Big Hand 2.2M Benchmark: Hand Pose Data Set and State of the Art Analysis

Shanxin Yuan[1], Qi Ye[1], Björn Stenger[2], Siddhant Jain[3], Tae-Kyun Kim[1]

[1]Imperial College London [2]Rakuten Institute of Technology [3]IIT Jodhpur

Abstract

In this paper we introduce a new large-scale hand pose dataset collected using a novel capture method. Existing datasets are either synthetic or real: the synthetic datasets exhibit a certain level of appearance difference from real data, and the real datasets are limited in a quantity and coverage, mainly due to the difficulty of annotations. A magnetic tracking system with six magnetic 6D sensors and inverse kinematics on a hand model are proposed to automatically obtain 21-joints hand pose annotations of real data, and in real-time with minimal restriction of the range of motion. The new dataset collected with a designed protocol attempts to cover all of the natural hand pose space. As shown in the embedding plots, the new dataset shows by far the widest and the most dense range of hand poses compared to the existing benchmarks. Current state-of-the-art methods are evaluated using the new dataset, and we demonstrate significant improvements in cross-benchmark evaluations. We also show significant improvements in egocentric hand pose estimation by training on the new dataset.

1. Introduction

The area of hand pose estimation has made significant progress in the recent past and a number of working systems have been proposed [2, 4, 7, 10, 11, 15, 22, 28]. As has been noted in [8], existing benchmarks [26, 30, 22, 15, 24, 32], however, are restricted in terms of the number of frames (mainly due to the difficulty of annotation), annotation accuracy, hand shape and viewpoint variations, and articulation coverage.

The current state-of-the-art for hand pose estimation employs deep neural networks to estimate hand pose from input data [30, 36]. It has been shown that these methods scale well with the size of the training data set without over fitting. The availability of a large-scale, accurately annotated data set is therefore a key factor for advancing the field. Manual annotation has been the bottleneck for creating large-scale benchmarks [15]. This method is not only labor-intensive, but can also result in inaccurate position labels. Semi-automatic capture methods have been devised where 3D joint locations are inferred from manually annotated 2D joint locations [17, 8]. Alternatives, which are still time-consuming, aim to track a hand model and manually refine the results, if necessary iterate these two steps [24, 26, 30]. Additional sensors can aid automatic capture significantly, but care must be taken not to restrict the range of motion, for example when using a dataglove [34]. More recently, less intrusive magnetic sensors have been employed for finger tip annotation in the *HandNet* data set [32] .

In this paper, we introduce our million-scale *Big Hand* data set that makes significant advancement in terms of completeness of hand data variations and quality of full annotations, see Figure 1 and Table 1. We detail the capture set-up and methodology that enables efficient hand pose capture with high accuracy. This enables us to capture the range of hand motions that can be adopted wihtout external forces. Our data set contains 2.2 million depth maps with accurately annotated joint locations. The data is captured by attaching six magnetic sensors on the hand, five on each finger nail and one on the back of the hand, where each sensor provides accurate 6D measurements. Locations of all joints are obtained by applying inverse kinematics on a hand model with 31 degrees of freedom (dof) with kinematic constraints. The *Big Hand* data set contains 290,000 frames of egocentric hand poses, which is 130 times larger than the currently largest egocentric hand pose data set so far. Training a Convolutional Neural Network (CNN) on the data shows significantly improved results.

The recent study by Supancic *et al.* on cross-benchmark testing showed that around 40% of poses are estimated with an error larger than 50mm. This is due to different capture set-up, hand shape variation, and annotation schemes. Training a CNN using the *Big Hand* dataset, we demonstrate state-of-the-art performance on existing benchmarks.

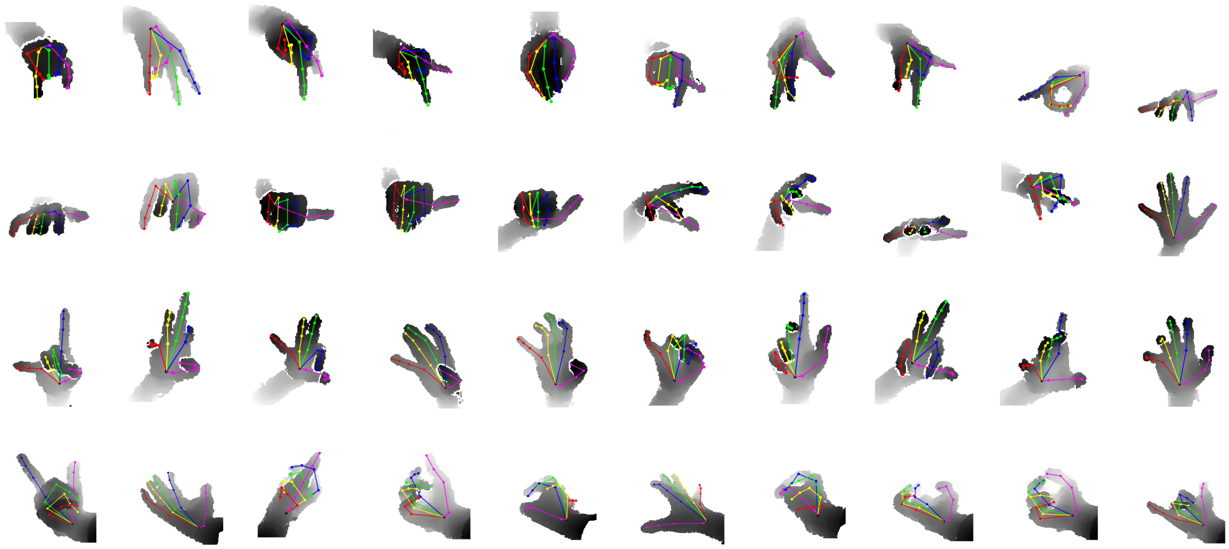

Figure 1. Examples images from the Big Hand data set. Our benchmark covers the pose space of a human hand that can be covered without external forces. This data set is the state of the art in terms of pose coverage and quality of full annotations.

