CBMM Memo No. 097 **December 2, 2018**

Partially Occluded Hands: A challenging new dataset for single-image hand pose estimation

**Battushig Myanganbayar, Cristina Mata, Gil Dekel,
Boris Katz, Guy Ben-Yosef, Andrei Barbu**

Abstract

Recognizing the pose of hands matters most when hands are interacting with other objects. To understand how well both machines and humans perform on single-image 2D hand-pose reconstruction from RGB images, we collected a challenging dataset of hands interacting with 148 objects. We used a novel methodology that provides the same hand in the same pose both with the object being present and occluding the hand and without the object occluding the hand. Additionally, we collected a wide range of grasps for each object designing the data collection methodology to ensure this diversity. Using this dataset we measured the performance of two state-of-the-art hand-pose recognition methods showing that both are extremely brittle when faced with even light occlusion from an object. This is not evident in previous datasets because they often avoid hand- object occlusions and because they are collected from videos where hands are often between objects and mostly unoccluded. We annotated a subset of the dataset and used that to show that humans are robust with respect to occlusion, and also to characterize human hand perception, the space of grasps that seem to be considered, and the accuracy of reconstructing occluded portions of hands. We expect that such data will be of interest to both the vision community for developing more robust hand-pose algorithms and to the robotic grasp planning community for learning such grasps. The dataset is available at occludedhands.com

This work was supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216.

Partially Occluded Hands: A challenging new dataset for single-image hand pose estimation

Battushig Myanganbayar, Cristina Mata, Gil Dekel
Boris Katz, Guy Ben-Yosef, Andrei Barbu

CSAIL, MIT, Cambridge MA 02139, USA
{btushig,cfmata,dekelg,boris,gby,abarbu}@mit.edu

Abstract. Recognizing the pose of hands matters most when hands are interacting with other objects. To understand how well both machines and humans perform on single-image 2D hand-pose reconstruction from RGB images, we collected a challenging dataset of hands interacting with 148 objects. We used a novel methodology that provides the same hand in the same pose both with the object being present and occluding the hand and without the object occluding the hand. Additionally, we collected a wide range of grasps for each object designing the data collection methodology to ensure this diversity. Using this dataset we measured the performance of two state-of-the-art hand-pose recognition methods showing that both are extremely brittle when faced with even light occlusion from an object. This is not evident in previous datasets because they often avoid hand-object occlusions and because they are collected from videos where hands are often between objects and mostly unoccluded. We annotated a subset of the dataset and used that to show that humans are robust with respect to occlusion, and also to characterize human hand perception, the space of grasps that seem to be considered, and the accuracy of reconstructing occluded portions of hands. We expect that such data will be of interest to both the vision community for developing more robust hand-pose algorithms and to the robotic grasp planning community for learning such grasps. The dataset is available at occludedhands.com

Keywords: Partial occlusion · dataset · RGB hand-pose reconstruction

This work was supported, in part, by the Center for Brains, Minds and Machines (CBMM) NSF STC award 1231216, the Toyota Research Institute, and the MIT-IBM Brain-Inspired Multimedia Comprehension project.

1 Introduction

Understanding what humans are doing nearly always requires reconstructing the poses of their bodies and in particular their hands. While body pose recognition from single RGB images has advanced significantly [1–5], hand-pose recognition from this type of data has received far less attention. Despite this, recent publications have shown results that reconstruct hand models to within a few pixels of human annotations [6, 7]. Here we investigate the performance of such models, introducing a challenging new dataset consisting of common grasps of common objects where those grasps are shown both with the objects occluding the hand and without the object present. We also provide the first measurements of human hand-pose reconstruction accuracy as a function of the proportion of the hand that is occluded against ground-truth annotations. In brief, we discover that existing machine vision systems are brittle with respect to occlusion while humans are robust and vastly outperform machines; we also introduce a new benchmark and target for hand-pose estimation algorithms.

