3D Hand Shape and Pose Estimation based on 2D Hand Keypoints

Drosakis Drosakis
drosakis@ics.forth.gr
University of Crete and FORTH
Heraklion, Crete, Greece

Antonis Argyros
argyros@ics.forth.gr
University of Crete and FORTH
Heraklion, Crete, Greece

ABSTRACT

We present a method for simultaneous 3D hand shape and pose estimation on a single RGB image frame. Specifically, our method fits the MANO 3D hand model to 2D hand keypoints. Fitting is achieved based on a novel 2D objective function that exploits anatomical joint limits, combined with shape regularization on the MANO hand model, jointly optimizing the 3D shape and pose of the hand in a single frame. In a series of quantitative experiments on well-established datasets annotated with ground truth, we show that it is possible to obtain reconstructions that are competitive and, in some cases, superior to existing 3D hand pose estimation approaches.

KEYWORDS

3D Hand Shape, 3D Hand Pose, Optimization

ACM Reference Format:
Drosakis Drosakis and Antonis Argyros. 2023. 3D Hand Shape and Pose Estimation based on 2D Hand Keypoints. In *Proceedings of the 16th International Conference on PErvasive Technologies Related to Assistive Environments (PETRA '23), July 05–07, 2023, Corfu, Greece.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3594806.3594838

1 INTRODUCTION

Hands are the most important tools for humans. They are the primary means of interacting with the world and assist humans in performing a multitude of tasks. Machines that are capable of observing and understanding the motion of human hands will be able to interact effectively with humans and they will have the capacity to mimic such motions to automate certain repetitive tasks. Moreover, with the rise of Virtual Reality headsets that use egocentric cameras for tracking, 3D hand pose estimation can be deployed for a more realistic experience in the virtual world, so that the manipulation of virtual objects is performed with a non-instrumented hand instead of a controller.

Many recent approaches [1, 2, 6, 8, 9, 15, 21, 23–25] try to solve the 3D hand pose estimation problem using either RGB or RGBD input. With RGBD input, 3D hand pose is typically estimated by fitting a 3D hand model to the depth observations. Depth is a powerful cue for inferring the shape and articulation of a visible hand and methods which use such input have great accuracy [9, 12, 15, 17–20].

PETRA '23, July 05–07, 2023, Corfu, Greece

ACM ISBN 979-8-4007-0069-9/23/07.
https://doi.org/10.1145/3594806.3594838

If only RGB information is available, the problem becomes more difficult due to ambiguities, self occlusions and uniform hand texture. However, using RGB input alone is highly preferable compared to RGBD camera-based solutions because RGB cameras are cheaper and much more common. Moreover, such solutions are operational in outdoor environments where RGBD sensors typically do not perform appropriately due to the direct sunlight.

State of the art methods with RGB input, use neural networks, such as Convolutional Neural Network (CNN) to directly regress the 3D joint locations of the hand from the RGB image. Moreover, they either predict 2D joint locations on the image as an intermediate step and then produce the 3D joints [2, 23, 24] or they produce the 3D joints directly from the input image [6, 25]. Because the training data with annotated 3D ground truth is somewhat limited, the use of 2D supervision as an intermediate step improves accuracy. However, these methods fail to completely generalize to unseen hand poses and shapes or in the presence of strong occlusions.

In this work, we rely on 2D hand keypoints that we lift into 3D via an optimization scheme. More specifically, we fit a 3D hand model, jointly optimizing its shape and pose parameters, to the 2D observations. The employed optimization scheme minimizes the 3D reprojection error, that is, the distances of the reprojected 3D hand joints estimations and the 2D keypoints detected by a state of the art 2D keypoint estimation [13]. We show that our method performs competitively to the state of the art on 3D hand pose estimation. This is demonstrated qualitatively, but also quantitatively on standard datasets that are annotated with ground truth.

2 RELATED WORK

We discuss computational methods that use single image RGBD or RGB input for 3D hand pose estimation. Methods with RGBD input typically fit a generative hand model to the depth data. Optimization methods for RGB fit a 3D hand model to 2D evidence in the image [2, 10]. On the contrary, learning-based methods directly regress joint locations or joint rotations and hand shape parameters.

