VNU Journal of Science:
Computer Science and Communication Engineering
Journal homepage: http://www.jcsce.vnu.edu.vn/index.php/jcsce

Original Article

Robustify Hand Tracking by Fusing Generative and Discriminative Methods

Nguyen Duc Thao[1], Nguyen Viet Anh[2], Le Thanh Ha[1], Ngo Thi Duyen[1,*]

[1]*VNU University of Engineering and Technology, 144 Xuan Thuy, Cau Giay, Hanoi, Vietnam*
[2]*AI Academy Vietnam, 489 Hoang Quoc Viet, Bac Tu Liem, Hanoi, Vietnam*

Received 14 August 2020
Revised 04 September 2020; Accepted 04 September 2020

Abstract: With the development of virtual reality (VR) technology and its applications in many fields, creating simulated hands in the virtual environment is an e ective way to replace the controller as well as to enhance user experience in interactive processes. Therefore, hand tracking problem is gaining a lot of research attention, making an important contribution in recognizing hand postures as well as tracking hand motions for VR's input or human machine interaction applications. In order to create a markerless real-time hand tracking system suitable for natural human machine interaction, we propose a new method that combines generative and discriminative methods to solve the hand tracking problem using a single RGBD camera. Our system removes the requirement of the user having to wear to color wrist band and robustifies the hand localization even in di cult tracking scenarios.

Keywords: Hand tracking, generative method, discriminative method, human performance capture.

1. Introduction

Hand tracking is a fundamental research topic and has been widely studied for decades because of its wide range of applications. The exact reconstruction of the shape and articulation of the human hand is one of the particularly important questions when solving that problem. Virtual reality technology has become more popular in recent years and moreover become an e ective way to enhance the experience as well as the sense of presence and immersion. To make the interaction more "realistic", there are many ways to optimize the input mechanism such as using controllers, keyboards. However, the fully articulated hand tracking is at the top of the expectation. Although recent works have focused on creating systems that allow hand tracking in real-time, we cannot deny that extracting hand motions is still a challenge with many factors such as fast movements, data noises, and self-occlusions [1]. These challenges require us to constantly explore more powerful methods of recreating the hand model, increasing input clarity.

* Corresponding author
E-mail address: duyennt@vnu.edu.vn

https://doi.org/10.25073/2588-1086/vnucsce.261

Approaching by the input devices, we can categorize existing works in the field of hand tracking into the invasive and the non-invasive. For the invasive works, using gloves for recording hand pose is the most representative that directly reconstruct the hand model from gloves data [2, 3]. These methods provide high accuracy but are di cult to deploy for the reasons of complexity in calibration, cost of expensive commercial gloves and user movement impediment due to additional hardware. Therefore, non-invasive methods are being developed in recent years, towards systems that require less cost, easy to deploy and make further improvement in flexibility.

With non-invasive works, these methods mainly use imaging devices or depth sensors to implement tracking algorithms. These works can be divided into two main classes: discriminative (appearance-based) methods and generative (model-based) methods. Discriminative methods build a machine learning model that relies on large amounts of data to automatically detect hand pose from each individual frame. Without the need of relying on temporal coherence, discriminative methods are still successfully employed for real-time hand tracking [4]. In another approach, generative (model-based) methods apply iterative model-fitting optimization to sensor information. These methods describe the hand with an articulated model and typically minimizes the discrepancy between the data synthesized from the model and the data observed by the sensor [5].