Dataset	Annotation	No. frames	No. joints	No. subjects	View point	Resolution
Dexter1 [23]	manual	3,157	5	1	3rd	-
MSRA14 [15]	manual	2,400	21	6	3rd	320×240
ICVL [26]	track + refine	17,604	16	10	3rd	320×240
NYU [30]	track + refine	81,009	36	2	3rd	640×480
MSRA15 [24]	track + refine	76,375	21	9	3rd	320×240
UCI-EGO [17]	semi-auto	400	26	2	ego	320×240
Graz16 [8]	semi-auto	2,166	21	6	ego	320×240
ASTAR [34]	automatic	870	20	30	3rd	320×240
HandNet [32]	automatic	212,928	6	10	3rd	320×240
MSRC [22]	synthetic	102,000	22	1	3rd	512×424
BigHand	automatic	2.2M	21	10	full	640×480

Table 1. Benchmark Comparison. Existing benchmarks are limited in the number of frames, which is restricted by annotation methods. Our automatic annotation method allows us to collect unlimited number of fully annotated depth images. Our benchmark are collected with the latest Intel RealSense SR300 camera [1], which can produce high quality depth images of high resolution.

2. Existing Benchmarks

Despite the intensive efforts in the field [11, 15, 22, 10, 7, 28, 2, 28, 2, 4], a few existing benchmarks for evaluation and comparison [26, 30, 22, 15, 24, 32] are significantly limited in scale (from a few hundred to tens of thousand), annotation accuracy, articulation, view point, and hand shape.

The bottleneck for building a large scale real benchmark is the lack of a fast and accurate annotation method. Manual annotation has been applied early to build small benchmarks [15, 23], but it's labor-intensive and can result in inaccurate annotations. These benchmarks are small in size, e.g. MSRA14[15] and Dexter1 [23] has only 2,400 and 3,157 frames, making them not suitable for training algorithms and only suited to evaluating model based hand tracking methods.

Alternative annotation method, which are still labor-intensive and time-consuming, aim to track a hand model and manually refine the results, if necessary they have to iterate these two steps [24, 26, 30]. ICVL benchmark [26] is a small and simple benchmark, which is firstly annotated using 3D skeletal tracking method [5] and then manually refined. Despite the manual efforts, the annotation accuracy is compromised, and its respective issues have been noted in literature [24, 8]. NYU benchmark [30] is much larger and has a larger range of view points. Its annotations were done by a model based hand tracking on depth images from three cameras. Particle Swarm Optimization is used to find out the final annotation. This method often drifts to wrong poses, where manual correction is needed to restart the tracking process. MSRA15 benchmark [24] is currently the largest and most complex [8], it is annotated in an iterative way, where an optimization method [15] and manual re-adjustment alternate till a convergence. The annotation yet contains a high level of errors, such as missing annotations on certain fingers (especially the thumb). This benchmark has a large view point coverage, but it has small variations in articulation. It captures 17 base articulations and each of them varies little in a 500-frame sequence.

Semi-automatic annotation were pursued but only small benchmarks were produced [17, 8]. UCI-EGO benchmark [17] was annotated by iteratively searching for the closest synthetic example in synthetic set and manually refining. Graz16 benchmark [8] was annotated by iteratively annotating visible joints in some key frames and automaticly inferring the whole sequence using optimization method, where the appearance, temporal, and distances constraints are exploit. But it is difficult to annotate fast moving hands. It also requires manual corrections when the inference fails. This semi-automatic method offerred a 2,000 frame egocentric benchmark successfully annotated, however, not sized enough to train hand pose estimation algorithms.

Additional sensors can aid automatic capture significantly [34, 32, 31, 14], but care must be taken not to restrict the range of motion. ASTAR benchmark [34] used a data-glove called ShapeHand [21], but wearing the glove significantly distorts captured hand images, and hinders free hand movements. In the works of [31, 14], the human body pose were treated as a state estimation problem given magnetic sensor and depth data. More recently, less intrusive magnetic sensors have been used for finger tip annotation in HandNet benchmark [32], which exploits a similar annotation setting as our benchmark with trakSTAR magnetic sensors [6]. However, this benchmark only provides fingertips not the full annotations, used for evaluting fingertip detection methods.

Synthetic data has been exploit to do training [16, 18, 35], or both training and testing [22]. Even though we can get un-limited amount of accurately annotated synthetic data, there is a gap between the synthetic and real data. Apart from differences in hand characteristics and the lack of sensor noise, synthetically generated images tend to produce kinematically implausible hand, see Figure 10. MSRC benchmark [22] is an synthetic benchmark, where data is uniformly distributed in the 3D view point space. However, the data is limited in articulations, which are generated by randomly sampling from six articulations, and there is a significant level of gap between the synthetic data and real hand images.

3. Full Hand Pose Annotation

In this part, we present our method to do accurate full hand pose annotations using the trakSTAR tracking system with 6D magnetic sensors.

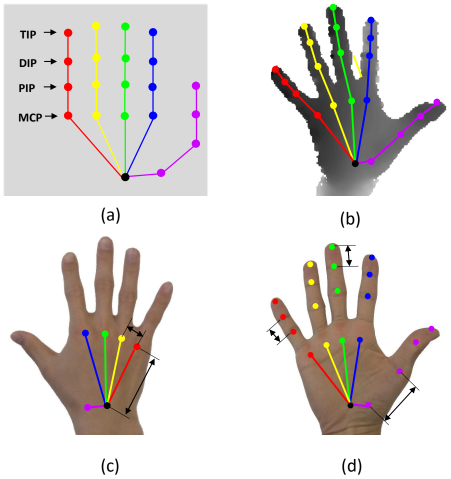

Figure 2. 31-D hand model and hand measurement. (a) Our hand model has 31 Degree of Freedom (DoF), including 6 DoF of global location and orientation for the wrist. Each finger has 5 DoF (flexion for DIP and PIP, flexion, abduction and twist for MCP). (b) 31-D model with hand shape give a 21-joint skeleton model. (c) and (d) shows that we manually measure the hand shape for each person.

