Hand tracking from monocular RGB with dense semantic labels

Peter Thompson and Aphrodite Galata
Department of Computer Science, School of Engineering, University of Manchester, UK

Abstract— **Recent years have seen a renewed interest in RGB-based hand tracking, as opposed to the depth-based tracking that has dominated the field since the introduction of commodity depth cameras. This trend has been driven by the ability of convolutional neural networks to process large quantities of image data. In this paper, we propose an approach to hand tracking that operates on sets of dense semantic labels. A full pipeline for RGB-based hand tracking is presented. This pipeline uses convolutional neural networks to produce a per-pixel semantic map of the scene before optimising the state of a kinematic model according to this semantic map using a tracking algorithm based on Different Evolution (DE). This technique allows us to simultaneously localise the hand in 3D space and recover the pose, and requires only monocular RGB input. We apply our technique to a benchmark dataset, reporting semantic segmentation and 3D pose tracking results, which we compare to the current state of the art. We also compare our DE-based algorithm to an equivalent one based on Particle Swarm Optimisation (PSO) and show that it is superior.**

I. INTRODUCTION

Since the introduction of commodity depth cameras and the success of depth-based human body tracking[1], [9], [12], [36], [54], [42], [48], [59], hand pose estimation and tracking has been dominated by depth-based algorithms[6], [11], [10], [13], [20], [22], [24], [30], [37], [41], [44], [49], [50], [51], [53], [55], [58], [60], [61], [64]. However, the relative ubiquity and low cost of standard RGB cameras is a compelling reason to seek approaches to hand tracking that do not require depth data.

Many of the successful depth-based algorithms take a generative approach in which a kinematic hand model is used to evaluate a hypothesis pose according to the input depth data. The problem with applying this approach to RGB data is that, even when the physical properties of the hand are known, the appearance of the RGB image is grossly underdetermined by the hand pose.

One way to address this problem is to apply a feature transform that reduces the ambiguity of the image while retaining the information necessary to recover the hand pose. This approach was common in the era before depth cameras, when simple features, such as edges[47], colour patches[45], silhouette features[18], Haar-like features[4], and simple patch descriptors, were used. Such features are severely limited in that they do not specifically resolve the ambiguity in mapping the pose to the image.

Our approach is to perform hand tracking from dense semantic labels. These labels correspond to the different parts of the hand (i.e. the palm and phalanxes) and, as such, are unambiguous, strictly relevant to the hand pose, and unaffected by the lighting and reflectance factors that cause problems when simpler features are used. We propose a pipeline for RGB-based hand tracking in which convolution neural networks are used to semantically segment an input hand image before a kinematic model is fit to that segmentation in a generative manner. The model is fit using a tracking algorithm based on Differential Evolution (DE), and optimisation algorithm due to Storn and Price[46].

We evaluate our system on the Stereo Hand Tracking Benchmark (STB) and compare our results with the current state-of-the-art RGB-based hand pose reconstruction algorithms.

II. RELATED WORK

In recent years, there has been a resurgence of interest in hand pose estimation and tracking techniques that require only monocular RGB as input, breaking from the dominant tendency of the past decade in which depth camera input was generally required. These techniques rely heavily on deep learning.

Zimmerman and Brox[65] used multiple networks to segment the hand image and find 2D keypoint locations before simultaneously estimating the 3D keypoint locations and viewpoint. Pantileris et al.[33] used a pretrained CNN to produce 2D keypoint locations to which a 3D hand model was fit using inverse kinematics. Mueller et al.[25] took a similar approach but used generative adversarial networks to produce a large quantity of realistic synthetic data in order to train the 2D keypoint locator. These papers represent the closest work to ours in terms of the end result, since they both localise the hand in 3D space and estimate the pose from RGB input.

Iqbal et al.[15] simultaneously learned to produce 2D joint heatmaps and depth-maps then combined them to estimate the 3D joint locations. Our work is similar to these, insofar as the input is processed in image space before 3D reconstruction is considered, the distinction being that we use dense semantic information rather than sparse joint locations.

Cai et al.[3] and Dibra et al.[7] also used CNNs to predict a pose, which was then rendered into a depth image and compared with a ground truth depth image in order to refine the results, thus taking advantage of depth information during training without requiring that it be available during testing. Similarly, Rad et al.[38] used a depth to 2D keypoint approach, but also trained a separate encoder to map an RGB image to the features extracted from its corresponding depth image. Yuan et al.[62] used depth images to mask out convolutional features corresponding to non-hand regions of

This work was not supported by any organization

the image in a network that regressed directly on pose from RGB.

Nicodemou et al.[28] used a fully-convolutional architecture to map RGB to depth explicitly, taking the view that a predicted depth map of sufficient quality could be used in place of the depth camera input in any tracking system that requires it. This is similar to our approach in that the CNN is used to produce a representation of the scene to which a kinematic model can be fit, only the representation is a depth map rather than a semantic one.

Baek et al.[2] proposed a CNN architecture that predicted a mesh representation of the hand simultaneously with the hand pose. The mesh was then used to refine the pose according to extracted 2D features.

