., by taking the max over these distributions (e).

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 RGBD 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.

2 Related Work

Egocentric hand/object manipulation: 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. Previous work examined the problem of recognizing objects [9, 10] and interpreting American Sign Language poses [11] from wearable cameras. Much work has also focused on hand tracking [12–15], finger tracking [16], and hand-eye tracking [17] from wearable cameras. Often, hand pose estimation is examined during active object manipulations [18–22]. One commonality behind such previous work is the use of RGB sensor input. Motivated in part by biological evidence [8], we show that depth cues considerably aids the processing of such near-field interactions.

Depth-based pose estimation: Our technical approach is closely inspired by the Kinect system [4], which also makes use of synthetically generated depth maps for articulated pose estimation. Our approach differs in that we construct classifiers that classify entire poses rather than local landmarks or parts. We posit and verify that the numerous occlusions of articulated fingers from a wearable viewpoint requires a more global approach, since local information can be ambiguous due to occlusions. For

this reason, temporal reasoning is also particularly attractive because one can use dynamics to resolve such ambiguities. Much prior work on hand-pose estimation takes this route [23, 7, 24]. Our approach differs from these approaches in that we focus on single-image hand pose estimation, which is required to avoid manual (re)initialization. A notable exception is the recent work of [25], who also process single images but focus on third-person views.

Egocentric RGB-D: Depth-based wearable cameras are attractive because depth cues can be used to better reason about occlusions arising from egocentric viewpoints. There has been surprisingly little prior work in this vein, with notable exceptions focusing on targeted applications such as navigation for the blind [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.

Features: Many methods based on RGB images rely on color-based skin detection and segmentation. Examples for hand tracking from a moving camera can be found in [27] or more recently [17]. Earlier algorithms for hand pose estimation based on RGB images can be found in [28]. Recent work has exploited Kinect-like depth sensors [25]. RGB and Time-of-Flight (ToF) cameras have also been combined for real-time 3D hand gesture interaction [3] or near-realtime detailed hand pose estimation [29].

Generative vs discriminative: Generative model-based approaches have historically been more popular for hand pose estimation [30]. A detailed 3D model of the hand pose is usually employed for articulated pose tracking [31, 5] and detailed 3D pose estimation [32]. Discriminative approaches [7, 24] for hand pose estimation tend to require large datasets of training examples, synthetic, realistic or combined [7]. Learning formalisms include boosted classifier trees [33] and randomized decision forests[24], and regression forests [7]. We describe a discriminative approach based on [34], which uses a tree-structured multi-class cascade for pose estimation and detection. We specifically extend the work of [34] to use both RGB and Depth features.

3 Our method

3.1 Setting and Choice of the Device

We use a chest-mounted Time-of-Flight camera, an Intel Creative (see Fig. 2a), which is particularly well-suited for short-range hand-object interactions.

TOF vs Structured Light: Much recent work on depth-processing has been driven by the consumer-grade PrimeSense sensor [35], 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 [36].

3.2 Synthetic Training Exemplars

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` [37] and commercial Poser [38, 39]. 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 mimicing our physical data collection process. To generate data corresponding to different body and hand shapes, we make use of Poser's character library.

Viewpoint prior: To specify a viewpoint prior for floating hands, we simply 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). For our full character models, we generate small perturbations of the virtual chest camera mount. This simulates camera viewpoint and body variation between individuals wearing egocentric cameras. We use forward kinematics of the arm to naturally limit hand poses to realistic viewpoints, another benefit of full-character egocentric modeling.

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 $\{\theta_i\}$ from a collection of grasping motion capture data [40]. 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 be perturbing each original sample with Gaussian noise $\theta_i + \epsilon$, where $\epsilon \sim N(0, \sigma I)$. Notably, we also perturb the *entire arm* of the full character-body, which generates natural (egocentric) viewpoint variations of hand configurations. We remove those samples that result in poses that are self-intersecting or lie outside the field-of-view. Example poses are shown in Fig. 2b.