Depth-based methods leverage the rich depth data to fit a model of the hand. Sridhar *et al.* [15] perform hand tracking via a detection guided optimization using a single depth sensor. They first detect and classify pixels into parts of the hand and then they combine the detected part labels with a Gaussian mixture representation of the depth to fit a hand model pose. The hand model's shape is also defined with a mixture of Gaussians and they present an automatic procedure to fit the model to a specific user. The depth data provides strong constraints for the optimization leading to great accuracy of the method but the depth sensor is sensitive to direct sunlight making it harder to use for outdoor scenarios.

Using an RGB sensor can eliminate these issues while being a more accessible and lower cost solution. Optimization based methods for RGB predict the 2D hand keypoints in the image and then lift the 2D representation to 3D. Panteleris *et al.* [10] use a fixed hand model and optimize the joint rotations by minimizing the 2D reprojection error of the joints. Learning-based methods directly regress the 3D joint locations or the parameters of a 3D model. Zhang *et al.* [23] present a method to recover a hand mesh representation using the more expressive MANO hand model [11]. Using this representation makes it possible to also obtain the joint locations with linear interpolation of the mesh vertices as well as the 2D keypoints in the image by projecting the 3D joints computed. The fact that MANO parameterizes effectively hand shape and pose in a differentiable manner simplifies the training of neural networks that regress a mesh representation of the hand.

Jointly optimizing the hand shape and pose parameters of a hand model based only on 2D evidence though is challenging. A naive optimization of both parameters in a single frame is underconstrained, that is, there exist several optimal hand shape - pose pairs. Makris and Argyros [5] perform online shape adaptation while tracking the pose of the hand. They use the already computed poses of previous frames to infer a better shape and continue tracking the hand pose with the newly computed shape. This approach relies on progressively adapted poses computed based on a progressively adapted hand shape with the goal to both converge to accurate estimations. We use shape regularization provided by the MANO 3D hand model [11] which essentially makes the optimizer to converge to the shape that is closer to the mean hand shape. This way we can compute both the shape and pose from a single frame.

Boukhayma *et al.* [2] proposed a learning based approach to directly predict the shape and pose parameters of the MANO hand model [11] and they also tested an optimization scheme for hand pose estimation, similar to that of [10]. They compare the optimization scheme that performs 2D fitting on the 2D detections with their learning-based method, which also uses 2D keypoint detections from a separate CNN, showing that their method outperforms the fitting. In the optimization method though they optimize a pose space that has reduced dimensionality using PCA in an effort to remove implausible hand articulations by using regularization in that pose space. We show that by setting explicit anatomical limits in the full pose space instead, we outperform the optimization in the PCA pose space and perform similarly with the learning-based method that they proposed.

Several learning-based methods have been proposed that make substantial progress to generalization capabilities of neural networks. Zimmermann and Brox [25] use a CNN to directly regress 3D joint locations from the RGB image. Mueller *et al.* [6] propose the use of a Generative Adversarial Network to translate synthetic images of hands to real in an effort to bridge the real-synthetic domain gap when training networks to perform hand pose estimation from RGB input. Spurr *et al.* [14] propose the use of Variational Autoencoders, Baek *et al.* [1] proposes to use a neural network combined with a differentiable renderer and Iqbal *et al.* [3] propose to estimate the hand pose through a 2.5D representation, estimating the pose up to a scaling factor that can be determined if a prior of the hand size is known.

Learning and optimization can be combined to leverage the advantages of each one. Xiang *et al.* [21] present a method to jointly estimate the pose of a human body, face and hands and they predict a single representation that encodes the poses of all body parts from an RGB image. They use the predicted representation along with predicted joint confidence maps to fit a deformable 3D mesh model via optimization.

Zhou *et al.* [24] proposed a learning based approach that detects the 3D joint positions of the hand and then computes the shape and pose of the MANO hand model. Their method uses a learning based inverse kinematics (IK) solver to obtain the joint rotations from the joint positions and performs PSO optimization using the bone lengths to obtain the shape of the hand, thus getting the full pose and shape of the hand. Li *et al.* [4] propose a hybrid analytical and neural inverse kinematics solution for human pose estimation. For the hand however, the analytical part of the solution suffices for correctly computing inverse kinematics. So, with these two methods, one can obtain the full shape and pose of the MANO hand model from 3D joint positions by performing a PSO optimization for the shape and analytically computing the pose from the joint positions. This method achieves state of the art results on the well established datasets Dexter+Object [16] and EgoDexter [7].