There have also been several attempts to apply generative neural network approaches to hand pose estimation. Spurr et al.[43] used a variational network to encode multiple kind of data (specifically RGB, depth, and 2D and 3D hand pose) to the same latent space. The result is a generative model from which any form of data can be sampled or generated from any other, with the RGB to 3D hand pose estimation benefiting from the regularisation provided by the other modes. Yang and Yao[**?**] built upon this approach by attempting to factor out background and viewpoint when reconstructing the pose.

CNNs capable of producing dense semantic predictions were introduced several years ago[8], [34], [35], with the fully-convolutional encoder-decoder architecture becoming standard[19], [39], [29], as it did for several tasks that require per-pixel predictions. The technique has mostly been applied to scene understanding and medical imaging.

A small number of depth-based algorithms incorporated semantic information. Tang et al.[52] used labels output from a random forest to refine their pose estimates. Neverova et al.[26] and Chen et al.[5] learned semantic information as part of the intermediate representation in a CNN that directly regressed on hand pose. The key difference between these approaches and ours is that they rely on depth input and perform semantic labelling on a segmented point cloud, as opposed to an RGB image. The semantic segmentation of RGB hand images has also been considered in a context that was agnostic of the hand tracking approach that should be applied. Neverova et al.[27] took a semi-supervised approach to train a VGG-like architecture to segment hand parts in depth images. Saleh et al.[40] used multiscale low-level feature extraction and an FCN architecture on close-cropped RGB images to perform hand parts segmentation. Our work is somewhat complementary to these, as we propose a method for tracking from the semantic results provided by these techniques.

To our knowledge, the only application of DE to hand tracking is that of Li and Zhou [17], who combined DE with particle filtering in order to perform conventional depth-based hand tracking, with a mesh-model used to create a hypothesis depth image from a hypothesis pose. More generally, optimisation techniques used with depth tended to be based on the Iterative Closest Points (ICP) algorithm, which is not applicable in our case since there is no input point cloud.

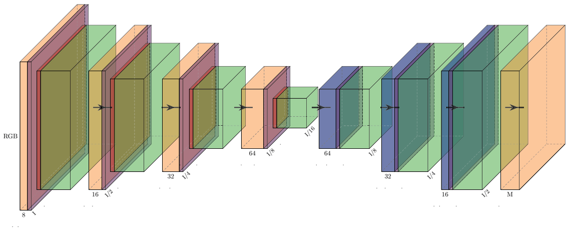


Fig. 1: The network architecture. Orange blocks represent convolution. Purple blocks represent sequences of batch normalisation, scaling, ReLU, and dropout. Red blocks represent max pooling. Green blocks represent four-layer residual units. Blue blocks represent interleaving deconvolution.

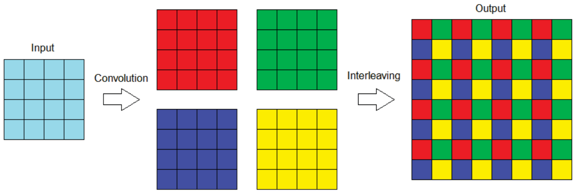


Fig. 2: Deconvolution through interleaving. Best viewed in colour.

III. Methodology

Our approach consists of two main parts. In the first, we use a fully-convolutional network to produce a semantic segmentation of the input hand image. In the second, we use a kinematic model and DE to find the pose that best explains the output of this network.

A. Semantic segmentation

We use an encoder-decoder architecture to perform semantic segmentation on our input hand images. The encoder has four downsampling blocks while the decoder has three upsampling blocks, resulting in an output half the height and width of the input. Residual units, as introduced by He et al.[14], are placed after each down or upsampling block. The architecture is shown in figure 1.

The deconvolutions are implemented through interleaving (see figure 2). This technique was suggested by Laina et al.[16] as a more efficient way of performing unpooling and convolution, as zero multiplications are avoided.

The loss is calculated per-pixel from the softmax output of the final layer. We use multinomial logistic loss with inverse-frequency class weighting, meaning each pixel in the output contributes to the loss to an extent that is inversely proportional to the frequency of its ground truth label in the batch. This is necessary because the classes are highly unbalanced in hand images, with hands typically representing less than 5% of the pixels in an image. The loss function can be written as follows,

$$E = -\frac{1}{N} \sum_{i}^{N} H_{c_i} log(\hat{p}_{c_i}), \tag{1}$$

where

$$H_c = \frac{M}{N} \sum_{i}^{N} \sum_{c'}^{M} \delta_{c'c}, \tag{2}$$

c_i is the ground truth label of pixel i, N is the total number of pixels in the batch, M is the number of classes, and $\delta_{c'c}$ is the Kronecker delta. This could be thought of as maximising the per-class accuracy (the metric generally used when evaluating semantic segmentation results) rather than per-pixel/per-example accuracy, as would be the case if the standard loss function were used.

Because the network struggles to learn the structure of the hand and separate it from the background simultaneously, we train two networks separately. One ignores any pixel labelled as background and only predicts hand-part labels. The other network only distinguishes the hand from the background. Each of these networks has the architecture described in figure 1. The final semantic map is compiled from the output of these networks as follows,

$$p(b) = kp_1(b), \tag{3}$$

$$p(c) = (1 - kp_1(b))p_2(c), \forall c \neq b, \tag{4}$$

where p_1 and p_2 are the outputs of the background and hand part networks respectively, b is the background label, and k is a constant between 0 and 1, which we refer to as the *background factor*. The effect of this constant the segmentation results will be discussed quantitatively in section IV-A. The general effect of using two networks is to allow the parts estimator to learn the composition of the hand, while the background estimator learns to distinguish the hand in the image. The background estimator does not explicitly learn any particular background and generalises well to previously unseen ones.