Our dataset is large, consisting of 11,840 images of grasps of 148 object instances. Unlike in most other prior datasets, here each image is collected individually and not from a longer video. We eschew shooting videos and collecting frames from them despite the convenience of doing so because the resulting frames are highly correlated and, due to the mechanical constraints of humans the images, tend to display unoccluded hands moving from place to place. The correlation between the grasps and object types was minimized by asking subjects to perform multiple grasps with the same object. We annotated 400 images with 21 keypoints 4 times over in order to compute human inter-coder agreement. The dataset contains images of hands grasping objects followed by that same grasp but without the presence of the object. This allows us to compute the accuracy of human perception for partially occluded hands against the ground truth hand poses. We then provide a human benchmark on this dataset, finding that humans have 5.31 pixel agreement, which allows us to quantify the state of the art in hand-pose construction and the difference between human and machine perception. While on most other datasets hand-pose reconstruction works well, e.g., average Euclidean distance of 4 to 5 pixels on the Dexter datasets [8–10], the dataset we provide has an average error of 20 pixels making it far more challenging.

The contributions of this work are:

1. a new hand-pose dataset that focuses on recovering hand pose when it matters most: during hand-object interactions,
2. demonstrating that current hand-pose recognizers are brittle and fail when faced with occlusion despite scoring extremely well on previous datasets,
3. a novel methodology for gathering hand-pose datasets that allows us to produce, for the first time, a large set of pairs of hand grasps both with the object present and without the object occluding the hand,
4. an approach to gather many varied stable grasps per object,
5. the first baseline for human hand pose recognition accuracy showing that existing approaches vastly underperform humans.

2 Related work

Several hand-pose datasets already exist. Historically, most have focused on depth images rather than RGB images and few have had any hand-object interactions. Notably, the NYU Hand Pose Dataset [11] contains several long video sequences but is only geared toward use as a depth dataset as the RGB images are rectified rendering them unusuable for single-image hand-pose reconstruction. We do not discuss depth-only datasets further and point the reader to a recent survey of such [12]. The MPII Human Pose dataset [13] contains 11,701 test RGB images of humans engaging in common actions. While these are natural videos from YouTube and the frames are selected to be fairly far apart (at least five seconds, to ensure that they are decorrelated), in most frames hands do not grasp any object. We focus in particular on grasps because they naturally result in high occlusion and because they are so critical to understanding how others are manipulating objects. Few images in the MPII Human Pose dataset are annotated with hand poses. The Yale Human hand grasping dataset [14], while very large at 27.7 hours of video, contains 9100 RGB images of hands grasping 11 object classes shot from an egocentric point of view. As these are derived from a small number of video sequences where subjects performed repetitive tasks with a small number of objects, the same grasps reoccur many times.

Dexter+Object EgoDexter **Ours**

Fig. 1. A comparison of grasps from Dexter+Object, EgoDexter, and our dataset. Note that previous datasets are designed to remove occlusions and have subjects engage in careful grasps to do so. In our dataset, most subjects were naive having no background in computer vision and only a generic knowledge of how the dataset would be used. This leads to much more natural grasps that significantly occlude the hands.

The closest datasets to the one presented here are Dexter [8], Ego-Dexter [9], and Dexter+Object [10]. They are all collected using the same general procedures. Video sequences are shot of humans changing their hand pose, in the latter two cases while interacting with objects. Since these datasets are collected by shooting videos rather than images, both the grasps in adjacent frames and the pixel values of the frames themselves are highly correlated. This effectively reduces their dataset size significantly. More fundamentally, even though they contain hand-object interactions, hands must travel to arrive at objects and manipulate them. This means that many of the frames do not actually contain hand-object interactions. The three Dexter datasets were collected by subjects who were motivated to make their grasps and interactions as plain, simple, and obvious as

possible. See fig. 1 for an example of what we mean here; the grasps in Dexter avoid hand occlusion and are designed to be easy. Most of our subjects were naive, they were paid for their data collection services, essentially all of the data was collected by subjects not connected to this publication, and they were generally only vaguely aware of the purpose of the dataset. This resulted in grasps that are far more natural and, when combined with our methodology to increase the variety of grasps, resulted in a much larger range of hand poses. It is simply a consequence of natural human object grasps that they result in images where the hands are highly occluded. The biases of subjects and those inherent in data collection protocols have been shown to lead to a significant overstating of machine accuracy in other domains such as natural language processing [15].