3 METHOD

Our method takes as input one RGB image, detects the 2D hand keypoints and then fits a parametric 3D hand model to the 2D keypoints via optimization, minimizing an objective function with a joint reprojection error.

Hand Model: We use the MANO hand model [11] that takes as input 45 rotation parameters $\boldsymbol{\theta}$ and 10 shape parameters $\boldsymbol{\beta}$ and produces a 3D hand mesh. The MANO model is defined as a differentiable function as follows:

$$M(\boldsymbol{\beta}, \boldsymbol{\theta}) = W(T_p(\boldsymbol{\beta}, \boldsymbol{\theta}), J(\boldsymbol{\beta}), \boldsymbol{\theta}, \boldsymbol{W}) \tag{1}$$

and

$$T_p(\boldsymbol{\beta}, \boldsymbol{\theta}) = \bar{\boldsymbol{T}} + B_S(\boldsymbol{\beta}) + B_P(\boldsymbol{\theta}), \tag{2}$$

where $\boldsymbol{\beta} \in \mathbb{R}^{10}$ and $\boldsymbol{\theta} \in \mathbb{R}^{15x3}$ are the hand shape and pose parameters, respectively, and W a linear blend skinning function applied to a template hand mesh T_p which is obtained by deforming a mean hand mesh $\bar{\boldsymbol{T}}$ with shape and pose blend shape functions B_S and B_P, respectively. Moreover, joints are denoted with J and blend weights with $\boldsymbol{W}$. The resulting mesh $\boldsymbol{M} \in \mathbb{R}^{Nx3}$, where N the number of vertices, is then posed in space with orientation $\boldsymbol{r} \in \mathbb{R}^3$ and translation $\boldsymbol{t} \in \mathbb{R}^3$.

2D keypoint detector: We utilize the OpenPose hand keypoint detector [13], a CNN which takes as input an image containing a human hand and outputs the 21 hand keypoint locations along with a confidence score for each keypoint.

Error function: The error function that guides the optimization consists of 3 terms, a keypoint reprojection error E_{key}, an anatomical joint limits error E_{limits} and a shape regularization error E_{shape}:

$$E(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{r}) = E_{key}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{r}) + E_{limits}(\boldsymbol{\theta}) + E_{shape}(\boldsymbol{\beta}). \tag{3}$$

In more detail, the keypoint reprojection error E_{key} penalizes the difference between the 2D keypoints detected by the CNN and the projections of the 3D joints of the hand model to the image plane. In notation,

$$E_{key}(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{t}, \boldsymbol{r}) = \sum_{i=0}^{n} w_i \left\| p_i - k_i \right\|_2^2, \tag{4}$$

where k_i is the i$^{\text{th}}$ keypoint detected with confidence w_i and $p_i = j_i K^T$ with j_i being the 3D joint obtained with linear interpolation of the hand mesh vertices, projected onto the image with a camera intrinsics matrix K. It is noted that the camera is assumed to be at the world origin.

The anatomical joint limits error E_{limits} penalizes the 45-dimensional pose vector (3 Degrees of Freedom for each joint) when it exceeds some experimentally defined anatomical limits for obtaining plausible hand articulations, similar to [20]. This error is defined as a soft constraint with two exponential functions with starting positions at the lower and upper bound of the limit, respectively. Specifically,

$$E_{limits}(\boldsymbol{\theta}) = a_{limits} \sum_{i=0}^{m} (e^{l_i - \theta_i} + e^{\theta_i - u_i}), \tag{5}$$

where $[l_i, u_i]$ are the joint limits for joint angle θ_i and $a_{limits} = 10^3$ is an experimentally identified weight factor.

The shape regularization error E_{shape} penalizes the shape parameters and forces the optimizer to converge to a plausible shape as close to the mean (zero) hand shape as possible. In notation,

$$E_{shape}(\boldsymbol{\beta}) = a_{shape} \left\| \boldsymbol{\beta} \right\|_2^2, \tag{6}$$

with $a_{shape} = 10^3$.