B. Pose optimisation

To find the hand pose, we optimise a kinematic model of the hand according to the semantic information from the CNN. The hand model was adapted from a low-polygon mesh that is freely available online [23] and is shown in figure 3.

To evaluate a given pose hypothesis, we first render the model to an image of semantic labels. For each pixel, we then look up probability of the label given by the network. We then calculate a cost that is the sum of the complement of these probabilities. The cost function can be written as follows,

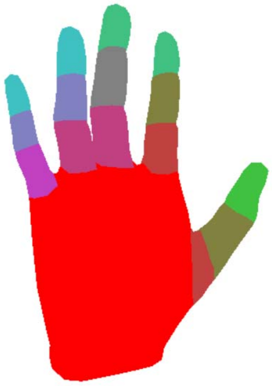


Fig. 3: Hand model with parts labels.

$$C = \sum_{i}^{N} 1 - p_i(R_i), \tag{5}$$

where p_i is the probability distribution over the possible labels for the ith pixel in the network output, and R_i is the label of the ith pixel in the rendered semantic image.

We minimise this cost function using Differential Evolution (DE), an algorithm due to Storn and Price[46]. This is a simple, population-based metaheuristic that is robust to noise and does not require a gradient, which is a necessary feature in our case, since the relationship between the pose and semantic projection is not smooth. Optimisation methods such as Levenberg-Marquardt cannot be used for this reason. The ICP-like algorithms used with depth input are also inapplicable in this case, since there is no point cloud. Particle Swarm Optimisation (PSO) is an applicable algorithm and has been used extensively in depth-based hand tracking[31], [37], [41], [55]. However, we find it does not perform well in this context (see section IV-B).

For each frame, the DE algorithm runs for a fixed number of generations. We use a variation of DE in which the current best agent is selected for mutation with a probability of 50% with a random agent being selected otherwise. The selected agent is then mutated according to the difference between two others. We refer to this variant of DE in the notation suggested by Storn and Price as DE/(best-rand)/1/bin. The best candidate solution found on each frame is taken as the solution. The next frame inherits the entire population with the subset representing the worst quality candidates replaced by mutations of the best. These mutations are generated by perturbing each parameter according to a normal distribution, the standard deviation of which is a constant fraction of its range. This prevents the population from converging too closely to a particular solution, thus losing the ability to adapt to the next frame, without losing too much information. Pseudocode for this algorithm is shown in Algorithm 1.

We optimise over the 26 degree-of-freedom parameter space of the hand model and enforce boundary conditions

	IOU	Prec.	Rec.	F1
Ours	0.633	0.733	0.822	0.775
Baek et al.[2]	0.651	0.828	0.753	0.789
Zimmermann et al.[65]	0.354	0.365	0.921	0.523
Urooj and Borji[57]	0.527	0.717	0.666	0.690

TABLE I: Comparison of the background estimation with examples from the literature on the RHD testing set. The metric are intersect-over-union (IOU), precision, recall, and f1-score. Higher is better in all cases.

corresponding to the biomechanical constraints on the hand joints. We also bound root of the model inside a reasonable depth range and the camera's frustum volume (i.e. the boundaries of the image).

Algorithm 1 DE-based tracking algorithm.

```
randomly initialise population of N agents
for each frame do
    run DE/(best-rand)/1/bin for G generations
    sort agents according to loss
    select best agent as result for current frame
    replace worst M agents with mutations of best
```

IV. Experiments

We followed the experimental setup of Zimmerman and Brox[65], using two sequences ('B1Counting' and 'B1Random') from the Stereo Hand Tracking Benchmark (STB)[63] for testing, with the rest of the sequences plus the training set from the Rendered Handpose Dataset (RHD) used for training. We use STB because it provides the focal length and field of view of the camera used to capture the RGB sequences, both of which we need in order to accurately render our model in the original image space and compare our keypoint locations to the ground truth. The RHD and Back To RGB[32] datasets do not provide this information. The Hands In Action dataset[56] provides camera information but does not provide 3D keypoints.

We used the semantic ground truth provided with RHD, with the labels transformed to correspond to those in our model, including mapping the right hand labels to left hand ones, since only left hands are present in STB. To acquire a semantic ground truth for the STB data, we fit our model to the ground truth 3D joint locations (using DE) and rendered the scene as a semantic image. We also use this technique to obtain approximate ground truth keypoint positions, since our model's keypoints are in slightly different positions to the provided ground truth.

A. Segmentation results

The background and parts estimator were both trained using stochastic gradient descent with a learning rate starting at 0.01 and lowered every 50000 iterations by factor of ten until convergence. The background estimator was trained on both the STB and RHD training sets, whereas the parts estimator was only trained on STB. One epoch is equal 3514 and 936 iterations for these training sets respectively. The results from each were then combined using a background factor, as described in section III-A.