With such complex tasks, most non-invasive works in the past have made a number of assumptions that greatly simplify the problem of detecting or tracking a hand. However, the above generative methods still face certain limitations such as:

i) Delimitation of the wrist: In most previous works, the hand in the image (video) can be easily segmented from the arm by wearing a colored wristband [5, 6]. Besides, if a person is wearing long non-skin coloured sleeves shirt, the extraction of the hand silhouette is much simplified [7]. However, in reality, these solutions are easily broken when a person wears a shirt that has the same color as the wristband or when the background behind the person is cluttered. In addition to that, the preparation of a wristband also makes hand tracking not very flexible;

ii) Using static background: A blank or uniform background will allow the hand to be segmented more easily [7]. However, using a static background is considered too restrictive for a general system, and the technique of background subtraction is unreliable when there are strong shadows in the scene, when background objects move or if the illumination changes, especially with hand tracking systems in reality;

Figure 1. In complex mixed backgrounds, previous methods [5] might misrecognize the wrist position.

iii) Some works localize and segment the hand using a learned pixel-wise classifier [8, 9]. This approach does not fully exploit the context information of other body parts, therefore it has limited robustness and has di culty in disambiguating left and right hands of the same person.

Although the restrictions listed above only occur in some certain circumstances, they are still considered to be limited in general cases. For example, with a given input image (video), no hands are identified (Figure 1) or it is impossible to guarantee whether the hand is located correctly. This proves that a stronger method of hand recognition is needed. Moreover, the process to extract hands in some

generative methods is still sketchy, requiring wrist determination in many practical cases.

In this paper, we present a method with the synergy of a discriminative method and a generative method to recover articulated hand motions. The body part locations in the discriminative method will be used to remove the dependence of the wrist position in the generative method. We aim to achieve more accurate and robust tracking results and overcome some drawbacks of existing works. The contributions of this paper are as follows:

i) A new approach for hand localization and segmentation in hand tracking, therefore removes the wristband preparation and wristband segmentation process;

ii) We fully disclose our source code to ensure reproducibility of our results and facilitate future research in this domain: https://github.com/thaond98/robust-htrack.

The paper is organized as follows. Section 2 provides the background on the two mainstream methods which are the base for developing our hand tracking system. The proposed method is described in Section 3, we will explain the synergy of the generative and discriminative methods which improves upon existing works [5]. In Section 4, we qualitatively analyze the performance of our hand tracking system and discuss the results. Section 5 concludes the paper and opens the discussion for future works.

2. Related Works

As mentioned, studies on hand tracking problem can be divided into two main categories: generative and discriminative methods. In this section, we present an overview of these two methods, which is the basis for proposing our new solution.

2.1. Generative Method

The generative method describes a model of the hand and search for the optimal solution in the model's continuous parameter space by minimize the energy functions which are defined to measure the fit between the model and observed data. The optimal solution comes from local optimization around the estimate for the previous frame.

A typical study in the generative method is the work of Tagliasacchi et al [5]. The system is called htrack, and it is the foundation for us to propose new solutions. Htrack is a method for real-time capturing articulated hand poses and motions using a single RGBD sensor. The system is based on a real-time registration process that fits a 3D articulated hand model to depth images to accurately reconstruct hand poses. The most common problems for generative methods are a good enough input preprocessing and initialization, an expressive and e cient enough hand model, and an objective function that minimizes the error between the 3D hand model and the observed data.

Initialization: There are many initialization methods that provide an alignment for the first frame. Some works [10] initializes by extracting the sensor color image and performing a skin color segmentation. Besides, [6] also initializes by fingertip detection. Htrack [5] detects a color wristband by color segmentation. After that, it gets the 3D points in the proximity of the wristband and compute the principal axis. The hand point cloud is segmented by conjunction of this axis and the wristband centroid. Any depth pixel within the hand point cloud is labelled as belonging to the silhouette image as shown in Figure 2. However, the method of Tagliasacchi et al. [5] has some drawbacks such as being prone to the color similarity between the wristband and nearby objects of the same color, or it can be impractical for those gestures which contain di cult hand orientations, and it requires inconvenient wristband preparation.

Figure 2. Htrack's wrist localization and hand segmentation [5].