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 EveryDayHands animation library [41], which contains 50 canonical hand grasps and objects. This package was originally designed as a computer animation tool, but we find the library to cover a reasonable taxonomy of grasps and objects for egocentric recognition. Objects include general shapes

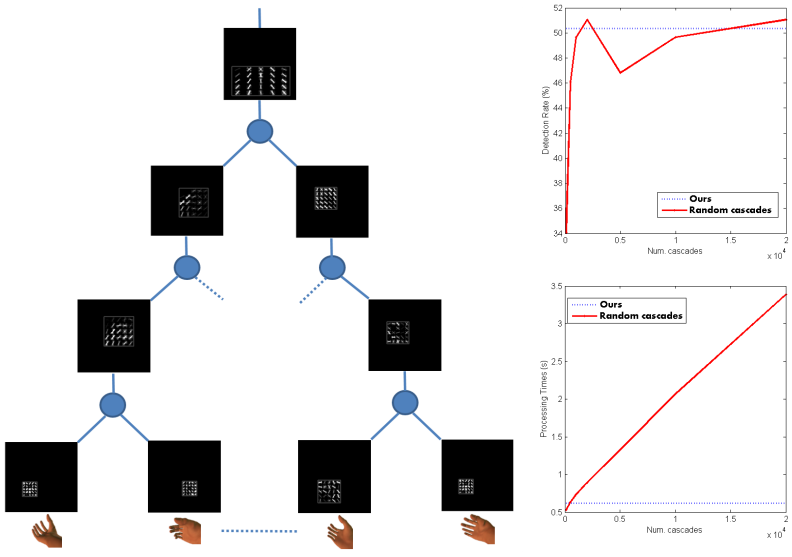

Fig. 3. (left) Hierarchical cascade of parts. (right) Detection rates and processing time varying the number of random cascades in the ensemble vs an exponential number of cascades. The hierarchy shows the coarse to fine detection, showing only one part per branch for clarity. We show on the upper right that our new detector is equivalent to an infinite number of Random Cascades (RC) from [34]. In the bottom right 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.

such as balls and cylinders of varying size, as well as common everyday objects including utensils, phones, cups, etc. We apply our rejection-sampling technique to generate additional valid interacting object-hand grasp configurations, yielding a final dataset of 10,000 synthetic egocentric hand-object examples (see examples in Fig. 1).

3.3 Hierarchical Cascades (past work)

We would like a hand pose detector that simultaneously performs hand detection and pose estimation. We describe an approach based on the multi-class rejection-cascade classifiers of [34]. We review the basic formulation here, but refer the reader to [34] for further detail. From a high-level, both detection and pose estimation are treated as a K-way classification problem, with classes specifying one of K discrete poses or the background. The $K+1$-way classifier is trained using an boosting-like algorithm where weak-classifiers are "parts" classifiers trained using linear SVMs defined on localized HOG features within a scanning-window coordinate frame. A multi-way classification strategy may require considerable amounts of training data and may be slow at test time, since K can be large. [34] describe an approach with three crucial properties that address these limitations, discussed below. We then describe various improvements that apply in our problem domain.

Coarse-to-fine sharing: Parts are *shared* across all K pose-classes through a hierarchical coarse-to-fine tree. Specifically, hierarchical K-means is used to cluster our set of training poses into K quantized pose classes, which are naturally arranged in a hierarchical tree with K leaves. The i^{th} node in this tree represents a coarse pose class; a visualization is shown in Fig. 3. Given an image window x, a binary classifier tuned for coarse pose-class i is defined as:

$$f_i(x) = \prod_{j \in \text{ancestors}(i)} h_j(x) \quad \text{where} \quad h_j(x) = \mathbf{1}_{[w_j^T x > b_j]} \tag{1}$$

where the ancestors of node i include i, and $\mathbf{1}$ is the indicator function (evaluating to 1 or 0). Each binary "weak classifier" $h_j(x)$ is a thresholded linear function (trained with a linear SVM) that is defined on localized HOG features extracted from a subwindow of x. This allows us to interpret w_j as a zero-padded "part" template. Parts higher in the tree tend to be generic, and used across many pose classes. Parts lower in the tree, toward the leaves, tend to capture pose-specific details.

Rejection cascades: At test time, the above set of hierarchical classifiers can naturally be implemented as a rejection cascade with a breadth-first search through the tree. This can be readily seen by rewriting (1) recursively as $f_i(x) = f_p(x)h_i(x)$, where p is the parent of i. We only need to evaluate the descendants of node i if $h_i(x)$ evaluates to 1. This means we can quickly prune away large portions of the pose-space when evaluating region x, making scanning-window evaluation at test-time quite efficient. Finally, a notable byproduct is that multiple leaf classes might fire in a given image region, each with a different leaf score. We generally report the highest-scoring pose as the final result, but show that alternate high-scoring hypotheses can still be useful (since they can be later refined using say, a tracker).