We find that varying the background factor represents a trade-off between the accurate labelling of the background pixels adjacent to the hand and that of important details in the periphery of the hand region. This trade-off can be demonstrated quantitatively by comparing per-class and per-pixel accuracy. As figure 5 shows, per-pixel accuracy tends to increase with the background factor, as more background pixels are labelled correctly, but per-class accuracy tends to decrease as parts of the hand start to be labelled as background. The IOU and F1 values also tend to increase with background factor, though also decrease as it approaches 1. Figure 6 show a qualitative example of this, with segmentation results for higher background factors being less blurry, with the boundary between the background and hand more precisely defined, but also tending to lose important details around the edge of the hand region, such as the shape of the palm and protruding fingers.

We choose a background factor of 0.5 for subsequent results, as it is balances these two concerns. Various examples from across the testing set with this background factor are shown in the middle columns of figures 10 and 11.

We also evaluate the background estimation against current state-of-the-art hand tracking systems that use hand segmentation as part of the preprocessing. Table I shows that our architecture is approximately as good or better than the current state-of-the-art when applied to the RHD testing set.

B. Tracking results

Our tracking algorithm was applied to the STB testing sequences with 16 agents, 50 generations per frame, and 8 agents mutated after each frame. The crossover probability and differential weight were 0.1 and 0.3 respectively. All of the quantitative results are based five runs of the tracking algorithm, with the relevant quantities averaged.

The algorithm was first compared against an equivalent algorithm based on PSO on both the ground truth and predicted semantic maps. The results of this are shown in table II and figures 7 and 8. It can be seen that the DE-based algorithm outperforms the PSO-based one in all cases, with the joint errors of the DE-based algorithm operating on the ground truth semantic maps being very low.

The algorithm was also compared against some of the current state-of-the-art RGB-based hand pose reconstruction algorithms described in section II. Figure 9 shows comparison PCK curves. When operating on the ground truth semantic maps, the results are comparable to some of the contemporary state-of-the-art algorithms. When operating on the the estimated semantic maps, however, the algorithm falls short. The main reason for this appears to be ambiguity in the global orientation of the hand. This is not an issue when the semantic ground truth is used, as the palm in the ground truth is the same shape as in the model. When using estimated semantic maps, however, the shape of the palm corresponds to that of the subject, meaning the model tends

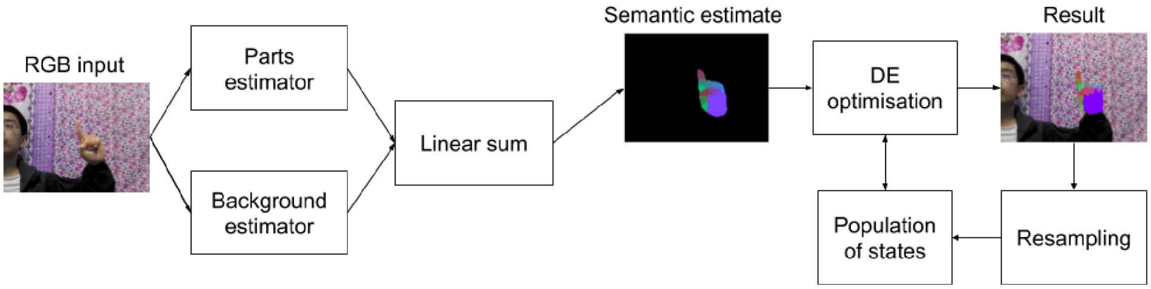
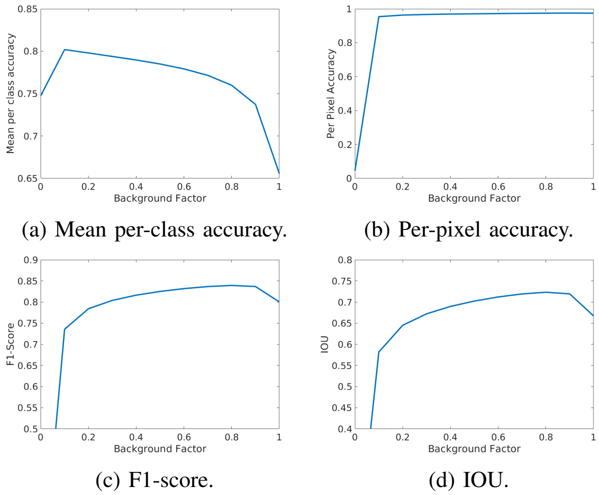


Fig. 4: The proposed hand tracking pipeline. The population of states is maintained between frames, with some agents being replaced with ones sampled from around the previous result.



(a) Mean per-class accuracy.

(b) Per-pixel accuracy.

(c) F1-score.

(d) IOU.

Fig. 5: The performance of the segmentation networks on STB in terms of several different metrics as a function of the background factor.

to be slightly out of position when fit according to these maps. To demonstrate that this is the main cause of error, a global orientation oracle was acquired by fitting the model to the ground truth joint locations and noting the global orientation values. The results of running the algorithm with these values imputed are also shown in figure 9. It can be seen that, with this information made available, the results are greatly improved and comparable to the best of the current state-of-the-art.

Figures 10 and 11 show qualitative results from different frames in both sequences of the dataset. These results were produced by the algorithm operating on the estimated semantic maps, without accessing the global orientation oracle.