Tracking model: The articulated hand model serves as the mean of the fitting procedure and the presentation of tracking results [6]. It also encodes geometric priors for shape completion and constrains hand morphology. Human hands are highly articulated and therefore require models with su ciently many degrees of freedom to adequately describe the corresponding motion space. In order to model the hand, there are a variety of options depending on the balance required between accuracy and performance. There are some hand models have been proposed. Melax et al. [11] used a union of convex bodies for hand tracking. Qian et al. [6] built the hand model using a number of spheres. Tagliasacchi et al. [5] found that a hand model using only cylinder primitives works well for tracking in terms of accuracy and e ciency. They register a template cylinder hand model to the sensor data with: 26 degrees of freedom, 6 for global rotation and translation and 20 for articulation. Not only the model can be quickly adjusted to the user by specifying global scale, palm size and finger lengths, but also this one can be used to drive a high-resolution skinned hand because they compute joint angles (including rigid transformation) in the widespread BVH motion sequence format (Figure 3).

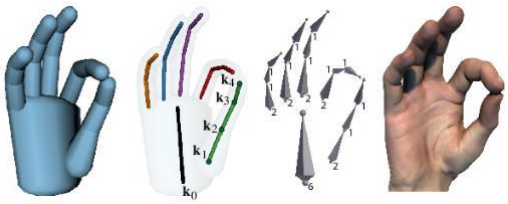

Figure 3. Htrack's hand tracking model. Left to right: the cylinder model used for tracking, the skeleton, the BVH skeleton exported to drive the rendering, the rendered hand model [5].

Objective function: The objective function measures the discrepancy between the hand model and input data, as well as the validity and plausibility of the hand pose [6]. The objective function is composed of data fitting terms and prior terms in general. With *htrack* [5], Tagliasacchi et al. also formulate this goal as a minimum of the following objective function:

$$\min_{\theta} \underbrace{E_{3D} + E_{2D} + E_{wrist} +}_{fitting\,term}$$
$$\underbrace{E_{pose} + E_{kinematic} + E_{\text{temporal}}}_{prior\,term}$$

where E_{3D}, E_{2D} *and* E_{wrist} are the alignment energies corresponding to 3D point cloud, 2D silhouette and wrist joint. In *htrack* [5], authors use publicly available database of recorded hand poses to build a low-dimensional subspace of plausible poses. Then they enforce the hand parameters to closely match with this subspace using the projection energy E_{pose}. Kinematic prior is used to deal with hand poses that have unrealistic joint angle limit. $E_{kinematic}$ is kinematic prior energy, including $E_{collision}$ and E_{bounds} where $E_{\text{collision}}$ is an energy that accounts for the inter-penetration between each pair of cyninders in the hand model while E_{bounds} helps preventing the hand from overbending the joint to impossible postures. Temporal prior helps with jitter in hand motion, therefore increase smoothness. $E_{temporal}$ is an energy penalizing the velocity and acceleration of points attached to the kinematic chain.

Let F be the sensor input data consisting of a 3D point cloud X_s and 2D silhouette S_s (see Fig. 1). Given M a 3D hand model with joint parameters $\theta = \{\theta_1, \theta_2, ..., \theta_{26}\}$, they aim at recovering the pose θ of the user's hand that matches the sensor input data F . Fitting terms measure how well the hand parameters explain the input frames F and prior terms regularize the solution to produce realistic hand poses.

2.2. Discriminative Method

Discriminative method aims to extract the hand poses directly by train a classifier or a regressor to map image features of hand appearance to hand poses. Over the past few years, many works based on discriminative method have been developed for hand tracking. Approaches based on nearest neighbor search [12], decision trees [13], or convolutional

networks [14], or CNN models [15, 16] have demonstrated that appearance-based methods can be successfully employed for real-time hand tracking. They are usually implemented with machine learning and require a large amount of training data to automatically detect the position of joints in each frame.

A typical study in the discriminative method is that of Cao et al [4]. The system is called OpenPose, and it is the foundation for us to propose new solutions. OpenPose provides a real-time method for multi-person 2D pose estimation based on its bottom-up approach. With their various related research works, they can extend their work into the real-time multi-person system to jointly detect human body, hand, facial, and foot key-points on single images. The pipeline of the system is depicted in Figure 4.