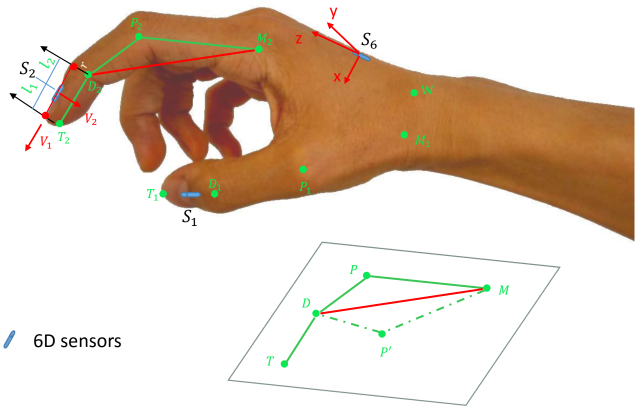

Figure 3. Full hand pose (21 joints) inference using six 6-D magnetic sensors and hand model. Given the location and orientation of sensor S6 and the hand model, the six joints on the palm can be inferred. Each sensor on the nail is used to infer the TIP and DIP joints. Each PIP joint can be calculated using the physical constraints and bones lengths, e.g., P(instead of $P^{'}$) is kept considering PIP and TIP should be on different sides of the line connecting DIP and MCP.

3.1. Annotation by Inverse Kinematics

Given the six magnetic sensors, each with 6D data (locations and orientations), along with a hand model, we use Inverse Kinematic to infer the full hand pose, i.e. the locations of 21 joints. We choose the 21-joints hand model, as shown in Figure 2. The physical constraints we used are as follows: 1) the wrist and 5 MCP joints are relatively fixed,

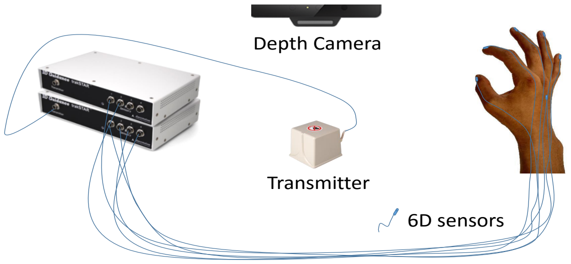

Figure 4. Annotation settings. The equipment used in our annotation system are: Two hardware synchronized electromagnetic tracking units, six 6D magnetic sensors, one "Mid-Range Transmitter", and Intel SR300 camera.

2) bone lengths are kept, and 3) MCP, PIP, DIP, and TIP for each finger are located on the same plane.

Similar to [20], as shown in Figure 3, five magnetic sensors (from thumb to pinky, the sensors are S1, S2, S3, S4, S5) are attached on the five fingers' tips, the sixth one (S6) is attached on the back of the palm. Given the location and orientation of S6, as well as the hand model, the wrist (W) and five MCPs (M1, M2, M3, M4, M5) are inferred. For each finger, given the sensor's location and orientation, the TIP and DIP are calculated in the following way (as shown in Figure 3, take the Index finger as an example): sensor's orientation is used to find the three orthogonal axes, V_1 is along the finger, V_2 is pointing forward from the finger tip. TIP's location (T) and DIP's location (D) are:

$$T = L(S) + l_1 * V_1 + r * V_2 \quad (1)$$

$$D = L(S) - l_2 * V_1 + r * V_2 \quad (2)$$

where b is the bone length connecting DIP and TIP, $L(S)$ denotes the sensor location, and r half of the finger thickness. l_1+l_2 = b.

The last joint to infer is the PIP, whose location P, as shown in Figure 3, is calcuated using the following conditions: (1) T, M, D are given, (2) $\|P - D\|$ and $\|P - M\|$ are fixed, (3) T, D, P, M are on the same plane, and (4) T and P should be on different sides of the line connecting M and D. There exits only one solution for P, that meets all the constraints.

The full hand pose annotation is therefore inferred using the six sensor locations and orientations, as well as the predefined hand model, which is directly measured from each user's hand.

3.2. Time Synchronization and Coordinate Calibration

To build and annotate our benchmark, we use a trakSTAR tracking system [6] combined with an Intel Realsense SR300 camera [1], which is the latest version depth camera using Fast VGA technology and with high resolution. See Figure 4, we used a trakSTAR tracking system [6] with two hardware synchronized electromagnetic tracking units, each of which can track at most four 6D magnetic sensors. The 6D sensor we used is called "Model 180" and is 2mm wide with a flexible 1.2mm wide and 3.3m long cable. When the cable is attached to the hand using tight elastic loops the depth profile and hand movements are not affected. The transmitter is the "Mid-Range Transmitter" that has a maximum tracking distance of 660mm, which is suitable for hand pose tracking. The tracking system is stable and without drift in continuous operation.

The trakSTAR tracking system captures the locations and orientations of the six magnetic sensors at a speed of 720 fps. The depth camera captures images with a resolution of 640*480 and runs at a maximum speed of 60 fps. Since the trakSTAR tracking system and the depth camera are at different frame rates, we do synchronization by finding the nearest neighboring time stamps. The time gap between the depth image and the magnetic sensors in this way is 0.7 millisecond at most.