V. Conclusion

In this paper, we presented a novel approach to hand tracking in which dense semantic labels are used to fit a kinematic hand model in a generative manner and proposed a complete pipeline for performing hand tracking on RGB images. The pipeline consisted of CNN-based semantic segmentation and DE-based tracking algorithm. The semantic segmentation networks performed well, with the results being accurate and physically plausible. The semantic segmentation results were also shown to be roughly as good as the state-of-the-art where a quantitative comparison was possible. The tracking algorithm was shown to perform favourably to an equivalent algorithm based on PSO, which is a common optimisation algorithm used in hand tracking when operating on both ground truth and estimated semantic maps. When compared to the current state-of-the-art in RGB-based hand tracking, the algorithm is comparable when operating on the ground

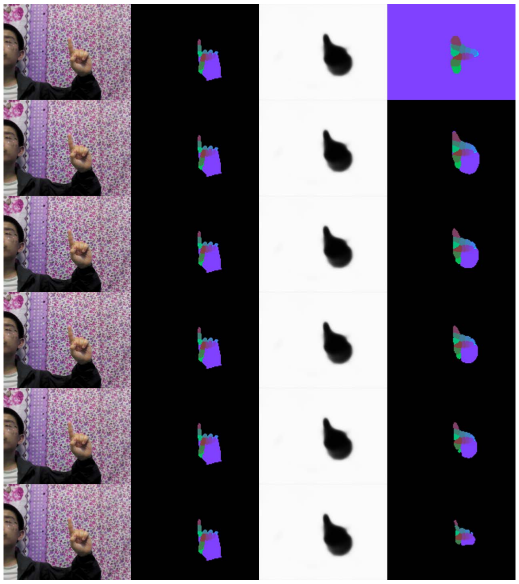


Fig. 6: Segmentation results for different background factors. The columns from left to right show the input RGB image, the estimation semantic ground truth, the output of the background network, and the final segmentation result.

	Counting	Random
DE estimate	34.9	32.7
DE ground	16.1	24.0
PSO estimate	71.5	93.7
PSO ground	93.7	106.7

TABLE II: Mean relative joint error in millimetres across each sequence in the STB testing set for DE and PSO-based tracking.

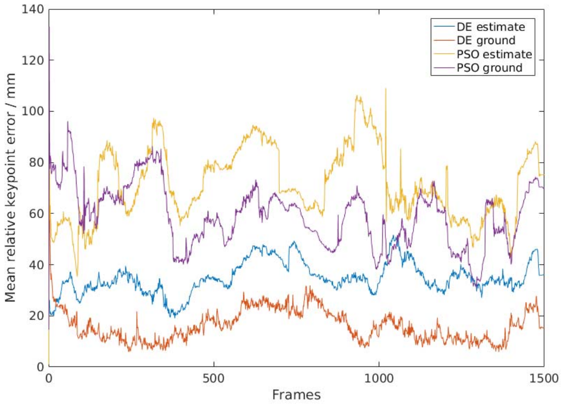


Fig. 7: Mean relative keypoint error for "B1Counting" for DE and PSO-based tracking.

truth semantic maps. When operating on estimated semantic maps, the algorithm falls short. The main reasons for this was determined to be ambiguity in global orientation. This could be addressed in future work in several ways. One way would be to learn to predict the global orientation directly in a discriminative manner. Another would be to learn features that allow the global orientation to be ascertained more easily, such as landmarks on the surface of the hand, and consider these features alongside the semantic labels in the optimisation procedure. The use of a hand model that more closely resembles the subject's hand would also improve the tracking results. A mesh representation of the subjects hand could be determined at deploy time using a deep learning models, in a manner similar to the proposed approaches of Malik et al.[21] and Baek et al.[2].

References

[1] Andreas Baak, Meinard Müller, Gaurav Bharaj, Hans-Peter Seidel, and Christian Theobalt. A data-driven approach for real-time full body pose reconstruction from a depth camera. In *Consumer Depth Cameras for Computer Vision*, pages 71–98. Springer, 2013.

[2] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. Pushing the envelope for rgb-based dense 3d hand pose estimation via neural rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1067–1076, 2019.

[3] Yujun Cai, Liuhao Ge, Jianfei Cai, and Junsong Yuan. Weakly-supervised 3d hand pose estimation from monocular rgb images. In *European Conference on Computer Vision*, pages 1–17. Springer, Cham, 2018.

[4] Qing Chen, Nicolas D Georganas, and Emil M Petriu. Real-time vision-based hand gesture recognition using haar-like features. In *2007 IEEE instrumentation & measurement technology conference IMTC 2007*, pages 1–6. IEEE, 2007.

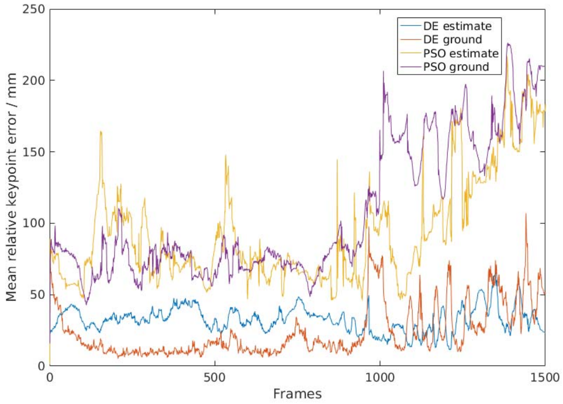


Fig. 8: Mean relative keypoint error for "B1Random" for DE and PSO-based tracking.