Table 1. A comparison of statistics of RGB hand-pose datasets. Within a sequence, poses are highly correlated meaning that the datasets are effectively far smaller than it may first appear. This is in part because most previous datasets are collections of video snippets. In our case, dataset size is effectively reduced by a factor of two because the dataset contains the same hand pose both occluded and unoccluded by an object. The dataset presented here is much larger than previous datasets, with many more decorrelated hand poses and with more controls to provide a variety of hand poses

Dataset	# of sequences	# of frames	# of objects
Dexter	7 videos	1,750	N/A
Dexter+Object	6 videos	3,151	1
Ego-Dexter	4 videos	3,194	17
Ours	**5,920**	**11,840**	**148**

Previous investigations have attempted to characterize human hand perception although no concrete results exist showing the agreement and accuracy of human hand-pose recognition. Santello *et al.* provide a recent overview of the neural and biological basis for hand control [16]. Existing datasets have at most one annotation and rarely have any ground truth data. This means measuring the accuracy of human perception on this task, and, by extension, determining how far away machine results are from human-level performance, is not possible.

In robotics, grasp planning has been investigated [17]. This has included large datasets of grasps but they consist generally of images or videos of attempted robotic grasps [18]. In some cases, such datasets have been synthesized automatically, a potentially useful approach for human hand-pose recognition [19]. The stability of a grasp is of key importance to robotics and its prediction plays a role in some datasets [20]. While we do not consider physical stability in this publication, we do intend to, in the future, investigate if perception is affected by notions of stability using this data.

3 A dataset of partially occluded hands

We collected a dataset of 11,840 images of hands holding objects. We chose 148 object instances, initially 150 objects but two were damaged during imaging thereby changing their appearance and prompting their removal. Human subjects were asked to hold those objects. Each time a subject held an object, it was then taken from them while they remained in the same pose and another image was shot. This provides the same grasp, up to some minor error, both as it would appear naturally occluded by the object being grasped and unoccluded by the object. Since many objects have a default use, humans tend to grasp them in a consistent manner. To avoid this, and to prevent merely learning and evaluating the default grasp thereby giving the appearance of hand-pose recognition without there being any, we asked subjects to then hold the object in a different way, as determined by the subjects. Each trial then consists of a quad of a hand holding an object in two ways, each both with the object and then without. In what follows we describe the rationale, methodology, and contents of the dataset.

The 148 object instances, almost all from different object classes, were chosen to reflect common objects present in homes that might be of interest to both activity recognition in households and to robot grasping. The chosen objects were opaque to ensure that they could properly occlude the hand and large enough to do so; this ruled out some common but far too small objects like most cutlery. Both deforamble and non-deformable objects were included; with 20 out of the 148 objects being deformable. An overview of the objects selected is shown in fig. 2. The dataset is balanced with respect to the object instances with 80 images for each object. The objects have a diverse set of possible grasps since they serve different purposes; they were chosen to have different topologies, and have different weights putting constraints on the possible stable grasps.

We designed the data collection procedure to increase the set of grasps that were employed by the subjects. First, an image was collected of a subject grasping an object. Next, that object was removed from the grasp of the subject and another image was collected. In this second image the grasp is evident as it is no longer occluded by the object. It proved to be critical that another individual remove the object from the grasp of the subject so that they could maintain their hand in the same pose. Then the subject was asked to grasp that same object but in a different manner. We did not control for this but subjects were given several instructions to help them identify other reasonable grasps such as imagining the object being much heavier, imagining that part of the object is hot, or that it is being given to someone else at some distance and orientation from the subject. Two more images were collected just as in the initial conditions with the subjects holding the object and then having the same grasp but without the object. This produces a quad where the first pair of the quad is likely a more intuitive grasp while the second is likely a more special-purpose grasp.