TrakSTAR tracking system and Intel Realsense SR300 camera have their own coordinate systems. We used ASPNP [37] to calibrate the coordinates as in [32]. Given a set of 3D locations of the magnetic sensors in the trakSTAR tracking system and the corresponding 2D locations captured by Intel Realsense camera as well as its intrinsic camera parameters, ASPNP algorithm establishes the transformation between these two coordinate systems.

4. New Big Hand Benchmark

We collected the *Big Hand* data set containing 2.2 million depth images of a single hand with joints automatically annotated (see Section 3). Ten subjects (7 male, 3 female) were captured for two hours each. We capture 31 dimensions in total, 6 dimensions for global pose and 25 articulation parameters, represented in the angle space. Each finger's pose is represented by five angles, including the twist angle, flexion angle, abduction angle for the MCP joint and flexion angles for the DIP and PIP joints. Similar to [33], we defined *extremal poses* as hand poses where each finger assumes a maximally bent or extended position, there are 32 such poses. For maximum coverage of the natural articulation space, we enumerate all $\binom{32}{2}$ = 496 possible pairs of these *extremal poses*, and capture the natural motion when transitioning between the two poses of each pair. In total the *Big Hand* data set consists of three parts: (1) Schemed poses: to cover all the articulations that a human hand can freely adopt, this contains has 1.534 million frames, captured as described above. (2) Random poses: 375K frames are captured with participants being encouraged to fully explore the pose space. (3) Egocentric poses: 290K frames of egocentric poses are captured with subjects carrying out the

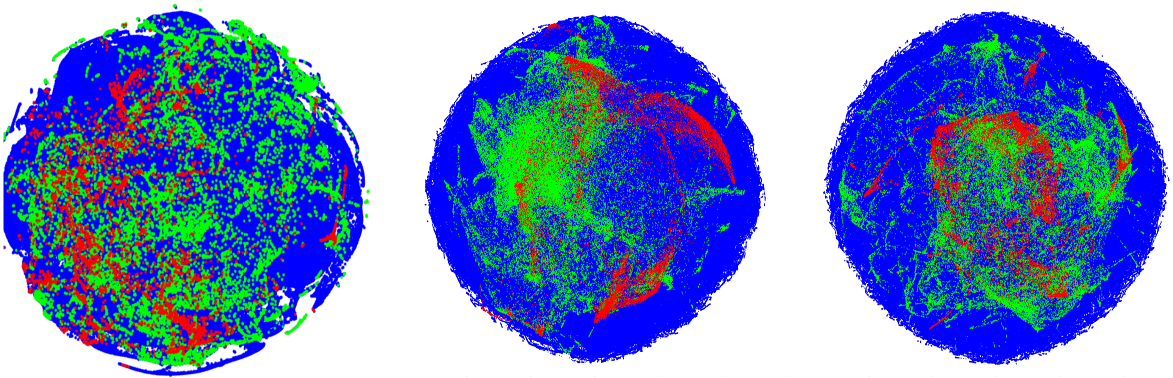

Figure 5. 2D t-SNE embedding of the hand pose space. Big Hand is represented by blue dots, ICVL is represented by red dots. NYU is represented by green dots. The figures show (left) global view point space coverage, (middle) articulation angle space (25D), and (right) hand angle (global orientation and articulation angles) coverage comparison. Compared with existing benchmarks, the *Big Hand* contains a wider range of variation.

Benchmarks	Rogez [19]	Oberweger [8]	*Big Hand* Egocentric
No. Frames	400	2166	290K

Table 2. Egocentric Benchmark size comparison. The egocentric subset of *Big Hand* dataset is 130 time larger than the next largest available dataset.

32 extremal poses combined with random movements.

4.1. Hand Articulation Space

In order to cover all view points, we introduce variation by changing the sensor height, the subject's position and arm orientation. The view point space is divided into 16 regions, and subjects are instructed to carry out random view point change within each region.

As the t-SNE visualization in Figure 5 shows, our benchmark data covers a significantly larger region of the pose articulation space than the public ICVL, NYU data sets.

4.2. Hand Shape Space

We select ten participants with different hand shapes. Of the 10 participants, 3 of them are females, 7 are males. All of them are between 25 and 35 years old.

Existing benchmarks also tried to involve different hand shapes, but are limited in their annotation methods. MSRC [22] synthetic benchmark has only one hand shape. ICVL [26] select ten different participants with varying hand sizes, but these ten person have very similar hand shapes, and they are annotated in the same one hand model. NYU [30] training data has one hand shape, its testing data has two hand shapes including one shape from the training set. MSRA15 has nine participants, but in the annotated ground truth, only three hand shapes are used. see Figure 6

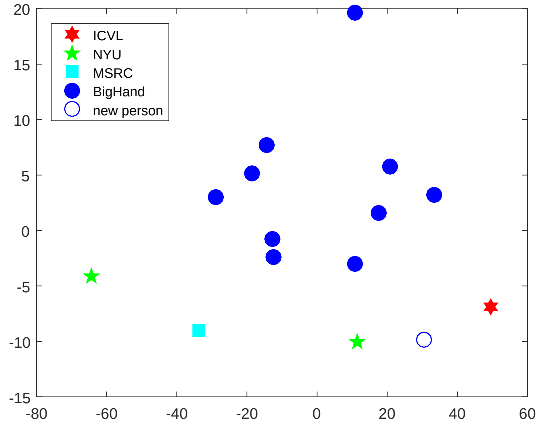

Figure 6. Hand shape variation. The *Big Hand* data set contains 10 hand shapes, and an additional subject's hand, used for testing baselines. ICVL contains one hand shape even though it has ten different persons, NYU has two hand shapes, MSRC has one synthetic hand shape. This figure is obtained by applying PCA to the five distances from wrist to finger tips.