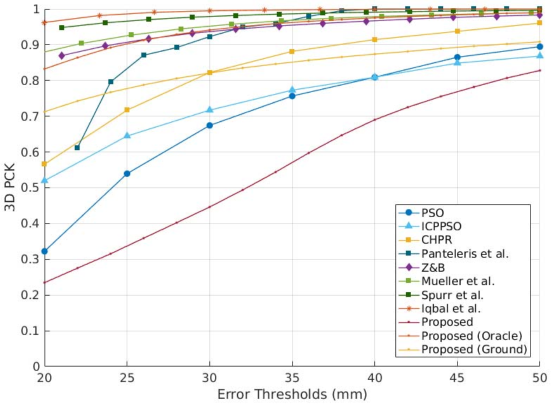


Fig. 9: Our PCK results shown alongside the current state of the art. Results are shown for the algorithm operating on estimated and ground truth semantic maps, as well as the estimated maps with access to the global orientation oracle.

[5] Xinghao Chen, Guijin Wang, Cairong Zhang, Tae-Kyun Kim, and Xiangyang Ji. Shpr-net: Deep semantic hand pose regression from point clouds. *IEEE Access*, 6:43425–43439, 2018.

[6] Xiaoming Deng, Shuo Yang, Yinda Zhang, Ping Tan, Liang Chang, and Hongan Wang. Hand3d: Hand pose estimation using 3d neural network. *arXiv preprint arXiv:1704.02224*, 2017.

[7] Endri Dibra, Silvan Melchior, Ali Balkis, Thomas Wolf, Cengiz Oztireli, and Markus Gross. Monocular rgb hand pose inference from unsupervised refinable nets. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1188–118810. IEEE, 2018.

[8] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun. Learning hierarchical features for scene labeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1915–1929, 2013.

[9] Varun Ganapathi, Christian Plagemann, Daphne Koller, and Sebastian Thrun. Real time motion capture using a single time-of-flight camera. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 755–762. IEEE, 2010.

[10] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8417–8426, 2018.

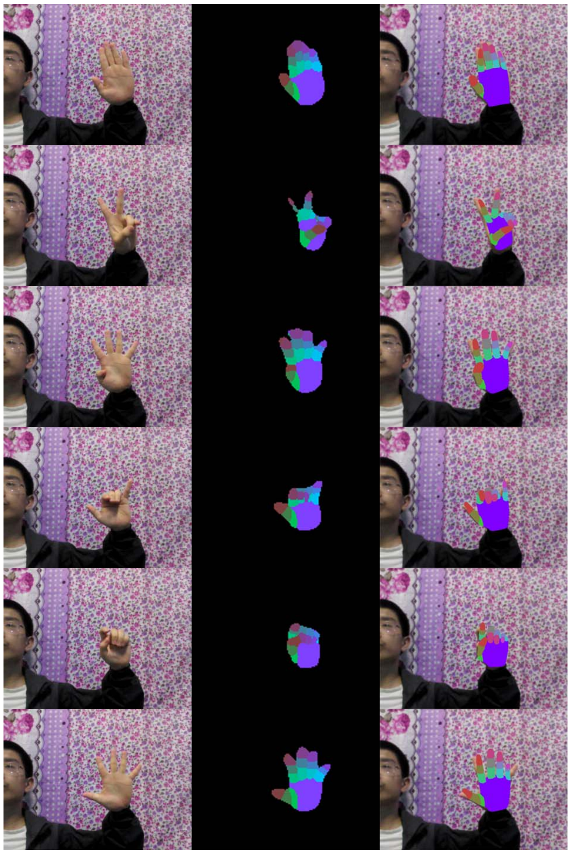


Fig. 10: Example tracking results from "B1Counting". From left to right, the columns show the input, semantic segmentation, and the final results overlaid on the input.

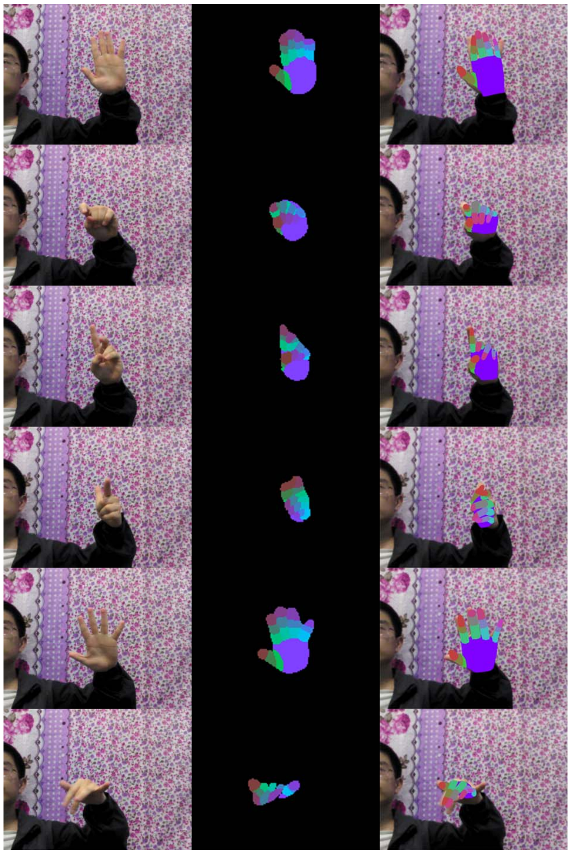


Fig. 11: Example tracking results from "B1Random". From left to right, the columns show the input, semantic segmentation, and the final results overlaid on the input.