The dataset consists of 10 quads for each object collected in 10 different locations by approximately two dozen subjects, although the dataset is not balanced with respect to the identity of the subject. However, it is balanced with respect to the locations, an equal number of images having been shot in each.

Fig. 2. The 148 objects imaged as part of this dataset. These are common objects found in households and offices with a wide range of purposes.

Within a location, we intentionally did not specify a position or viewpoint for the camera leading to more varied backgrounds but multiple images were shot from the same viewpoint in a location. Locations are areas such as hallways, rooms, or outdoor patios. Examples from the dataset are shown in fig. 3.

Images were collected using portable tripods and subjects' cellphone cameras. Subjects were allowed to use both hands when grasping and were not instructed about the space of allowable grasps. Our intention is to collect images of as a varied set of grasps as possible. Additionally, subjects were chosen such that a variety of skin tones is represented, although we did not balance with respect to skin tone. Most existing hand datasets feature almost exclusively light skin tones which both biases learning and the evaluation of hand pose recognizers. We de-identified subjects by using a face detector and coarsely pixelating their faces. Faces almost never overlapped with grasps.

Since our dataset features the hand poses unobstructed by the object and due to the fairly good performance of existing hand-pose recognizers in such favorable conditions, approximate ground truth exists for every image; we merely copy over the annotation from the unoccluded hand to the occluded hand. This allows us to validate the accuracy of the hand-pose recognizers on all images in the dataset even without any human annotation. Yet this would not allow us to understand how well humans perform on this task of reconstructing interacting and partially-occluded hands. To do so subjects annotated 400 images, 200 pairs of unoccluded and occluded images, with anatomical 21-keypoints — 4 for each finger and one at the wrist. Multiple annotators provided judgments for each image, with four annotations per image. We use this to compute human agreement on partially-occluded hand annotations. With our unique design that results in pairs of images of the same hand occluded and unoccluded images, we can test the robustness of human perception of partially-occluded hands. These two experiments help characterize human performance on hand-pose recognition and have not been performed before. Finally, we can also use human annotations to verify the performance of machines where we show that on unoccluded hands, performance is quite good and then quickly decreases with any occlusion.

The methodology described above results in a novel dataset with both the occluded and unoccluded hands in the same poses, which is balanced with respect to the object instances, while encouraging a varied set of grasps.

4 Experiments

After discussing the statistics of the dataset, in section 4.1, we evaluate the performance of state-of-the-art computer vision systems on our dataset, in section 4.2, provide the first account for human hand-pose recognition performance, in section 4.3, and then compare machines against humans putting the state of computer vision into context, in section 4.4. Two recently-published single-image hand-pose recognition systems are evaluated: that of Zimmermann and Brox [6], which we refer to as ZB, and OpenPose [7]. We extensively surveyed the past three years of major computer vision conferences and these were the only two

Fig. 3. Eight examples from the dataset. Each row is a quad, a series of four images taken in quick succession. In the first two images, the leftmost two, the subject uses their default grasp of the object. In the next two, the rightmost two, subjects are asked to choose another unspecified grasp. Each pair of images has one image in which the hand is holding the object followed by another in which the hand is in the same pose but without the object.

such systems with publicly available runnable source code and pretrained models. Note that we do not fine-tune on this dataset, much as many other systems do not fine tune on the Dexter datasets, and to discourage fine-tuning on this data we do not provide a training and test split.