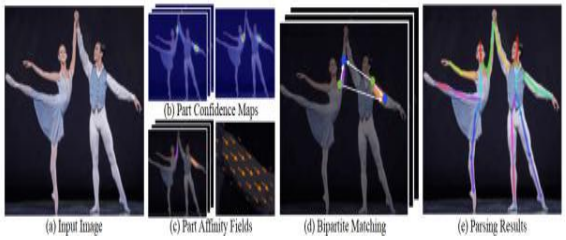

Figure 4. Overall pipeline of OpenPose system [4].

First, they let the frame pass through the first 10 layers of VGG-19 model to extract feature maps. Then, the feature maps are processed with multiple CNN stages to predicts the confidence maps (Figure 4b) of di erent body parts location and predicts the a nity fields (Figure 4c), which represents a degree of association between di erent body parts. Finally, the confidence maps and part a nity fields are processed by a greedy algorithm to obtain the poses for each person in the image.

The greedy algorithm which is used to parse poses of multiple people from confidence maps and part a nity fields can be summarized as:

Step 1: Find all body part locations using the confidence maps.

Step 2: Find which body parts go together to form pairs using the part a nity fields and joints in step 1.

Step 3: Associate limbs that belong to the same person and get the final list of human poses.

Multi-people parsing and tracking are not really important in our goal, and the result when identifying a key point belonging to the main person becomes easier when there is only one person in the scene (Figure 5).

Figure 5. Using OpenPose to detect body and hand keypoints.

3. Proposed Method

Robust hand tracking requires a strategy for hand localization and segmentation during tracking, as well as finding hand when tracking is lost. This is highly important for human computer interaction applications, in which hand motion can be fast and it is di cult to accurately predict the motion. As mentioned in the related works, color-based approaches such as [5] are easily broken in cluttered background or lighting changes. We address this issue by building a combination method based on the use of OpenPose [4] and htrack [5]. The method currently only allows one person to stand in the background and it is able to track one hand. The combination follows the pipeline in Figure 6.

An RGB image and its corresponding depth image are taken from the depth sensor. The RGB image will go through OpenPose to get the wrist position as coordinates on the color image. The wrist data from the depth image will be used to retrieve the 3D point cloud of the hand and the 2D distance transform of its silhouette. An energy minimization problem with the iterative closest point procedure will be solved to best align the hand model to the processed data.

3.1. Acquisition Device

Microsoft Kinect V2 is one of the most popular depth sensors because of its price and image quality. Unfortunately, this device is not supported by htrack. We therefore build a new module in htrack to retrieve the Kinect sensor data using libfreenect 2 as the driver. The Kinect V2 has a depth resolution of 512x424 pixels, a color resolution of 1920x1080 pixels, acquired at 30 fps. For real-time performance, we down-sample the color image by a factor of 2 and synthesize a new depth image in the view of the color camera to bring the sensor data into a single coordinate system. The system will make use of both color and depth data for tracking hand poses.

3.2. Wrist Localization and Hand Segmentation

We use OpenPose C++ API and its pre-trained model BODY 25 to detect 25 body key-points of possibly multiple people on the color image. The network inference is run on GPU in parallel with the energy optimization process of the previous frame. By detecting whole body key-points, we fully exploit the context information of other body parts to infer the wrist location and can easily disambiguate left and right hands. This whole body information is also extremely helpful if we want to build a whole body 3-D performance capture system including body, face, and hands.

We use the wrist location of the main user to replace the color-based wristband identification and segmentation. The hand silhouette is extracted by retrieving the wrist depth value and all depth information in a fixed depth range around the wrist. After calculating the axis of the first principle component at the wrist position, we will crop a circle above the axis to identify the hand as well as the 2D hand silhouette. This way, we remove the requirement of wearing a wristband and performing a color calibration at the start.