[11] Liuhao Ge, Hui Liang, Junsong Yuan, and Daniel Thalmann. 3d convolutional neural networks for efficient and robust hand pose estimation from single depth images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, page 5, 2017.

[12] Ross Girshick, Jamie Shotton, Pushmeet Kohli, Antonio Criminisi, and Andrew Fitzgibbon. Efficient regression of general-activity human poses from depth images. In *2011 International Conference on Computer Vision*, pages 415–422. IEEE, 2011.

[13] Hengkai Guo, Guijin Wang, Xinghao Chen, Cairong Zhang, Fei Qiao, and Huazhong Yang. Region ensemble network: Improving convolutional network for hand pose estimation. In *Image Processing (ICIP), 2017 IEEE International Conference on*, pages 4512–4516. IEEE, 2017.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[15] Umar Iqbal, Pavlo Molchanov, Thomas Breuel Juergen Gall, and Jan Kautz. Hand pose estimation via latent 2.5 d heatmap regression. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 118–134, 2018.

[16] Iro Laina, Christian Rupprecht, Vasileios Belagiannis, Federico Tombari, and Nassir Navab. Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*, pages 239–248. IEEE, 2016.

[17] Dongnian Li and Yiqi Zhou. Combining differential evolution with particle filtering for articulated hand tracking from single depth images. *Int J Signal Process Image Process Pattern Recognit*, 8(4):237–248, 2015.

[18] John Y Lin, Ying Wu, and Thomas S Huang. 3d model-based hand tracking using stochastic direct search method. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 693–698. IEEE, 2004.

[19] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[20] Meysam Madadi, Sergio Escalera, Xavier Baró, and Jordi Gonzalez. End-to-end global to local cnn learning for hand pose recovery in depth data. *arXiv preprint arXiv:1705.09606*, 2017.

[21] Jameel Malik, Ahmed Elhayek, Fabrizio Nunnari, Kiran Varanasi, Kiarash Tamaddon, Alexis Heloir, and Didier Stricker. Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth. In *2018 International Conference on 3D Vision (3DV)*, pages 110–119. IEEE, 2018.

[22] Jameel Malik, Ahmed Elhayek, and Didier Stricker. Structure-aware 3d hand pose regression from a single depth image. In *International Conference on Virtual Reality and Augmented Reality*, pages 3–17. Springer, 2018.

[23] MatSoft. human-hands male basemesh. https://www.turbosquid.com/3d-models/free-basemesh-human-hands-3d-model/770920, retrieved 2019-07-22, 2013.

[24] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2018.

[25] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. Ganerated

hands for real-time 3d hand tracking from monocular rgb. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 49–59, 2018.

[26] Natalia Neverova, Christian Wolf, Florian Nebout, and Graham W Taylor. Hand pose estimation through semi-supervised and weakly-supervised learning. *Computer Vision and Image Understanding*, 164:56–67, 2017.

[27] Natalia Neverova, Christian Wolf, Graham W Taylor, and Florian Nebout. Hand segmentation with structured convolutional learning. In *Asian Conference on Computer Vision*, pages 687–702. Springer, 2014.

[28] Vassilis C Nicodemou, Iason Oikonomidis, Georgios Tzimiropoulos, and Antonis Argyros. Learning to infer the depth map of a hand from its color image. *arXiv preprint arXiv:1812.02486*, 2018.

[29] Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1520–1528, 2015.

[30] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. Hands deep in deep learning for hand pose estimation. *arXiv preprint arXiv:1502.06807*, 2015.

[31] Iason Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Efficient model-based 3d tracking of hand articulations using kinect. In *BmVC*, volume 1, page 3, 2011.

[32] Paschalis Panteleris and Antonis Argyros. Back to rgb: 3d tracking of hands and hand-object interactions based on short-baseline stereo. *Hand*, 2(63):39, 2017.

[33] Paschalis Panteleris, Iason Oikonomidis, and Antonis Argyros. Using a single rgb frame for real time 3d hand pose estimation in the wild. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 436–445. IEEE, 2018.

[34] George Papandreou, Liang-Chieh Chen, Kevin P Murphy, and Alan L Yuille. Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation. In *Proceedings of the IEEE international conference on computer vision*, pages 1742–1750, 2015.

[35] Pedro O Pinheiro and Ronan Collobert. Weakly supervised semantic segmentation with convolutional networks. In *CVPR*, volume 2, page 6. Citeseer, 2015.

[36] Gerard Pons-Moll12, Jonathan Taylor13, Jamie Shotton, Aaron Hertzmann14, and Andrew Fitzgibbon. Metric regression forests for human pose estimation. BMVC, 2013.

[37] Chen Qian, Xiao Sun, Yichen Wei, Xiaoou Tang, and Jian Sun. Realtime and robust hand tracking from depth. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1106–1113, 2014.

[38] Mahdi Rad, Markus Oberweger, and Vincent Lepetit. Domain transfer for 3d pose estimation from color images without manual annotations. In *Asian Conference on Computer Vision*, pages 69–84. Springer, 2018.