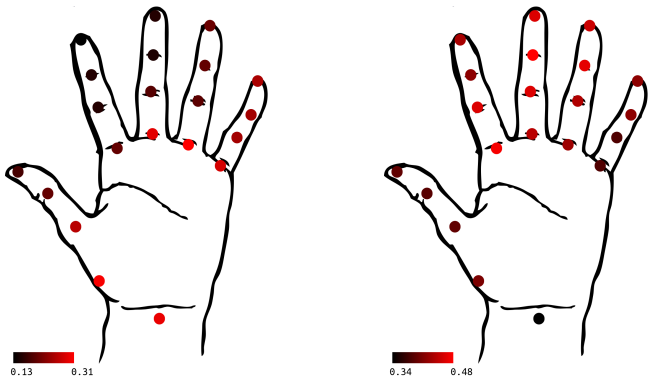
4.1 Dataset statistics

The dataset consists of 11,840 images of the 148 object instances shown in fig. 2 with 80 images per instance half of which are of grasps holding an object and the other half are of those same grasps without holding the object. The dataset is balanced with respect to the object identity. Deformable objects account for 15% of the data. Hands were annotated with 21 keypoints over 400 images which were annotated 4 times over to measure human inter-coder agreement. The likelihood of any one keypoint being under self-occlusion, i.e., being occluded in the nominally unoccluded view, is 21.9% and is fairly uniform, shown in fig. 4(a), as is the likelihood of a keypoint being occluded by the object which is 42.3%, shown in fig. 4(b). We did not control for the distribution over occlusions per keypoint, as there is no practical way to do so; we merely report the statistics of the grasps that humans employ. While high-resolution images will be provided, on average 4065×2972, for consistency all numbers reported here use the same image dimensions as the Dexter datasets [8–10], namely 640×480. Note that both systems were run at their full native resolution with the full image; we merely report distances rescaled to this standard resolution. The average area of a hand in the dataset was approximately 2400 pixels, roughly 48x48, as computed by a bounding box encompassing the extremities of the hand, while the average size of an object is approximately 2200 pixels, roughly 47x47, about the same size as the hand.

4.2 Machine performance

First we evaluate how well machines perform against their own judgments. Given that we have both an occluded and unoccluded image, we can directly ask, assuming that the reconstructed pose in the unoccluded images are correct: how much does occlusion degrade machine performance? We find that occlusion is a major contributor to poor machine performance, i.e., it changes how machines interpret hands, leading to a 20 and 75 average pixel Euclidean distance for OpenPose and ZB respectively. The PCK, the probability that a keypoint is within a given threshold distance in the occluded image relative to the unoccluded one, is shown in fig. 5(a).

The overall distribution of distances for each keypoint is shown in fig. 5(b). Error on visible keypoints is around 15 pixels while many of the occluded keypoints are never identified, representing the peak at 48 pixels. Since the average hand is about 48x48 pixels we fix the maximum penalty for not finding a hand to this dimension — not scoring these would lead to perfect performance while penalizing them the entire length of the image is arbitrary. It is telling that in any one image roughly half of the keypoints are occluded. At first it may seem



(a) (b)

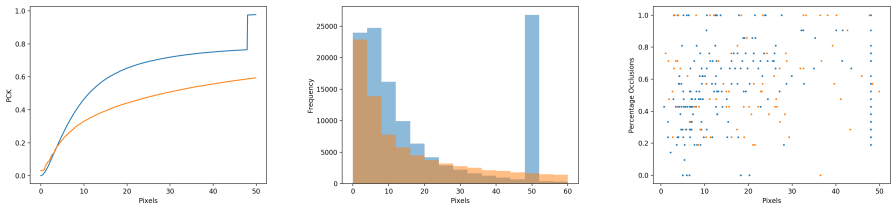
Fig. 4. The 21 annotated keypoints. (a) The likelihood that a keypoint is under self-occlusion by the hand. (b) The likelihood that a keypoint is occluded by an object. The likelihood that any one keypoint is occluded is fairly uniform regardless of the source of the occlusion. For any one grasp, 21.9% of the keypoints are occluded by the hand and 42.3% are occluded by the object. Note that the latter usually also includes the points from the former meaning the two sources of occlusion are not mutually-exclusive.

like ZB has lower error from fig. 5(b), but note that the tail is extremely long. ZB makes make confident but wrong predictions for hands which are spurious. We did not cap the maximum error of either OpenPose or ZB, but had we capped both, ZB would have a mean error closer to 40.