5. State of the art analysis

We adopt the *Holi* CNN [36] as a representative of the current state of the art. The detailed structure is shown in the supplementary material. The input for the CNN model is acquired by cropping hand area using ground truth joint locations. The cropped hand is normalized to 96 × 96 pixels. The normalized image with its two downsampled images of size 48 × 48 and 24 × 24 is fed into the CNN. The cost function is the mean squared distance between the location estimates and the ground truth locations.

The CNN model is implemented using Theano [29] and is trained on a workstation with a Nvidia GeForce GTX TITAN Black and a 32 core Intel processor. The whole

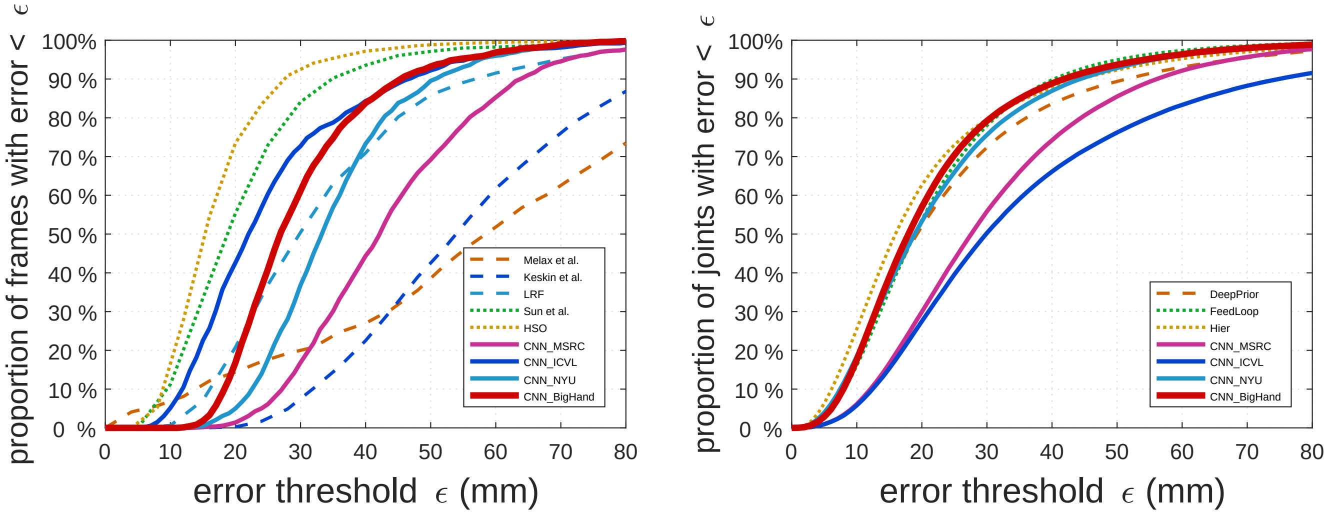

Figure 7. Cross benchmark performances. CNN models are trained on ICVL, NYU, MSRC, and our Big Hand benchmark. Cross benchmark evaluation is performed among the four benchmarks, the CNN model trained on Big Hand can achieve state-of-the-art performance on ICVL and NYU, while these CNNs trained on ICVL, NYU, and MSRC can not generalize well to other benchmarks. The left and right figure show the testing results on ICVL and NYU testing data. "CNN_MSRC", "CNN_ICVL", "CNN_NYU", and "CNN_BigHand" are CNNs trained on the training data of MSRC, ICVL, NYU, and Big Hand, respectively.

test train	ICVL	NYU	MSRC	Bighand
ICVL	**12.3**	35.1	65.8	46.3
NYU	20.1	**21.4**	64.1	49.6
MSRC	25.3	30.8	**21.3**	49.7
BigHand	**14.9**	**20.6**	**43.7**	**17.1**

Table 3. Cross Benchmark comparison. Cross-benchmark average errors, trained with the *Big Hand* data set, the model performs well on ICVL and NYU, while training on ICVL, NYU, and MSRC does not generalize well to other benchmarks.

BigHand dataset is split into a training set and a validation set by a 9:1 ratio. The model is trained using Adam [3], with β_1 being 0.9, β_2 being 0.999 and α being 0.0003. We stop the training process when the cost of the validation set reaches the minimum, which takes 30 epochs and each training epoch takes about 40 minutes. When training the CNN model on smaller BigHand subsets in Section 5.1 and Section 5.3, ICVL, NYU and MSRC dataset, we keep the CNN structure and β_1, β_2, α of Adam unchanged.

Along with our training data in Big Hand, a challenging testing sequence of 37K frames of a previously unseen person was recorded and automatically annotated, see "new person" in Figure 6.

5.1. Cross-benchmark Performance

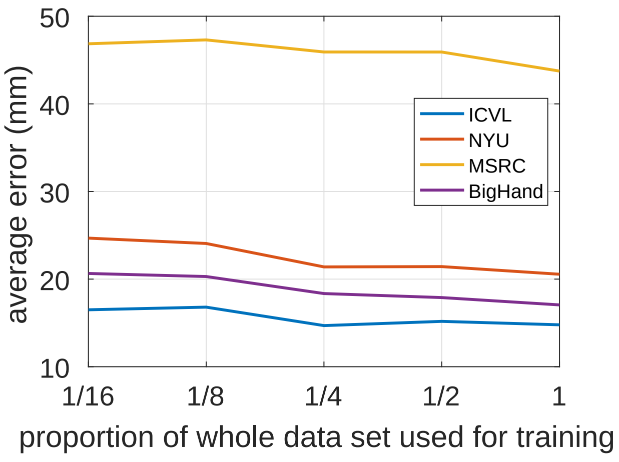

Figure 8. Data size effect on cross benchmark evaluation. When the CNN model is trained on $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$, $\frac{1}{2}$, and all of the benchmark data, the testing results on ICVL, NYU, MSRC, and BigHand keep improving.