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[40] Adel Saleh, Hatem Rashwan, Mohamed Abdel-Nasser, Vivek Singh, Saddam Abdulwahab, Md. Mostafa Kamal Sarker, Miguel Garca, and Domenec Puig. Finseg: Finger parts semantic segmentation using multi-scale feature maps aggregation of fcn. 02 2019.

[41] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, et al. Accurate, robust, and flexible real-time hand tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3633–3642. ACM, 2015.

[42] Jamie Shotton, Andrew Fitzgibbon, Andrew Blake, Alex Kipman, Mark Finocchio, Bob Moore, and Toby Sharp. Real-time human pose recognition in parts from a single depth image. 2011.

[43] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. Cross-modal deep variational hand pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 89–98, 2018.

[44] Srinath Sridhar, Antti Oulasvirta, and Christian Theobalt. Interactive markerless articulated hand motion tracking using rgb and depth data. In *Proceedings of the IEEE international conference on computer vision*, pages 2456–2463, 2013.

[45] Björn Stenger, Arasanathan Thayananthan, Philip HS Torr, and Roberto Cipolla. Model-based hand tracking using a hierarchical bayesian filter. *IEEE transactions on pattern analysis and machine intelligence*, 28(9):1372–1384, 2006.

[46] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[47] Erik B Sudderth, Michael I Mandel, William T Freeman, and Alan S Willsky. Visual hand tracking using nonparametric belief propagation. In *Computer Vision and Pattern Recognition Workshop, 2004. CVPRW'04. Conference on*, pages 189–189. IEEE, 2004.

[48] Min Sun, Pushmeet Kohli, and Jamie Shotton. Conditional regression forests for human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3394–3401. IEEE, 2012.

[49] Xiao Sun, Yichen Wei, Shuang Liang, Xiaoou Tang, and Jian Sun. Cascaded hand pose regression. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 824–832, 2015.

[50] Andrea Tagliasacchi, Matthias Schröder, Anastasia Tkach, Sofien Bouaziz, Mario Botsch, and Mark Pauly. Robust articulated-icp for real-time hand tracking. In *Computer Graphics Forum*, volume 34, pages 101–114. Wiley Online Library, 2015.

[51] Danhang Tang, Hyung Jin Chang, Alykhan Tejani, and Tae-Kyun Kim. Latent regression forest: Structured estimation of 3d articulated hand posture. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3786–3793, 2014.

[52] Danhang Tang, Tsz-Ho Yu, and Tae-Kyun Kim. Real-time articulated hand pose estimation using semi-supervised transductive regression forests. In *Proceedings of the IEEE international conference on computer vision*, pages 3224–3231, 2013.

[53] Jonathan Taylor, Lucas Bordeaux, Thomas Cashman, Bob Corish, Cem Keskin, Toby Sharp, Eduardo Soto, David Sweeney, Julien Valentin, Benjamin Luff, et al. Efficient and precise interactive hand tracking through joint, continuous optimization of pose and correspondences. *ACM Transactions on Graphics (TOG)*, 35(4):143, 2016.

[54] Jonathan Taylor, Jamie Shotton, Toby Sharp, and Andrew Fitzgibbon. The vitruvian manifold: Inferring dense correspondences for one-shot human pose estimation. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 103–110. IEEE, 2012.

[55] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics (ToG)*, 33(5):169, 2014.

[56] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. Capturing hands in action using discriminative salient points and physics simulation. *International Journal of Computer Vision*, 118(2):172–193, 2016.

[57] Aisha Urooj and Ali Borji. Analysis of hand segmentation in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4710–4719, 2018.

[58] Chengde Wan, Angela Yao, and Luc Van Gool. Hand pose estimation from local surface normals. In *European conference on computer vision*, pages 554–569. Springer, 2016.

[59] Mao Ye, Xianwang Wang, Ruigang Yang, Liu Ren, and Marc Pollefeys. Accurate 3d pose estimation from a single depth image. In *2011 International Conference on Computer Vision*, pages 731–738. IEEE, 2011.

[60] Qi Ye, Shanxin Yuan, and Tae-Kyun Kim. Spatial attention deep net with partial pso for hierarchical hybrid hand pose estimation. In *European conference on computer vision*, pages 346–361. Springer, 2016.

[61] Shanxin Yuan, Guillermo Garcia-Hernando, Björn Stenger, Gyeongsik Moon, Ju Yong Chang, Kyoung Mu Lee, Pavlo Molchanov, Jan Kautz, Sina Honari, Liuhao Ge, et al. Depth-based 3d hand pose estimation: From current achievements to future goals. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[62] Shanxin Yuan, Bjorn Stenger, and Tae-Kyun Kim. Rgb-based 3d hand pose estimation via privileged learning with depth images. *arXiv preprint arXiv:1811.07376*, 2018.

[63] Jiawei Zhang, Jianbo Jiao, Mingliang Chen, Liangqiong Qu, Xiaobin Xu, and Qingxiong Yang. 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*, 2016.

[64] Xingyi Zhou, Qingfu Wan, Wei Zhang, Xiangyang Xue, and Yichen Wei. Model-based deep hand pose estimation. *arXiv preprint arXiv:1606.06854*, 2016.

[65] Christian Zimmermann and Thomas Brox. Learning to estimate 3d hand pose from single rgb images. In *International Conference on Computer Vision*, volume 1, page 3, 2017.