To investigate the source of errors further, in fig. 5(c) we show pixel-wise error as a function of the occlusion of each keypoint in each image. We restricted the plot to within 50 pixels of the correct value after which the data is fairly uniform. Again the line at 48 represents the pixels which could not be detected. There are far more of these failed detections at higher occlusion levels but the plot hides this information. OpenPose seems significantly more resilient to occlusion although the trend where additional occlusion worsens results is clear. Whether a keypoint was occluded was determined by humans, but otherwise we do not use any human annotations in this experiment. Overall the robustness of machines to occlusion is quite poor; we will return to this in section 4.4 when machines are compared to humans.

4.3 Human performance

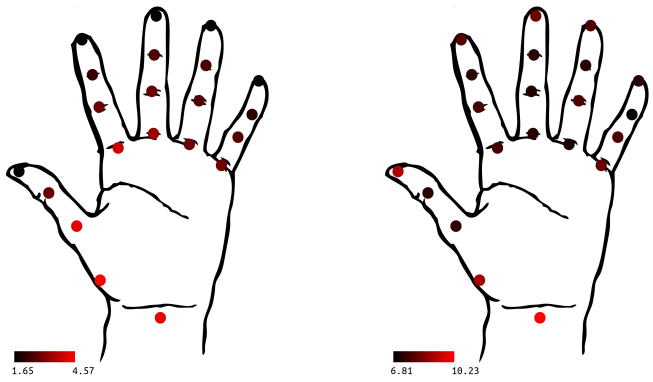
We provide the first quantitative measurements of humans on hand-pose reconstruction. On 400 images, 200 pairs of occluded and unoccluded images, we collected 4 annotations. One was collected in-house while three were collected on Mechanical Turk. Overall humans agree strongly with each other having a mean distance of 5.3 pixels and standard deviation of 1.7. In fig. 6, we show the inter-annotator agreement by keypoint separating (a) unoccluded form (b) occluded keypoints. Agreement is far higher on unoccluded keypoints with a



(a) (b) (c)

Fig. 5. Performance for OpenPose, shown in blue, and ZB, shown in red. (a) The distance between keypoints in the occluded hand and those in the same hand pose unoccluded by any object. (b) The distribution of the errors by distance for each keypoint. (c) The error as a function of the occlusion of the entire hand up to 50 pixels.

mean error of just 2.7 pixels (variance 1.0) while it is significantly lower on occluded keypoints with a mean of 7.8 (variance 2.5). Performance depends on the keypoint in part because some keypoints are more likely to be occluded than others and in part because some keypoints, particularly the wrist, are less well defined anatomically. This indicates that humans might be much more robust to occlusion since they agree with each other, although at the same time all humans might share the same systematic biases.

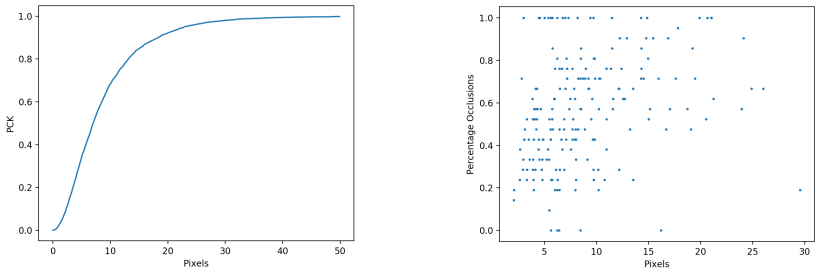


(a) (b)

Fig. 6. The inter-annotator agreement by hand keypoint showing that some keypoints have higher agreement than others. Overall agreement is far higher than in the machine case. (a) shows only the unoccluded keypoints while (b) shows only the occluded keypoints. The color encodes the mean pixel error for that keypoint across four annotators.