Ensembles of cascades: To increase robustness, we would like to average predictions across an ensemble of classifiers. [34] describes an approach that makes use of a pool of weak part classifiers for each coarse-pose class (node) i:

$$h_i(x) \in H_i \quad \text{where} \quad |H_i| = M, i \in N \tag{2}$$

One can instantiate a tree by selecting a weak classifier (from its candidate pool H_i) for each node i in the tree. This allows one to define an exponential number of instantiations M^N, where N is the number of nodes (coarse pose-classes) in the tree and M is the size of each candidate pool. In practice, [34] found that averaging predictions from a small random subset of trees significantly improved results.

3.4 Joint Training of Exponential Ensembles

In this section, we present several improvements to [34] 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 for jointly training weak classifiers and averaging predictions over exponentially-large sets of cascade ensembles.

Sequential training: Notably, [34] *independently* learned weak classifiers (w_i, b_i) by defining a positive/negative training set that is independent of other weak classifiers.

That is, *all* training images corresponding to node (coarse pose-class) i are treated as positives, and *all* other poses are treated as negatives. Instead, we use only the training examples that pass through the rejection cascade up to node i. This more accurately reflects the scenario at test-time. This requires classifiers to be trained in a sequential fashion, in a similar coarse-to-fine (breadth-first) search over nodes from the root to the leaves. This means weak learners are trained jointly rather than independently.,

Exponentially-large ensembles: Rogez et al. [34] average votes across a small number (around hundred) of explicitly-constructed trees. We now describe a simple procedure for exactly computing the average over the exponentially-large set of M^N. By averaging over a large set, we reduce the chance of a premature cascade rejection. Our insight is that one can compute an *implicit* summation (or average) over the set by caching partial summations. Assume nodes are numbered in breadth-first order, such that node 1 is the root. As before, we will iterate over nodes in a breadth-first, coarse-to-fine manner. We now apply *all* H_i weak classifiers and keep a record of the number that fire n_i:

$$t_i = t_p n_i \quad \text{where} \quad p = \text{parent}(i), \quad n_i = \sum_{h_i(x) \in H_i} h_i(x) \tag{3}$$

For any node i, the fraction of partial hierarchies (constructed from the root to node i) that vote for node i are given by the ratio $\frac{t_i}{M^{D(i)}}$, where $D(i)$ is the depth of node i. We omit the fairly straightforward proof due to lack of space. Hence the ratio for leaf nodes i yields the final set of (fine-scale) pose-class votes. Notably, once we reach an internal node for which no weak classifiers fire $n_i = 0$, then all of its descendants must generate votes $t_i = 0$, meaning that they need not be evaluated. This still allows for efficient run-time search (see Fig. 3).

Features: We experiment with two additional sets of features x. Rogez et al. [34] originally defined their model on oriented gradient histograms on RGB (HOG-RGB). 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) [42]. For depth we use 5x5 HOG blocks and 16 signed orientation bins.

3.5 Sparse Search

Two assumptions can be leveraged to effectively tackle egocentric hand detection in RGB-D images: 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 template classifier 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.

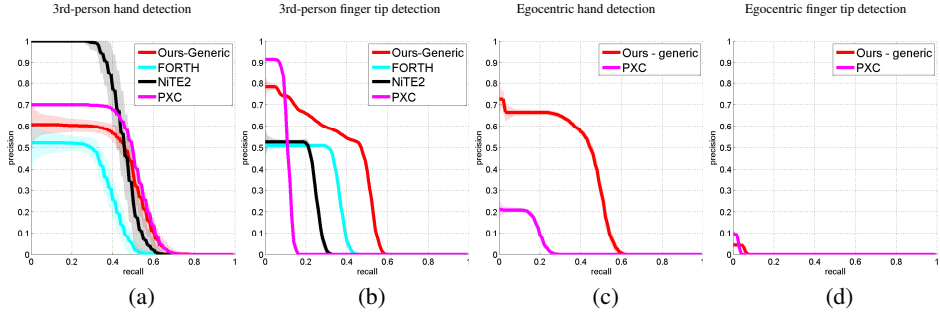

Fig. 4. Numerical results for 3rd-person (a-b) and egocentric (c-d) sequences. We compare our method (tuned for generic priors) to state-of-the-art techniques from industry (NITE2 [43] and PXC [36]) and academia (FORTH [23]) in terms of (a) hand detection and (b) finger tips detection. We shade the 95% confidence interval obtained from the statistical bootstrap. We refer the reader to the main text for additional decription, 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 Experiments

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 research. 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. We invite the reader to view videos in our supplementary material.