Cross-benchmark evaluation is a challenging and largely ignored problem in many fields, like face recognition [13] and hand pose estimation [25]. Due to the small number of training data, existing hand pose estimation systems rely on training data and performed poorly when tested on new unseen hand poses. As pointed out in [25], in existing datasets, "test poses remarkably resemble the training poses", and they proposed "a simple nearest-neighbor base line that outperforms most existing systems".

Table 3 and Figure 7 show that the estimation errors of the CNN models trained on ICVL, NYU, MSRC and BigHand when cross-tested. The performance of testing the CNN model trained on the *Big Hand* training set on other datasets is surprisingly good. On real testing datasets (ICVL

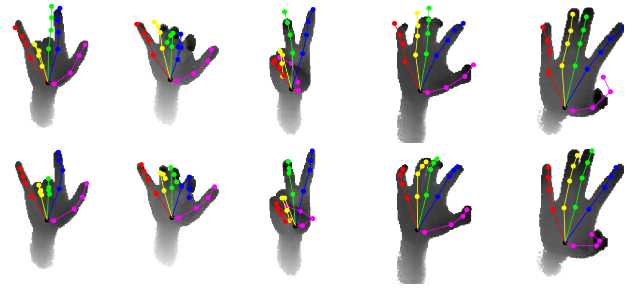

Figure 9. Generalization of the CNN Model Trained on Big Hand. The CNN model trained on our benchmark is able to generalize to existing benchmark ICVL, the estimation is even better than the poorly annotated ground truth. The top row shows the ground truth annotations, while the bottom row shows our estimation results.

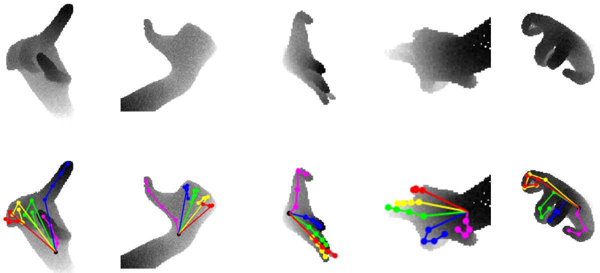

Figure 10. MSRC benchmark examples. Synthetic data lacks real hand shape and sensor noise, and tends to have kinematically implausible hand poses. The top row shows some depth images, the bottom row shows the corresponding groud truth annotation.

and NYU), it achieves comparable or better, performance, with models trained on corresponding training set. This confirms the annotation accuracy, hand shape and viewpoint variations, and articulation coverage of our dataset shown in previous sections and indicates that a CNN model trained on a large scale data set is able to generalize to new hand shapes and view points, while nearest neighbor method showed poor cross-testing performance [25].

The MSRC dataset is a synthetic data set with accurate annotations and the aim to of evenly distributed in viewpoints. When training the CNN on MSRC and testing on all real testing sets, the performance is worse than the CNN trained on NYU, and significantly worse than when trained on *Big Hand*, and is similar to that of the model trained on ICVL which is only one-sixth of the MSRC training set. On the other hand, the model trained on *Big Hand* having a consistently good performance across all real datasets has a bad performance on MSRC testing set. The reason behind this observation is that there is a gap between synthetic data and real data, and appearance gap hinders the cross-testing in both directions. Figure 10 shows some examples of the MSRC dataset. Apart from differences in hand shape and the lack of sensor noise, synthetically generated images tend to produce kinematically implausible hand poses.

Increasing the amount and training data improves the performance on cross benchmark evaluation, see Figure 8. When we train several CNN models with different subsets of Big Hand, and testing them on ICVL, NYU, MSRC, and Big Hand's testing sequence, the performance keeps increasing. These observations confirm that large amount of training data can enable CNNs to generalize to new unseen data.

5.2. State-of-the-art Comparison

In this section, we compared our CNN model trained on Big Hand with 8 state-of-the-art methods including HSO [27], Sun *et al.* [24], Latent Regression Forest (LRF) [26], Keskin *et al.* [2], Melax *et al.* [5], DeepPrior [9], FeedLoop [10], and Hier [36].

When the CNN model trained on *Big Hand* is used for testing on NYU, it outperforms two recent methods, DeepPrior [9] and FeedLoop [10], and achieves comparable accuracy with Hier [36], even though the model has never seen any data from NYU benchmark, demonstrated in the left figure of Figure 7. Since the annotation scheme of NYU is different from ours, we choose a common subset of 11 joint locations for this comparison. We expect better results for consistent annotation schemes.

The ICVL test error curve of the CNN model trained on *Big Hand* is shown in Figure 7. Although it does not appear as good as that on NYU when compared to other methods, it shows better results than most other methods. Note that the estimation error for our CNN model is already as low as 14mm, which means that a small annotation discrepancy between the training and the testing data will have a large influence on the result. As has been noted in [8], the annotation of ICVL is not as accurate as that of NYU. Many frames of our estimation results look plausible but result in larger estimation errors because of inaccurate annotations, see Figure 9 for qualitative comparisons. Another reason is that the hand measurement scheme is different from ours. In our benchmark, each subject's hand shape is recorded by manually measuring joint distances. In ICVL, the same synthetic model is used for all subjects and the MCP joints tend to slide towards the fingers rather than remaining on the physical joints.