Implementation details: We use PyTorch for optimization. For the 2D keypoint estimation we feed a crop of the frame containing the hand to OpenPose [13]. An important detail is how we detect the bounding box of the hand. For that, we employ SRHandNet [22] bounding box detection module and feed that to OpenPose hand keypoint detection CNN.

We use the LBFGS optimizer to fit the MANO model parameters to the detected 2D keypoints, minimizing the objective function described above, obtaining the 3D shape and pose of the hand in the image.

4 EVALUATION

We compare our method to the state of the art methods and to optimization methods competitive to ours.

Datasets: We evaluate our method with the well established datasets EgoDexter (ED) [7], Dexter+Object (DO) [16] and the Hands in Action (HIC) [20]. EgoDexter is an egocentric hand manipulation dataset containing 4 sequences and has heavy hand occlusions both from objects and self occlusions. Dexter+Object is a dataset containing 6 sequences showing a hand manipulating an object from a third view. The Hands in Action dataset contains sequences both from hand-only motion and hand object manipulations and exhibits less occlusions compared to EgoDexter and

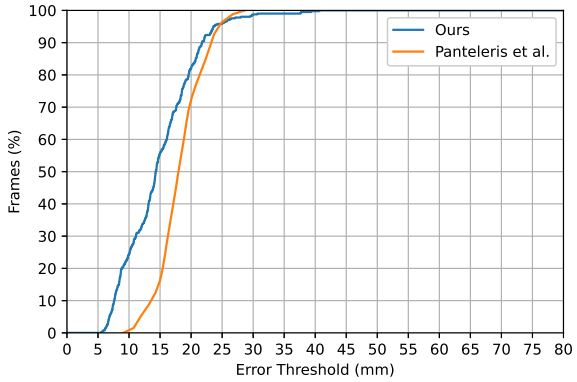

Figure 1: Percentage of correct frames on the HIC dataset.

Dexter+Object. This dataset is used to directly compare our method to the baseline method of Panteleris *et al.* [10].

Metrics: We evaluate our method on the above datasets using the metrics used by the state of the art methods for a direct comparison. We use the Percentage of Correct Keypoints (PCK) metric which describes how many keypoints are correctly detected when their distance from the ground truth is less than a threshold. We compute this metric for thresholds ranging from 20mm to 50mm and also measure the Area under Curve (AUC) for this range of thresholds. For direct comparison with Panteleris *et al.* [10], we use the Mean Joint Error metric and the percentage of frames that have a mean joint error less than a threshold. We also perform global alignment of the 3D joints and the ground truth. For [10] we use a *translation and rotation invariant* alignment based on [25] and for the rest of the methods we use a *translation invariant* alignment, aligning the centroids of the keypoints.

Quantitative evaluation: In Figure 1 we compare our method with the optimization method by Panteleris *et al.* [10] on the basis of the percentage of correct frames metric. The obtained results demonstrate that our method is preferable for lower error thresholds. For example, for the method of Panteleris *et al.* [10], 30% of the sequence frames have an error below 15mm while for our approach the corresponding percentage of frames is raised to almost 60%.

In Figure 2 we compare our method with state of the art methods on the Dexter+Object dataset [16]. It is clear that our method performs favourably compared to the optimization method by Boukhayma *et al.* [2] and competitive with the learning-based state of the art.

Similarly, in Figure 3 we measure the performance of our method on the EgoDexter dataset [7] and we see that it performs better than Boukhayma *et al.*[2] optimization method and competitively with the state of the art.

In Table 1 we provide the Area under the Curve (AUC) for each 3D PCK curve in Figures 2 and 3 for a more compact comparison of our approach to the state of the art methods.

Boukhayma *et al.* [2] optimization scheme performs worse than Panteleris *et al.* [10] and their learning based approach performs

Table 1: Comparison with the state of the art methods on DO and ED datasets.