4. Results

Our system is evaluated on a laptop running Ubuntu 16.04 with Intel Core i7 3.5GHz CPU and NVIDIA GTX 1060 GPU. Similar to htrack, we perform 1 iteration for rigid alignment and 7 full iterations for articulation alignment, at 4.5ms per iteration. We run body key point detection, closed form closest point correspondences and Jacobian computation for the fitting energies on GPU. Hand pose estimation with detected wrist on a frame is shown in Figure 7.

Since our main contribution is the removal of the wrist band requirement and its color calibration by using discriminative body key point detection, it is enough to compare qualitatively our system against the original htrack implementation under the same tracking conditions. We come up with the following evaluation scenarios:

The user does not wear a color wrist band. The user wears a color wrist band and the tracking performance is compared between our method and the previous one htrack [5] in various tracking scenarios.

- The background is normal.

- The background has a similar color as the wristband.

- There are the lighting changes during the tracking.

- The background has an object whose color is roughly the same as the wristband color.

- The background is cluttered.

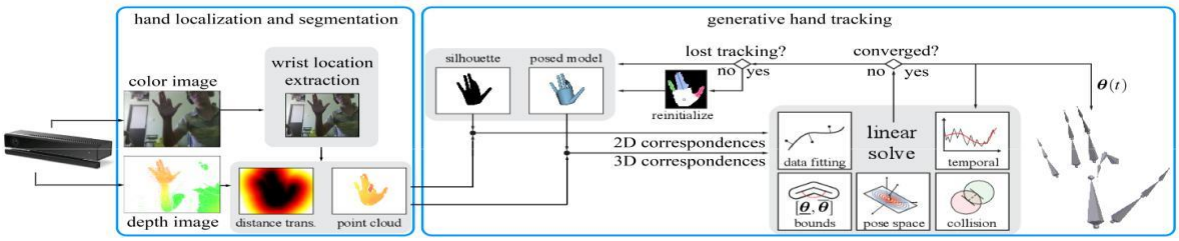

Figure 6. Our improved hand tracking pipeline upon [5].

Figure 7. The tracking is stable in complex backgrounds, even when colorful objects are around.

When the user does not wear a wristband, that means the hand recognition of the original htrack is disabled, as expected htrack fails to identify the hand as shown in Figure 8. However, our method can still track the hand successfully in various scenarios, see Figure 9.

Figure 8. htrack [5] does not work without a wristband.

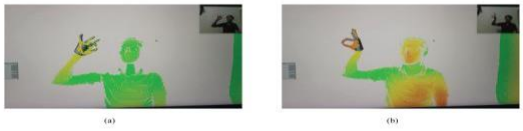

Figure 9. Tracking results of our system in no-wristband scenarios.

We then qualitatively show our real-time performance when the user wears a wristband in Figure 10, with a normal background. We can see that our system performs well on various hand poses and the rendered hand model looks almost the same as the input depth image of the hand. Compared with the results of htrack, the preprocessing and initialization through OpenPose gives equivalent results. The first column shows the result of htrack, while the second one shows the result of the new method. All results are displayed in our computer in real-time. We also present the robustness of our system with occlusion in Figure 10, third and forth columns. With the occlusion, flipping and pointing down, our system still recognizes the wrist in such cases. Those qualitative comparisons also show comparable performance with the existing generative method [5].

In the scenario that the background color matches the color of the wristband as shown in Figure 11, the old method reveals its limitation. It misidentifies the wristband position when considering the background as an object that is likely to be wristband. However, in this scenario, our proposed method can still find exactly the position of the wrist, and our system still works normally.