5.3. Baselines on Big Hand

Three baselines are evaluated on our challenging 37K-frame testing sequence, the CNN trained on *Big Hand*, the Particle Swarm Optimization method (FORTH) [12] and the method by Intel [1]. The latter two are generative methods.The CNN model significantly outperforms the two generative methods, see the left figure of Figure 11. When we choose a training and validation data number ratio of 9:1, which similar to ICVL, NYU, and HandNet [32] and the validation result achieved significantly good result, with 90% of the joints have error smaller than 5mm, see the mid-

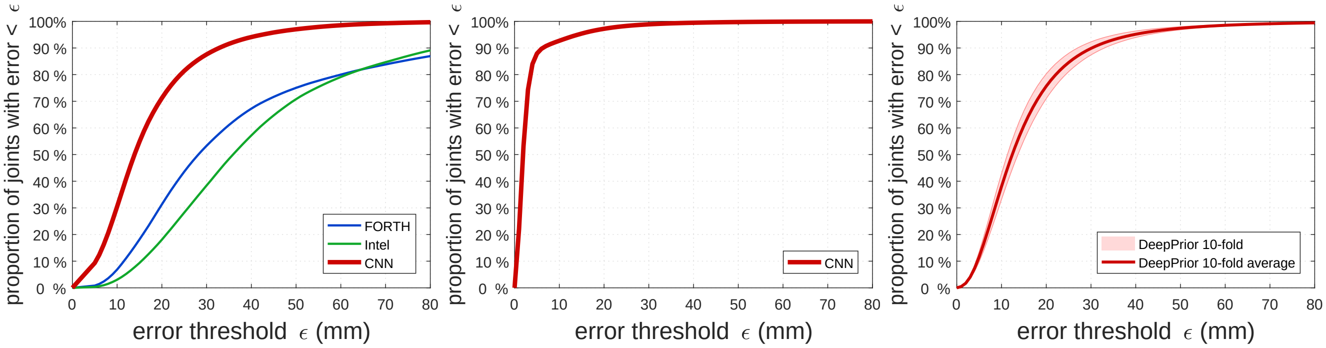

Figure 11. Hand pose estimation performance. The left figure shows the baselines performance on a new person's 37K frames of hand images, learning based CNN model significantly outperformed tracking based methods FORTH [12] and Intel [1]. The middle figure shows that when the CNN model is trained on 90% of our data, it can achieve significantly good estimation accuracy on the rest validation 10% images. The right figure shows that, when DeepPrior [9] model was used to do 10-fold cross–validation for egocentric hand pose estimation. For the first time, we achieved similar accuracy as that of 3rd view hand pose estimation.

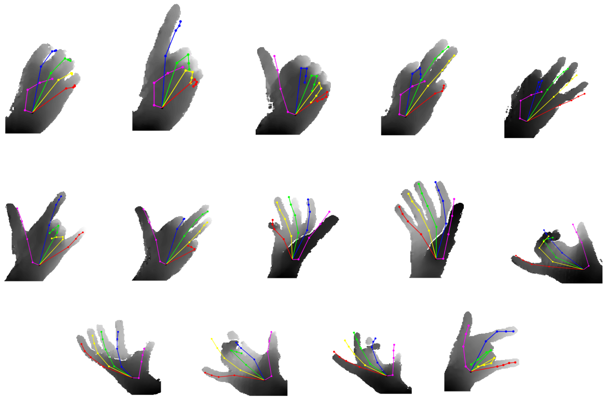

Figure 12. Examples of estimation result on the egocentric data set. A DeepPrior CNN model is trained and achieves state-of-the-art results on egocentric view hand pose estimation.

dle figure of Figure 11.

5.4. Egocentric Benchmark

The availability of a large-scale annotated data set has been a limiting factor for egocentric hand pose estimation. Existing egocentric benchmarks [19, 8] are very small, see Table 2. Rogez *et al.*[19] provided 400 frames and Oberwerger *et al.*[8] provided 2166 frames of fully annotated images. The *Big Hand* egocentric subset contains 290K frames of annotated frames. This data set enabled us to train a DeepPrior CNN model [9] and achieve a significant improvement for egocentric hand pose estimation. For the first time the performance is similar to that of 3rd view hand pose estimation. We perform 10-fold cross validation and obtain a mean error of 15.86mm with a standard deviation of 0.72mm. The right figure of Figure 11 shows the proportion of joints within a certain error threshold ϵ. Figure 12 shows some qualitative results.

6. Discussion and Conclusion

Hand pose estimation has attracted a lot of attention and some high quality working systems are produced, but the development in benchmark still lags behind the algorithm advancement. We propose an automatic full hand pose annotation method and create a million scale thorough and concise benchmark of real hand depth images, and exploit the benchmark for training and evaluating the algorithms, to lead to the next level of pose recognition accuracy.

Our benchmark was collected by a novel automatic annotation method. In our approach, a magnetic tracking system with six magnetic 6D sensors and inverse kinematics with a hand model are used to do full hand pose annotation, leading to real time full accurate annotation. To build a thorough yet concise benchmark, we systematically designed a hand movement scheme to capture all the natural hand poses.

Existing benchmarks are restricted to the number of frames due to difficulty of annotating, annotation accuracy, hand shape variations, viewpoint and articulation coverage, or resort to synthetic data. A large scale accurately annotated real benchmark is still an untamed problem. We introduce our Big Hand Benchmark that makes significant advancement in terms of completeness of hand data variations and quality of full annotations. The established benchmark includes about 290K frames of egocentric benchmark, to facilitate the advancement in egocentric hand pose estimation.

Current state-of-the-art methods are evaluated using the new benchmark, and we demonstrate significant improvements in cross-benchmark evaluations. We also show significant improvements in egocentric hand pose estimation by training on the new benchmark. It is our aim that the data set will help to further advance the research field, allowing the exploration of new CNN architectures or alternative learning approaches.