AUC of PCK	**Ours**	Zhou [24]	Zhang [23]	Baek [1]	Xiang [21]	Boukhayma [2]	Iqbal [3]	Spurr [14]	Mueller [6]	Z&B [25]
Dexter+Object	0.764	0.948	0.825	0.650	0.912	0.763	0.672	0.511	0.482	0.573
EgoDexter	0.563	0.811	-	-	-	0.674	0.543	0.466	-	0.552

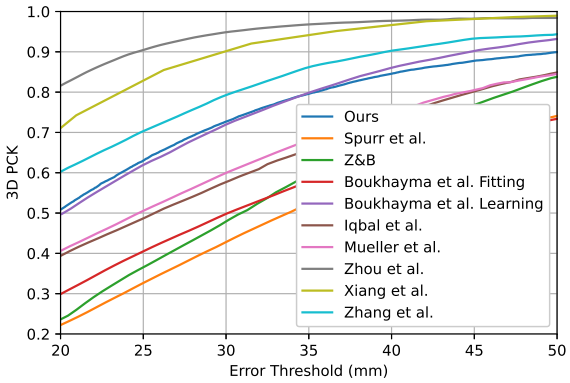

Figure 2: 3D Percentage of Correct Keypoints for Dexter+Object dataset.

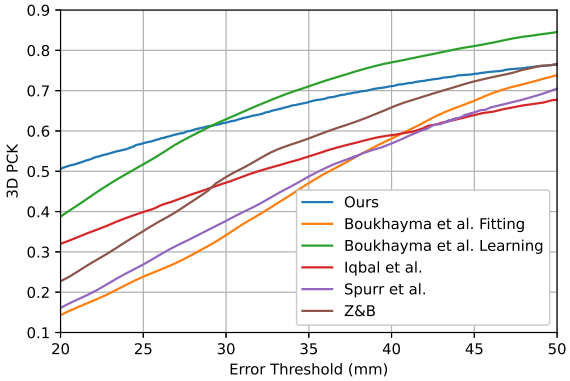

Figure 3: 3D Percentage of Correct Keypoints for EgoDexter dataset.

better. Our method outperforms their fitting method and performs similarly with the learning-based method that they proposed. The reason for this is that in order to produce plausible hand poses, they use a regularization term on the MANO pose PCA coefficients, which are a lossy representation of the pose space, but the anatomical joint limits that we use seem to limit more effectively the pose space while maintaining its whole representational power. The datasets chosen for evaluation are not used in training by any of the learning based methods thus getting a fair comparison with the state of the art.

Table 2: Robustness testing of our method on the HIC dataset. Optimization is performed either for pose+shape (1st column) or for pose alone (2nd column) with either an optimal shape (1st row) or a noisy shape initialization (2nd, 3rd, 4th row).

Mean Joint Error (mm)	Pose+Shape fitting	Pose fitting, only
Optimal Shape	14.73	14.59
Optimal Shape + $\mathcal{N}(0,1)$	14.69	15.04
Optimal Shape + $\mathcal{N}(0,2)$	14.92	15.58
Optimal Shape + $\mathcal{N}(0,3)$	14.89	16.93

We also perform a robustness testing of our method on the HIC dataset. More specifically, we evaluate the pose estimation error of our approach when both hand shape and pose is estimated (1st column of Table 2) and when a fixed hand model is used (2nd column of Table 2). This is performed for three different hand models, the ground truth one, as well as two noisy variants of it which is contaminated with increasing Gaussian noise (the three rows of Table 2, respectively). As it can be verified, when both shape and pose are optimized, the method performs satisfactorily even when the original hand model is quite noisy. On the contrary, if shape is not estimated, pose estimation becomes increasingly more inaccurate as a function of the discrepancy of the hand model from the actual one.

Qualitative evaluation: In Figure 4 we show example hand pose+shape estimations in frames of the EgoDexter (ED) [7], Dexter+Object (DO) [16] and the Hands in Action (HIC) [20] datasets. It can be verified that accurate estimations can be obtained even with strong occlusions, provided that the 2D keypoints are correctly detected in the image.

Limitations: The accuracy of the proposed approach depends highly on the accuracy of the estimated 2D keypoints. If several 2D keypoints are missing, e.g., due to occlusions, then the computed 3D hand pose will be invalid. This limitation could be addressed by adopting an occlusion-aware 2D keypoint detector, perhaps by passing, along with the image, the 2D mask of the object occluding the hand. Moreover, the 2D data is far more easily obtained than 3D ground truth and an even more powerful 2D keypoint detector could be created to directly improve the 3D accuracy as well. Another limitation is that, in some cases, an ambiguity is not pruned by the joint limits allowing the optimizer to converge, for example, to either positive or negative depth rotation of joints, in particular for joints near the tip of a finger. Some of these ambiguities could be pruned if the 2D mask of the hand is used in the optimization as well along with the 2D keypoints.