To investigate human performance, rather than just agreement, we use the occluded and unoccluded images of the same hand pose and compare human performance between the two. In essence this asks: how accurate are humans

when reconstructing occluded points? In fig. 7(a) we show human PCK and (b) distance as a function of occlusion. Humans are very accurate, robust to occlusion, and perform this task well leading to high agreement.



(a) (b)

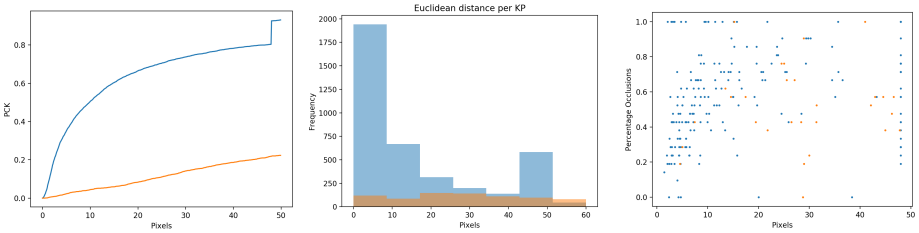
Fig. 7. The performance of humans when reconstructing partially-occluded hands showing (a) PCK and the (b) error as a function of occlusion.

4.4 Machine vs. Human performance

Finally, having established how well humans perform at hand-pose recognition, we compare machines to humans. We take the mean human annotation to be the gold standard and test OpenPose and ZB against it. In fig. 8(a) we show PCK and in (b) we show the distribution of distances for each keypoint. We report these numbers only for the occluded hand in each pair. Average keypoint error on this dataset is far higher than it is on other datasets, being roughly 20 pixels while it is on the order of 5 pixels in the Dexter datasets. This performance difference is further accentuated since our hands are smaller in the field of view, as can be seen in fig. 1. This high distance and distribution are explained by the fact that OpenPose fails to identify the occluded hand at all in many images, leading to a bimodal distribution with a peak near the correct hand and another extremely high variance component that is essentially uniform. Since occluded keypoints are likely harder to infer than unoccluded keypoints, in (c) we show the performance as a function of the percentage occlusion of the hand. This confirms the earlier observation that occlusion is indeed the major contributor to poor performance.

5 Discussion

We describe a challenging new dataset for single-image hand pose estimation, made difficult by the fact that the natural grasps that humans use tend to occlude the hand, and make that data available. On other datasets per-keypoint error is around 4 to 5 pixels while on this dataset it is roughly 20 pixels with the mean



(a) (b) (c)

Fig. 8. Performance of OpenPose, shown in blue, and ZB, shown in red against human annotations. (a) PCK for the partially-occluded hands. (b) The distribution of the errors by distance. (c) The pixel error plotted against the percentage occlusion of the hand.

hand dimensions being 48x48 pixels when the images are rescaled to 640x480. The per-keypoint error is roughly half the size of a hand. Hand-pose recognizers do not seem to be robust to partial occlusions while in the real world human hands tend to be occluded when hand-object interactions occur. Resilience to partial occlusions is generally not exercised by current datasets, in hand pose recognition or other areas of computer vision. More datasets for object recognition and other tasks where occlusion plays a large role may drive research toward new approaches.

Humans were found to be highly robust to partial occlusions with performance only being weakly related with respect to the percentage of the keypoints which are occluded. We are following up with experiments to understand how much viewing time is required for this robustness to manifest in humans — a long processing time may indicate the necessity for more than feed-forward networks. We hope this dataset will lead to new approaches to robust hand-pose recognition given the importance of this task for action recognition and human-robot interactions.