Evaluation: We present numerous evaluations for both hand detection 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%. We evaluate pose estimation with *2D-RMS* re-projection error of keypoints. However, some baseline systems report the pose of only confident fingers. In such cases, we measure *finger-tip detection* accuracy as a proxy for pose estimation. For additional diagnosis, we categorize errors into detection failures, correct detections but incorrect viewpoint, and correct detection and viewpoint but incorrect articulated pose. Specifically, *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 reprojection 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 layer be re-ranked, by say, a tracker).

Parameters: We train a cascade model trained with $K = 100$ classes, a hierarchy of 6 levels and $M = 3$ weak classifiers per branch. We synthesize 100 training images per class.

Third-person vs egocentric: We first evaluate our method (trained with a generic prior) for the tasks of third-person and egocentric hand analysis (Fig. 4). We compare against state of the art techniques from industry [43, 36] and academia [23]. Because a general prior span a larger range of viewpoints and poses, we train a model with $K = 800$ classes for this experiment. We only present egocentric results for our method and PXC, since the other baselines (FORTH and NITE2) are trackers that fail catastrophically on egocentric sequences where hands frequently leave the field-of-view. Moreover, because some baselines only report positions of confident fingers, we use finger-tip detection as a proxy for pose estimation. We conclude that (1) hand pose estimation is considerable harder in the egocentric setting and (2) our (generic-prior) pose estimation system is a state-of-the-art starting point for our subsequent analysis.

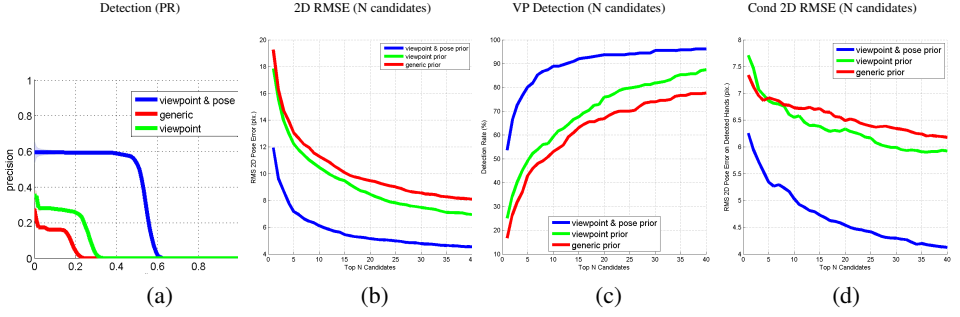

Fig. 5. Quantitative results for varying our priors evaluated with respect to (a) hand detection, (b) 2D RMS error, (c) viewpoint-consistent detections and (d) 2D RMS error conditioned on viewpoint-consistent detections. Please see text for detailed description of our evaluation criteria and analysis. In general, egocentric-pose priors considerably improve performance, validating our egocentric-synthesis engine from Sec. 3.2. When tuned for $N = 10$ candidates per image, our system produces pose hypotheses that appear accurate enough to initialize a tracker.

Pose+viewpoint prior: In Fig. 5, we show that an egocentric-specific pose and viewpoint prior outperforms the generic prior from Fig. 4. In general, a viewpoint prior produces a marginal improvement, while a pose prior considerably improves accuracy in all cases. 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. Our results suggest that our synthesis procedure from Sec. 3.2 correctly models both viewpoint and pose priors arising in egocentric settings.

