Integrated Driver Pose Estimation for Autonomous Driving

Xiao Cao[1,2], Wei Hu[2][a] and Hui Liu[3][b]
[1]Shanghai Electric Automation Group, 200023, Shanghai, China
[2]School of Mechanical and Aerospace Engineering, Nanyang Technological University, 639798, Singapore
[3]Cognitive System Lab, University of Bremen, 28359, Bremen, Germany

Keywords: Computer Vision, Autonomous Driving, Driver Posture Estimation, Human-Machine Interaction.

Abstract: Human-machine interaction, especially driver posture estimation is important to the development of autonomous driving, which can facilitate safe and smooth driving behaviours. Besides, it also contributes to ergonomics research and human-machine interaction design for automated vehicles. The existing studies have got great achievements in body estimation, hand pose estimation, and even face feature estimation thanks to the rapid development of deep learning approaches and the upgrade of hardware equipment. However, most existing models can only process body estimation or hand estimation separately, which will impede the research on driver-vehicle interaction in autonomous driving. This is because the driving process is highly dependent on the cooperation between the body and hands behaviours. In this study, five popular deep learning models, including Simple Faster R-CNN, RootNet, PoseNet, Yolo v3, and graph convolutional neural network, are combined through a cascade method to develop an integrated model which can estimate body and hand simultaneously during the driving process. The coordinate transform system is proposed to connect models in series. Experiment results demonstrate the proposed method can produce 2D and 3D reorganization of the human body and hands simultaneously with acceptable accuracy.

1 INTRODUCTION

Autonomous driving exhibits rapid progress in recent years due to its substantial application value and potential societal implications. Human pose estimation technology is crucial in autonomous driving, particularly with the growing possibility of automated vehicles navigating congested roads, which allows for instantaneous tracking of driver motion, enhancing driver requirements, and identifying potential safety risks. Many advanced functions including operation simplification, fatigue detection, and behaviour analysis can be developed based on driver gesture research. Besides, autonomous driving can be made to resemble human driving to the greatest extent possible through observing and recording driver behaviours.

The advancements in photograph acquisition technologies and deep learning approaches have led to significant progress in human pose estimation technology, which has been implemented in many domains like security systems and smart payment. However, for autonomous driving, most implementations only focus on specific body parts like the body or hand, which ignores the correlation and coordination among different human body parts during driving behaviour and results in limited progress in driver behaviour studies.

This research aims to develop an integrated model to estimate hands and body simultaneously by deploying the proposed cascade method on 5 mainstream computer vision models. Subsequently, 2D and 3D skeleton diagrams have been generated and the accuracy of the proposed method has been verified. However, due to the scarcity of public datasets on whole-body, the performance of the developed model can only be evaluated by visualization.

The rest of this article is organized as follows: Section 2 introduces the existing works in relevant fields. Section 3 presents the principles and pipeline of the proposed method. Section 4 illustrates the details of experiments and results while the conclusion and discussion are presented in Section 5.

[a] https://orcid.org/0000-0001-7058-939X
[b] https://orcid.org/0000-0002-6850-9570

Cao, X., Hu, W. and Liu, H.
Integrated Driver Pose Estimation for Autonomous Driving.
DOI: 10.5220/0012639400003657

In *Proceedings of the 17th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2024) - Volume 1*, pages 695-702
ISBN: 978-989-758-688-0; ISSN: 2184-4305

2 LITERATURE REVIEW

2.1 Human Detection

Human detection models are designed to identify the presence and location of humans in images or video frames, it is derived from objection models, which have been developed for decades. Current popular objection models contain Region-based Convolutional Neural Networks (R-CNN) and You Only Look Once (Yolo) families.

R-CNN family is the most popular framework for object detection and includes several categories like basic R-CNN, Fast R-CNN, and Mask R-CNN. R-CNN has founded the basis for the current region-based object detection methods (Girshick et al., 2014) and the main idea is to select a certain number of regions of interest to conduct the image classification randomly or empirically. The limitation is that the region size warp process may damage the original information and result in unexpected errors and unsatisfying accuracy. To address the issue, Spatial Pyramid Pooling in Deep Convolutional Networks (SPPnet) was developed utilizing grids meshing and features concatenating approach (He et al., 2014), and Fast R-CNN employed this approach to construct the Region of Interest (RoI) pooling layer, which brought the object detection into a new era. Besides, faster R-CNN introducing region proposal network (RPN) rather than conventional selective search to produce region proposals (Mueller et al., 2017), which improves the time-efficiency and facilities the development of multi-scale detection. Mask R-CNN is a convenient and flexible general object instance segmentation neural network (He et al., 2017), which can not only realize object detection but also generate segmentation results for each target.