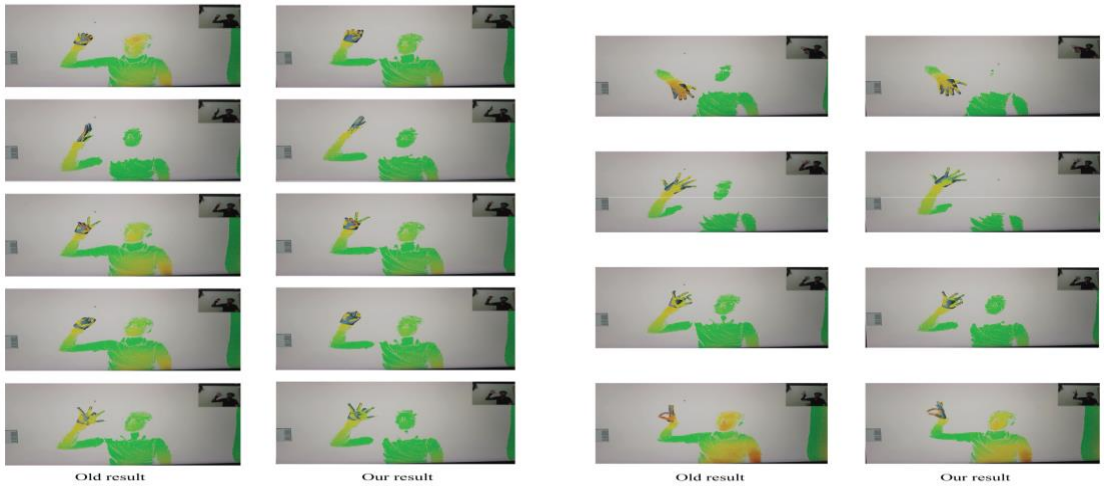

Figure 10. Tracking results of our system in comparison with htrack when the user wears a wrist band, with the normal background.

In the scenario in which the lighting is changed continuously, the hand recognition of Htrack is unstable, apparently due to the sudden lighting changes causing the color information and the depth information to be abnormally altered, the wrist center calculation becomes misleading, see Figure 12. Thanks to the discriminative wrist detection, our method is not negatively a ected by lighting changes and the system still works well in this condition.

In the case of having an object with a wristband-like color in the background, proposed method demonstrates better results in di erent hand tracking scenarios. When using a complex background or when not using a wristband, our method yields more stable results. Besides, the computational time of the proposed method is e ective enough to enable the system to work real time. Currently our method only allows one person to stand in the background. We can however extend our method to track person identity and support hand tracking even when multiple people are present in the scene.

Figure 11. Tracking results of our system in comparison with *htrack* when the background has a similar color as the wristband.

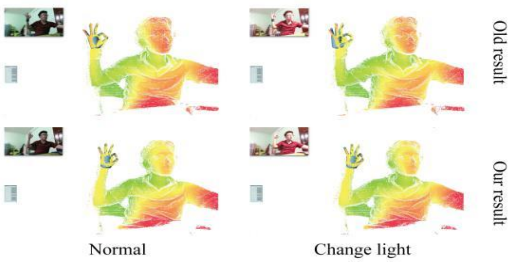

Figure 12. Tracking results of our system in comparison with htrack when there are lighting changes.

Figure 13. Tracking results of our system in comparison with *htrack* when there is an object in a wristband-like color in the background.

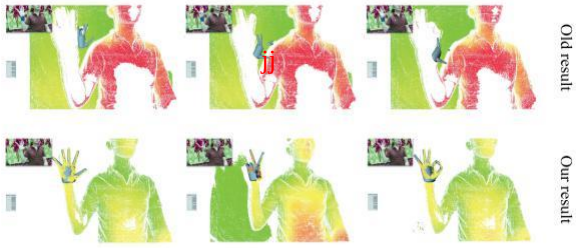

Figure 14. Tracking results of our system in comparison with htrack, with the cluttered background.

5. Conclusions

In this paper, we contributed a new method that combines the strengths of the two existing methods for real-time hand tracking. Our approach makes use of the discriminative method OpenPose's pose estimation [4] to provide preprocessing and initialization for the generative hand tracking method htrack [5].