Hands In Action [20]

Dexter+Object [16]

EgoDexter [7]

EgoDexter [7]

Figure 4: Qualitative results from HIC, DO and ED datasets.

5 SUMMARY

We presented a 3D hand shape and pose estimation method on RGB images that was based on the minimization of the 2D keypoint reprojection error using the MANO hand model. As was verified by several experiments on standard well-established datasets, the accuracy of the proposed method is competitive to the state of the art and superior to the accuracy of other optimization approaches. Future work can be directed to further improving the 2D keypoint estimation accuracy, leveraging the ease of obtaining 2D data over 3D, or incorporating other 2D evidence cues to the optimization such as holding object and hand masks in order to have some constraint about the occluded parts of the hand holding an object, and directly improve the 3D hand tracking accuracy as well.

ACKNOWLEDGMENTS

This research work was partially supported by the Hellenic Foundation for Research and Innovation (HFRI) under the "1st Call for HFRI Research Projects to support Faculty members and Researchers and the procurement of high-cost research equipment", project I.C.Humans, number 91. This work was also partially supported by the Greek Secretariat for Research and Innovation and the EU, Project SignGuide: Automated Museum Guidance using Sign Language T2EDK-00982 within the framework of "Competitiveness, Entrepreneurship and Innovation" (EPAnEK) Operational Programme 2014-2020.

REFERENCES

[1] Seungryul Baek, Kwang In Kim, and Tae-Kyun Kim. 2019. Pushing the Envelope for RGB-Based Dense 3D Hand Pose Estimation via Neural Rendering. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019), 1067–1076.

[2] Adnane Boukhayma, Rodrigo de Bem, and Philip HS Torr. 2019. 3d hand shape and pose from images in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 10843–10852.

[3] Umar Iqbal, Pavlo Molchanov, Thomas Breuel, Juergen Gall, and Jan Kautz. 2018. Hand Pose Estimation via Latent 2.5 D Heatmap Regression. *arXiv preprint arXiv:1804.09534* (2018).

[4] Jiefeng Li, Chao Xu, Zhicun Chen, Siyuan Bian, Lixin Yang, and Cewu Lu. 2021. HybrIK: A Hybrid Analytical-Neural Inverse Kinematics Solution for 3D Human Pose and Shape Estimation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2021), 3382–3392.

[5] Alexandros Makris and Antonis Argyros. 2015. Model-based 3D Hand Tracking with on-line Shape Adaptation. In *Proceedings of the British Machine Vision Conference (BMVC)*, Mark W. Jones Xianghua Xie and Gary K. L. Tam (Eds.). BMVA Press, Article 77, 12 pages. https://doi.org/10.5244/C.29.77

[6] Franziska Mueller, Florian Bernard, Oleksandr Sotnychenko, Dushyant Mehta, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2018. GANerated Hands for Real-Time 3D Hand Tracking from Monocular RGB. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*. 11 pages. https://handtracker.mpi-inf.mpg.de/projects/GANeratedHands/

[7] Franziska Mueller, Dushyant Mehta, Oleksandr Sotnychenko, Srinath Sridhar, Dan Casas, and Christian Theobalt. 2017. Real-Time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor. *2017 IEEE International Conference on Computer Vision (ICCV)* (2017), 1163–1172.

[8] Markus Oberweger, Paul Wohlhart, and Vincent Lepetit. 2015. Hands Deep in Deep Learning for Hand Pose Estimation. *ArXiv* abs/1502.06807 (2015).

[9] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A. Argyros. 2011. Efficient model-based 3D tracking of hand articulations using Kinect. In *British Machine Vision Conference*.

[10] Paschalis Panteleris, Iasonas Oikonomidis, and Antonis A. Argyros. 2017. Using a Single RGB Frame for Real Time 3D Hand Pose Estimation in the Wild. *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2017), 436–445.