Recently, the Yolo methodology also attracted the attention of researchers because of its excellent performance. Unlike R-CNN methods, Yolo v1 (Redmon et al., 2016) treats the object detection task as a regression problem instead of region detection. The main difference between Yolo and R-CNN families is that global information can be analysed rather than local information from sliding windows or region proposals approach. This allows for the acquisition of highly generalized features, which outperform previous object detection algorithms and can be migrated to related fields. Though the initial Yolo model had some limitations, such as lower accuracy and speed compared to some state-of-the-art object detection models, which were addressed in subsequent versions. Yolo v2 and v3 utilized improved network architectures and advanced training techniques, such as batch normalization and residual connections, to enhance detection accuracy. At present, Yolo v3 is a well-respected algorithm considering both maturity and training performance (Gkioxari et al., 2018). The Yolo family comprises a range of object detection algorithms that are well-known for their remarkable processing speed. Additionally, the CornerNet approach, which relies on key point-based object detection, has also demonstrated high efficiency and accuracy (Law & Deng, 2018). Notably, the recently proposed CornerNet-Lite, an improved version of CornerNet, has achieved both higher speed and superior performance compared to Yolo v3 (Hei Law, Yun Teng, Olga Russakovsky, 2019).

2.2 Body Pose Estimation

The estimation results of human body estimation models are always represented by the several key points on a specific skeleton and the methodology is roughly divided into three categories: 3D pose tracking, 2D-3D pose lifting, and pose regression from images. As the models selected in this project are all based on deep learning and neural network, the dissertation would focus on the last two methods, especially the deep learning-based methods. Pavllo et al.(Pavllo et al., 2019) processed the detected key points by a fully convolutional architecture that is compatible with the 2D joints detector to predict the coordinates effectively, while in (Ge et al., 2019) the 3D pose estimation is treated as a regression problem of Euclidean Distance Matrices (EDM) to capture more information about pairwise correlations between key points.

Pose Regression from Image can overcome the inherent ambiguity generated by encoding and decoding between 2D pose estimation and 2D to 3D lifting. Mehta et al.(Mehta et al., 2020) developed an estimation model to evaluate the level of similarity between the target 3D pose and the input image. Zhou et al.(Zhou et al., 2016) treated a kinematic object model as the prior knowledge in the neural network to optimize the articulated object pose estimation.

2.3 Hand Estimation

Hand estimation has great significance in the development of human-computer interaction with a long development period. Generally, hand estimation can be divided into three categories: discriminative approach, generative approach and hybrid approach. The discriminative method processes the image and predicts the pose of the hand from the image directly, while the generative method prepares a hand model previously and tries to match the hand model to the input image, and the Hybrid method is the combination of two approaches (Barsoum, 2016). Barsoum (Tompson et al., 2014) created labelled ground-truth data and developed the Pose Recovery model to estimate human hands from single-depth images. Oikonomidis et al.(Iasonas Oikonomidis, Nikolaos Kyriazis, 2011) treated the hand estimation as an optimization problem where the hand model parameters had to be determined to minimize the error between the preprepared models and the processed input image. Besides, Oikonomidis et al.(Oikonomidis et al., 2011) resented a similar method, where the discrepancy is quantified between the actual features and predicted features extracted from the observation and then minimized to the expected value by improving the parameters. Finally, the improved parameters are decoded to obtain the 3D hand pose.

2.4 Dataset

Benchmark datasets with ground truth annotations are critically important but the dataset establishment is usually difficult (Erol et al., 2007). Some popular datasets related to the human body and hands are listed below.

Table 1: Body datasets.

Dataset	Description
HumanEva (Mehta et al., 2018)	4 people, 8×10^4 samples, Marker-based MoCap in indoor
Human3.6M (von Marcard et al., 2018)	11 people, 360×10^4 samples, Marker-based MoCap in indoor
Total Capture (Sharp et al., 2015)	5 people, 190×10^4 samples, Marker-based MoCap along with IMUs in indoor
MPI-INF-3DHP (Barsoum, 2016)	8 people,130×10^4 samples, Marker less MoCap in both indoor and outdoor
3DPW (Oikonomidis et al., 2011)	5 people, 5×10^4 samples, 3D human poses captured with IMUs in outdoor

Table 2: Hands datasets.