Ablative analysis: To further analyze our system, we perform an ablative analysis that turns "off" different aspects of our system: sequential training, ensemble of cas-

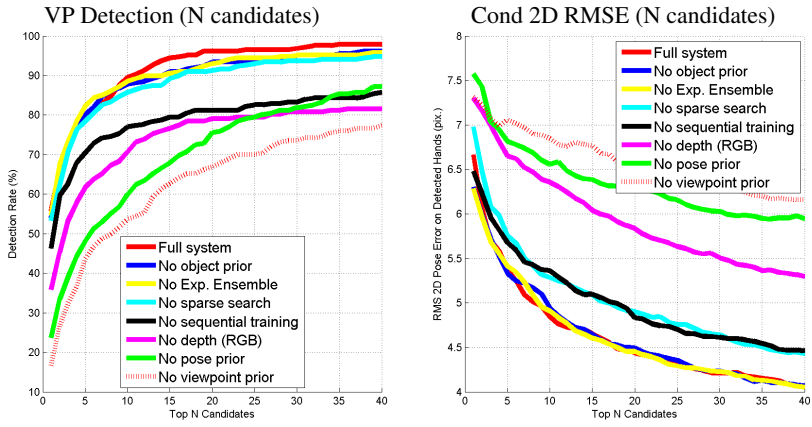

Fig. 6. We evaluate performance when turning off particular aspects of out system, considering both (a) viewpoint-consistent detections (b) 2D RMS error conditioned on well-detected hands. When turning off our exponentially-large ensemble or synthetic training, we use the default of 100 independently-trained cascades as in [34]. 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.

cades, depth feature, sparse search and different priors (viewpoint, pose, objects). Hand detection and conditional 2D hand RMS error are given in Fig. 6. Viewpoint prior, pose prior, depth HOG features and our new sequential training algorithm are key aspects in terms of performance. Turning these parameters off decreases the detection rate by a substantial amount (between 10 and 30%). 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.

Classifier design: Our new sequential training of parts significantly outperforms the independent training of [34] by 10-15% (Fig. 6). 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 [34]), while the sparse search is 3.15 times faster than a dense grid. Hence our final classifier significantly improves upon the accuracy and speed of [34], which uses a default of 100 random, independently-trained cascades evaluated on a dense grid of RGB features.

Qualitative results: We invite the reader to view our supplementary videos. We illustrate successes in difficult scenarios in Fig. 7 and analyze common failure modes in Fig. 8. Please see the figures for additional discussion.

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 have proposed to use task-specific synthetic training exemplars, trained with object interactions, in a discriminative detection framework. To do so efficiently, we have exploited a simple depth cue for fast detection. Finally, we have

Fig. 7. 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.

provided an insightful analysis of the performance of our algorithm on a new real-world annotated dataset of egocentric scenes.

Aknowledgements. This research was supported by the European Commission under FP7 Marie Curie IOF grant "Egovision4Health" (PIOF-GA-2012-328288).

References

1. 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)

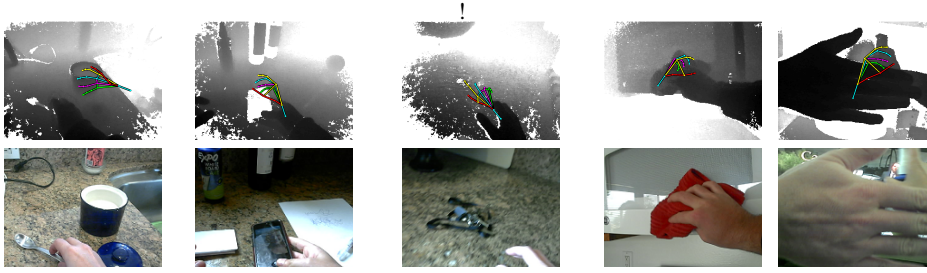

Fig. 8. Hard cases. We show frames where the hand is not correctly detected by our system, even with 40 candidates. These hard cases include excessively-noisy depth data, hands manipulating reflective material (phone) or unseen/deformable objects that look considerably different from those in our training set (e.g. keys, towels), and truncated hands.