References

[1] IntelSR300. http://click.intel.com/intelrrealsensetm-developer-kit-featuring-sr300.html. 2, 4, 7, 8

[2] C. Keskin, F. Kıraç, Y. E. Kara, and L. Akarun. Hand pose estimation and hand shape classification using multi-layered randomized decision forests. In *ECCV, 2012*. 1, 2, 7

[3] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[4] H. Liang, J. Yuan, and D. Thalmann. Parsing the hand in depth images. *TMM, 2014*. 1, 2

[5] S. Melax, L. Keselman, and S. Orsten. Dynamics based 3d skeletal hand tracking. In *Proceedings of Graphics Interface 2013*, pages 63–70. Canadian Information Processing Society, 2013. 2, 7

[6] NDItrakSTAR. http://www.ascension-tech.com/products/trakstar-2-drivebay-2/. 3, 4

[7] N. Neverova, C. Wolf, G. W. Taylor, and F. Nebout. Hand segmentation with structured convolutional learning. In *ACCV, 2014*. 1, 2

[8] M. Oberweger, G. Riegler, P. Wohlhart, and V. Lepetit. Efficiently creating 3d training data for fine hand pose estimation. In *CVPR, 2016*. 1, 2, 3, 5, 7, 8

[9] M. Oberweger, P. Wohlhart, and V. Lepetit. Hands deep in deep learning for hand pose estimation. In *CVWW, 2015*. 7, 8

[10] M. Oberweger, P. Wohlhart, and V. Lepetit. Training a feedback loop for hand pose estimation. In *ICCV, 2015*. 1, 2, 7

[11] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BMVC*, 2011. 1, 2

[12] I. Oikonomidis, N. Kyriazis, and A. A. Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *ICCV*, 2011. 7, 8

[13] O. M. Parkhi, A. Vedaldi, and A. Zisserman. Deep face recognition. In *BMVC, 2015*. 6

[14] G. Pons-Moll, A. Baak, J. Gall, L. Leal-Taixe, M. Mueller, H.-P. Seidel, and B. Rosenhahn. Outdoor human motion capture using inverse kinematics and von mises-fisher sampling. In *ICCV, 2011*. 3

[15] C. Qian, X. Sun, Y. Wei, X. Tang, and J. Sun. Realtime and robust hand tracking from depth. In *CVPR, 2014*. 1, 2, 3

[16] G. Riegler, D. Ferstl, M. Rüther, and H. Bischof. A framework for articulated hand pose estimation and evaluation. In *SCIA, 2015*. 3

[17] G. Rogez, J. S. Supancic, and D. Ramanan. First-person pose recognition using egocentric workspaces. In *CVPR, 2015*. 1, 2, 3

[18] G. Rogez, J. S. Supancic, and D. Ramanan. Understanding everyday hands in action from rgb-d images. In *ICCV, 2015*. 3

[19] G. Rogez, J. S. Supancic III, M. Khademi, J. M. M. Montiel, and D. Ramanan. 3d hand pose detection in egocentric rgb-d images. *ECCV, 2014*. 5, 8

[20] S. Schaffelhofer and H. Scherberger. A new method of accurate hand-and arm-tracking for small primates. *Journal of neural engineering*, 9(2):026025, 2012. 4

[21] ShapeHand. http://www.shapehand.com/shapehand.html. 2009. 3

[22] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. K. C. R. I. Leichter, A. V. Y. Wei, D. F. P. K. E. Krupka, A. Fitzgibbon, and S. Izadi. Accurate, robust, and flexible real-time hand tracking. In *CHI, 2015*. 1, 2, 3, 5

[23] S. Sridhar, A. Oulasvirta, and C. Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *ICCV, 2013*. 2

[24] X. Sun, Y. Wei, S. Liang, X. Tang, and J. Sun. Cascaded hand pose regression. In *CVPR, 2015*. 1, 2, 3, 7

[25] J. S. Supancic III, G. Rogez, Y. Yang, J. Shotton, and D. Ramanan. Depth-based hand pose estimation: methods, data, and challenges. *ICCV, 2015*. 6, 7

[26] D. Tang, H. J. Chang, A. Tejani, and T.-K. Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *CVPR, 2014*. 1, 2, 5, 7

[27] D. Tang, J. Taylor, P. Kohli, C. Keskin, T.-K. Kim, and J. Shotton. Opening the black box: Hierarchical sampling optimization for estimating human hand pose. In *ICCV, 2015*. 7

[28] D. Tang, T.-H. Yu, and T.-K. Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *ICCV*, 2013. 1, 2

[29] T. T. D. Team, R. Al-Rfou, G. Alain, A. Almahairi, C. Angermueller, D. Bahdanau, N. Ballas, F. Bastien, J. Bayer, A. Belikov, et al. Theano: A python framework for fast computation of mathematical expressions. *arXiv preprint arXiv:1605.02688*, 2016. 5

[30] J. Tompson, M. Stein, Y. Lecun, and K. Perlin. Real-time continuous pose recovery of human hands using convolutional networks. In *TOG, 2014*. 1, 2, 5

[31] T. von Marcard, G. Pons-Moll, and B. Rosenhahn. Human pose estimation from video and imus. *PAMI, 2016*. 3

[32] A. Wetzler, R. Slossberg, and R. Kimmel. Rule of thumb: Deep derotation for improved fingertip detection. In *BMVC, 2015*. 1, 2, 3, 4, 7

[33] Y. Wu, J. Y. Lin, and T. S. Huang. Capturing natural hand articulation. In *ICCV, 2001*. 4

[34] C. Xu, N. Ashwin, X. Zhang, and L. Cheng. Estimate hand poses efficiently from single depth images. *IJCV*, 2015. 1, 2, 3

[35] C. Xu and L. Cheng. Efficient hand pose estimation from a single depth image. In *ICCV, 2013*. 3

[36] Q. Ye, S. Yuan, and T.-K. Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *ECCV, 2016*. 1, 5, 7

[37] Y. Zheng, S. Sugimoto, and M. Okutomi. Aspnp: An accurate and scalable solution to the perspective-n-point problem. *IEICE TRANSACTIONS on Information and Systems*, 96(7):1525–1535, 2013. 4