References

1. Presti, L.L., La Cascia, M.: 3d skeleton-based human action classification: A survey. Pattern Recognition **53** (2016) 130–147
2. Perez-Sala, X., Escalera, S., Angulo, C., Gonzalez, J.: A survey on model based approaches for 2d and 3d visual human pose recovery. Sensors **14** (2014) 4189–4210
3. Wei, S.E., Ramakrishna, V., Kanade, T., Sheikh, Y.: Convolutional pose machines. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2016) 4724–4732
4. Cao, Z., Simon, T., Wei, S.E., Sheikh, Y.: Realtime multi-person 2d pose estimation using part affinity fields. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 7291–7299
5. Papandreou, G., Zhu, T., Kanazawa, N., Toshev, A., Tompson, J., Bregler, C., Murphy, K.: Towards accurate multi-person pose estimation in the wild. In: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE (2017) 3711–3719

6. Zimmermann, C., Brox, T.: Learning to estimate 3d hand pose from single RGB images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 4903–4911
7. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. (2017) 1145–1153
8. Sridhar, S., Oulasvirta, A., Theobalt, C.: Interactive markerless articulated hand motion tracking using RGB and depth data. In: Proceedings of the IEEE international conference on computer vision. (2013) 2456–2463
9. Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., Theobalt, C.: Real-time hand tracking under occlusion from an egocentric RGB-D sensor. In: Proceedings of International Conference on Computer Vision (ICCV). (2017)
10. Sridhar, S., Mueller, F., Zollhöfer, M., Casas, D., Oulasvirta, A., Theobalt, C.: Real-time joint tracking of a hand manipulating an object from RGB-D input. In: European Conference on Computer Vision, Springer (2016) 294–310
11. Tompson, J., Stein, M., Lecun, Y., Perlin, K.: Real-time continuous pose recovery of human hands using convolutional networks. ACM Transactions on Graphics **33** (2014)
12. Huang, Y., Bianchi, M., Liarokapis, M., Sun, Y.: Recent data sets on object manipulation: A survey. Big data **4** (2016) 197–216
13. Andriluka, M., Pishchulin, L., Gehler, P., Schiele, B.: 2d human pose estimation: New benchmark and state of the art analysis. In: Proceedings of the IEEE Conference on computer Vision and Pattern Recognition. (2014) 3686–3693
14. Bullock, I.M., Feix, T., Dollar, A.M.: The yale human grasping dataset: Grasp, object, and task data in household and machine shop environments. The International Journal of Robotics Research **34** (2015) 251–255
15. Berzak, Y., Huang, Y., Barbu, A., Korhonen, A., Katz, B.: Anchoring and agreement in syntactic annotations. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. (2016) 2215–2224
16. Santello, M., Bianchi, M., Gabiccini, M., Ricciardi, E., Salvietti, G., Prattichizzo, D., Ernst, M., Moscatelli, A., Jörntell, H., Kappers, A.M., et al.: Hand synergies: integration of robotics and neuroscience for understanding the control of biological and artificial hands. Physics of life reviews **17** (2016) 1–23
17. Bohg, J., Morales, A., Asfour, T., Kragic, D.: Data-driven grasp synthesis—a survey. IEEE Transactions on Robotics **30** (2014) 289–309
18. Levine, S., Pastor, P., Krizhevsky, A., Ibarz, J., Quillen, D.: Learning hand-eye coordination for robotic grasping with deep learning and large-scale data collection. The International Journal of Robotics Research **37** (2018) 421–436
19. Goldfeder, C., Ciocarlie, M., Dang, H., Allen, P.K.: The columbia grasp database. In: Robotics and Automation, 2009. ICRA'09. IEEE International Conference on, IEEE (2009) 1710–1716
20. Chebotar, Y., Hausman, K., Su, Z., Molchanov, A., Kroemer, O., Sukhatme, G., Schaal, S.: Bigs: Biotac grasp stability dataset. In: ICRA 2016 Workshop on Grasping and Manipulation Datasets. (2016)