[11] Javier Romero, Dimitrios Tzionas, and Michael J. Black. 2017. Embodied Hands: Modeling and Capturing Hands and Bodies Together. *ACM Trans. Graph.* 36, 6, Article 245 (nov 2017), 17 pages. https://doi.org/10.1145/3130800.3130883

[12] Toby Sharp, Cem Keskin, Duncan Robertson, Jonathan Taylor, Jamie Shotton, David Kim, Christoph Rhemann, Ido Leichter, Alon Vinnikov, Yichen Wei, Daniel Freedman, Pushmeet Kohli, Eyal Krupka, Andrew Fitzgibbon, and Shahram Izadi. 2015. Accurate, Robust, and Flexible Real-Time Hand Tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (Seoul, Republic of Korea) *(CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3633–3642. https://doi.org/10.1145/2702123.2702179

[13] Tomas Simon, Hanbyul Joo, Iain Matthews, and Yaser Sheikh. 2017. Hand Keypoint Detection in Single Images using Multiview Bootstrapping. In *CVPR*.

[14] Adrian Spurr, Jie Song, Seonwook Park, and Otmar Hilliges. 2018. Cross-modal Deep Variational Hand Pose Estimation. In *CVPR* (Salt Lake City, USA).

[15] Srinath Sridhar, Franziska Mueller, Antti Oulasvirta, and Christian Theobalt. 2015. Fast and robust hand tracking using detection-guided optimization. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2015), 3213–3221.

[16] Srinath Sridhar, Franziska Mueller, Michael Zollhoefer, Dan Casas, Antti Oulasvirta, and Christian Theobalt. 2016. Real-time Joint Tracking of a Hand Manipulating an Object from RGB-D Input. In *Proceedings of European Conference on Computer Vision (ECCV)*. 17 pages. http://handtracker.mpi-inf.mpg.de/projects/RealtimeHO/

[17] Jonathan Taylor, Vladimir Tankovich, Danhang Tang, Cem Keskin, David Kim, Philip Davidson, Adarsh Kowdle, and Shahram Izadi. 2017. Articulated Distance Fields for Ultra-Fast Tracking of Hands Interacting. *ACM Trans. Graph.* 36, 6, Article 244 (nov 2017), 12 pages. https://doi.org/10.1145/3130800.3130853

[18] Anastasia Tkach, Mark Pauly, and Andrea Tagliasacchi. 2016. Sphere-Meshes for Real-Time Hand Modeling and Tracking. *ACM Trans. Graph.* 35, 6, Article 222 (dec 2016), 11 pages. https://doi.org/10.1145/2980179.2980226

[19] Jonathan Tompson, Murphy Stein, Yann Lecun, and Ken Perlin. 2014. Real-Time Continuous Pose Recovery of Human Hands Using Convolutional Networks. *ACM Trans. Graph.* 33, 5, Article 169 (sep 2014), 10 pages. https://doi.org/10.1145/2629500

[20] Dimitrios Tzionas, Luca Ballan, Abhilash Srikantha, Pablo Aponte, Marc Pollefeys, and Juergen Gall. 2016. Capturing Hands in Action using Discriminative Salient Points and Physics Simulation. *International Journal of Computer Vision (IJCV)* (2016). http://files.is.tue.mpg.de/dtzionas/Hand-Object-Capture

[21] Donglai Xiang, Hanbyul Joo, and Yaser Sheikh. 2019. Monocular total capture: Posing face, body, and hands in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

[22] Baowen Zhang Yangang Wang and Cong Peng. 2019. SRHandNet: Real-time 2D Hand Pose Estimation with Simultaneous Region Localization. *IEEE Transactions on Image Processing* 29, 1 (10 2019), 2977 – 2986. https://doi.org/10.1109/TIP.2019.2955280

[23] Xiong Zhang, Qiang Li, Wenbo Zhang, and Wen Zheng. 2019. End-to-End Hand Mesh Recovery From a Monocular RGB Image. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)* (2019), 2354–2364.

[24] Yuxiao Zhou, Marc Habermann, Weipeng Xu, Ikhsanul Habibie, Christian Theobalt, and Feng Xu. 2020. Monocular Real-Time Hand Shape and Motion Capture Using Multi-Modal Data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[25] Christian Zimmermann and Thomas Brox. 2017. Learning to Estimate 3D Hand Pose from Single RGB Images. In *IEEE International Conference on Computer Vision (ICCV)*. https://lmb.informatik.uni-freiburg.de/projects/hand3d/ https://arxiv.org/abs/1705.01389.