2. 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) (2010) 462–477
3. den Bergh, M.V., Gool, L.J.V.: Combining rgb and tof cameras for real-time 3d hand gesture interaction. In: WACV. (2011) 66–72
4. 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)
5. Oikonomidis, I., Kyriazis, N., Argyros, A.A.: Tracking the Articulated Motion of Two Strongly Interacting Hands. In: CVPR. (2012)
6. Romero, J., Kjellstrom, H., Ek, C.H., Kragic, D.: Non-parametric hand pose estimation with object context. Im. and Vision Comp. **31**(8) (2013) 555 – 564
7. Tang, D., Kim, T.H.Y.T.K.: Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In: ICCV. (2013)
8. 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) (1999) 160–169
9. Fathi, A., Ren, X., Rehg, J.: Learning to recognize objects in egocentric activities. In: CVPR. (2011)
10. Pirsiavash, H., Ramanan, D.: Detecting activities of daily living in first-person camera views. In: CVPR. (2012)
11. Starner, T., Schiele, B., Pentland, A.: Visual contextual awareness in wearable computing. In: International Symposium on Wearable Computing. (1998)
12. 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. (2002) 84–89
13. Kölsch, M., Turk, M.: Hand tracking with flocks of features. In: CVPR (2). (2005) 1187
14. Kölsch, M.: An appearance-based prior for hand tracking. In: ACIVS (2). (2010) 292–303
15. Morerio, P., Marcenaro, L., Regazzoni, C.S.: Hand detection in first person vision. In: FUSION. (2013)
16. Dominguez, S., Keaton, T., Sayed, A.: A robust finger tracking method for multimodal wearable computer interfacing. Multimedia, IEEE Transactions on **8**(5) (2006) 956–972
17. Ryoo, M.S., Matthies, L.: First-person activity recognition: What are they doing to me? In: CVPR. (2013)

18. Mayol, W., Davison, A., Tordoff, B., Molton, N., Murray, D.: Interaction between hand and wearable camera in 2d and 3d environments. In: BMVC. (2004)
19. Ren, X., Philipose, M.: Egocentric recognition of handled objects: Benchmark and analysis. In: IEEE Workshop on Egocentric Vision. (2009)
20. Damen, D., Gee, A.P., Mayol-Cuevas, W.W., Calway, A.: Egocentric real-time workspace monitoring using an rgb-d camera. In: IROS. (2012)
21. Ren, X., Gu, C.: Figure-ground segmentation improves handled object recognition in egocentric video. In: CVPR, IEEE (2010) 3137–3144
22. Fathi, A., Farhadi, A., Rehg, J.: Understanding egocentric activities. In: ICCV. (2011)
23. Oikonomidis, I., Kyriazis, N., Argyros, A.: Efficient model-based 3d tracking of hand articulations using kinect. In: BMVC. (2011)
24. 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). (2012) 852–863
25. Xu, C., Cheng, L.: Efficient hand pose estimation from a single depth image. In: ICCV. (2013)
26. Mann, S., Huang, J., Janzen, R., Lo, R., Rampersad, V., Chen, A., Doha, T.: Blind navigation with a wearable range camera and vibrotactile helmet. In: ACM International Conf. on Multimedia. MM '11 (2011)
27. Argyros, A., Lourakis, M.: Real-time tracking of multiple skin-colored objects with a possibly moving camera. In: ECCV (3). (2004) 368–379
28. Erol, A., Bebis, G., Nicolescu, M., Boyle, R.D., Twombly, X.: Vision-based hand pose estimation: A review. CVIU **108**(1-2) (2007) 52–73
29. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive markerless articulated hand motion tracking using rgb and depth data. In: ICCV. (2013)
30. Stenger, B., Thayananthan, A., Torr, P., Cipolla, R.: Model-based hand tracking using a hierarchical bayesian filter. PAMI **28**(9) (2006) 1372–1384,
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)
32. de La Gorce, M., Fleet, D.J., Paragios, N.: Model-based 3d hand pose estimation from monocular video. IEEE PAMI **33**(9) (2011) 1793–1805
33. Ong, E.J., Bowden, R.: A boosted classifier tree for hand shape detection. In: FGR. (2004)
34. Rogez, G., Rihan, J., Orrite, C., Torr, P.H.S.: Fast human pose detection using randomized hierarchical cascades of rejectors. IJCV **99**(1) (2012) 25–52
35. Sense, P.: The primesensortmreference design 1.08. Prime Sense (2011)
36. Intel: Perceptual computing sdk (2013)
37. Šarić, M.: Libhand: A library for hand articulation (2011) Version 0.9.
38. SmithMicro: Poser10. http://poser.smithmicro.com/ (2010)
39. Shakhnarovich, G., Viola, P., Darrell, T.: Fast pose estimation with parameter-sensitive hashing. In: Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE (2003) 750–757
40. Romero, J., Feix, T., Kjellstrom, H., Kragic, D.: Spatio-temporal modeling of grasping actions. In: IROS. (2010)
41. Daz3D: Every-hands pose library. http://www.daz3d.com/everyday-hands-poses-for-v4-and-m4 (2013)
42. Spinello, L., Arras, K.O.: People detection in rgb-d data. In: IROS. (2011)
43. PrimeSense: Nite2 middleware (2013)