The synergy helps to remove the requirement of a color wristband and the color calibration step. Our system demonstrates good performance in various tracking scenarios and complex backgrounds.

In the future work, we aim to use the discriminative method to intervene more deeply in the optimization process of the generative method, meaning that for each finger joint location received from OpenPose, we will adjust the hand model's joints so that it aligns accordingly. It is expected to improve the tracking performance in di cult poses such as rotating fist, which lacks of depth features when being acquired by current commodity depth sensors.

Acknowledgments

This work has been supported by VNU University of Engineering and Technology under project number CN18.08.

References

[1] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, Tamaddon, A. Heloir, D. Stricker, Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth, CoRR abs/1808.09208, 2018.
URL http://arxiv.org/abs/1808.09208

[2] O. Glauser, S. Wu, D. Panozzo, O. Hilliges, Sorkine-Hornung, Interactive hand pose estimation using a stretch-sensing soft glove, ACM Trans, Graph. 38(4) (2019) 1-15.

[3] L. Jiang, H. Xia, C. Guo, A model-based system for real-time articulated hand tracking using a simple data glove and a depth camera, Sensors 19 (2019) 4680. https://doi.org/10.3390/s19214680.

[4] Z. Cao, G. Hidalgo, T. Simon, S. Wei, Y. Sheikh, Openpose: Realtime multi-person 2d pose estimation using part a nity fields, CoRR abs/1812.08008, 2018.

[5] A. Tagliasacchi, M. Schroder, A. Tkach, S. Bouaziz, M. Botsch, M. Pauly, Robust articulated-icp for real-time hand tracking, Computer Graphics Forum 34, 2015.

[6] C. Qian, X. Sun, Y. Wei, X. Tang, J. Sun, Realtime and robust hand tracking from depth, in: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[7] Tomasi, Petrov, Sastry, 3d tracking = classification + interpolation, in: Proceedings Ninth IEEE International Conference on Computer Vision 2 (2003) 1441-1448.

[8] T. Sharp, C. Keskin, D. Robertson, J. Taylor, J. Shotton, D. Kim, C. Rhemann, I. Leichter, A. Vinnikov, Y. 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, 2015, pp. 3633-3642.

[9] S. Sridhar, F. Mueller, A. Oulasvirta, C. Theobalt, Fast and robust hand tracking using detection-guided optimization, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[10] I. Oikonomidis, N. Kyriazis, A.A. Argyros, Tracking the articulated motion of two strongly interacting hands, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 2012, pp. 1862-1869.

[11] S. Melax, L. Keselman, S. Orsten, Dynamics based 3d skeletal hand tracking, CoRR abs/1705.07640, 2017.

[12] R. Wang, S. Paris, J. Popovic, 6d hands: Markerless hand tracking for computer aided design, 2011, pp. 549-558.
https://doi.org/10.1145/2047196.2047269.

[13] D. Tang, T. Yu, T. Kim, Real-time articulated hand pose estimation using semi-supervised transductive regression forests, in: 2013 IEEE International Conference on Computer Vision, 2013, pp. 3224-3231.

[14] M. Oberweger, P. Wohlhart, V. Lepetit, Generalized feedback loop for joint hand-object pose estimation, 2019, CoRR abs/1903.10883.
URL http://arxiv.org/abs/1903.10883.

[15] J. Malik, A. Elhayek, F. Nunnari, K. Varanasi, K. Tamaddon, A. Heloir,´ D. Stricker, Deephps: End-to-end estimation of 3d hand pose and shape by learning from synthetic depth, 2018, pp. 110-119. https://doi.org/10.1109/3DV.2018.00023.

[16] A. Mohammed, J.L.M. Islam, A deep learning-based end-to-end composite system for hand detection and gesture recognition, Sensors 19 (2019) 5282. https://doi.org/10.3390/s19235282.