Dataset	Description
Hand-Object Interaction (Hamer et al., 2010)	Hand-Object, rigid & articulated objects, 60 sequences, 10 objects shapes
ETHZ (Ballan et al., 2012)	Hand-Hand/Hand-Object, rigid & articulated objects, 7 sequences
Hands in Action (Tzionas et al., 2016)	Hand-Hand/Hand-Object, rigid & articulated objects, 29 sequences with a large variety of interactions
Dexter & Object (Sridhar et al., 2016)	Hand-Object, rigid objects, simple object shape(cube), 6 sequences with 2 actors and with 2 objects shapes
EgoDexter (Mueller et al., 2017)	Hand-Object, rigid & articulated objects, 4 sequences with 4 actors, various objects and cluttered background

3 METHOD

3.1 Overview

To realize the estimation of the body and hand simultaneously by one integrated model, the following pipeline (Fig. 1) has been designed: At first, the image is processed by Fast RCNN and Yolo v3 models to determine the bounding boxes for the human body and hands, respectively. Simultaneously, the original image is fed into RootNet to predict the root depth, which represents the absolute distance between the human and the camera. Subsequently, based on the bounding boxes, the image is cropped to isolate the corresponding sections comprising the human body and hands. Then PoseNet and HandNet models are utilized to estimate the body and hands using cropped images and root depth, and then their outputs are decoded to obtain 2D estimation and 3D skeleton of the hand and body. Finally, the integrated results are generated utilizing the overlapping and the coordinate transformation approach.

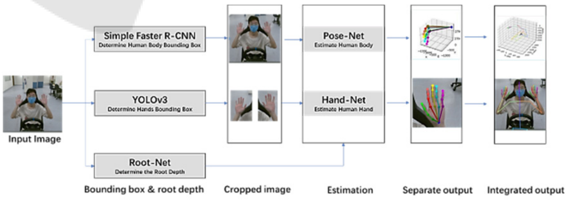

Figure 1: The pipeline of the proposed model.

3.2 Bounding Box

3.2.1 Body Bounding Box

The body bounding box should be generated before the estimation to clear the object for the subsequent model, which can avoid the error caused by the difference between the size of the image and the human. In this research, the Simple Faster R-CNN

(Ren et al., 2017) model is utilized to identify the bounding box in the image, which contains a fully convolutional network for feature map generation and a regional proposal network for processing. The processed feature is fed into a box-regression layer and a box-classification layer, and then the original image is cropped based on the bounding box parameters to get the image of the human.

3.2.2 Hand Bounding Box

Yolo v1 is utilized to predict large-size objects, whereas Yolo v2 and Yolo v3 are better suitable for medium and small-size objects. So, the Yolo v3 (Redmon & Farhadi, 2018) is utilized to predict the hands-bounding box, which is composed of the backbone and Darknet Building Block (DBL). The backbone contains the convolutional and residual network for object features extraction, while the DBL is composed of convolutional layers, batch normalization, and activation layers, which are used to predict the object and generate the bounding box.

3.3 Root Depth

The root depth estimation model is used to predict the camera-centred coordinate of key points of humans from the cropped image of people processed by the detection model. And the RootNet (Moon et al., 2019) proposed by Moon et al. has been utilized in this section. The network contains three modules for feature extraction, coordinate estimation, and depth estimation. The loss function is defined as follows:

$$L_{root} = \|R - R^{*}\|_1 \tag{1}$$

Where R is the predicted root depth while R^{*} represents the ground-truth root depth.

3.4 Estimation Net

3.4.1 PoseNet

The input is the cropped image based on the body bounding box and the integral regression method (Sun et al., 2018) is applied. It contains the backbone modules for feature extraction and poses estimation for 3D heatmap generations. The PoseNet is trained by minimizing L_1 distance between groundtruth coordinates and the predicted results. The loss function is defined as follows:

$$L_{pose} = \frac{1}{J}\left\|P_j^{rel} - P_j^{rel*}\right\|_1 \tag{2}$$

Where represents P_j^{rel} predicted coordinates and P_j^{rel*}represents the ground-truth coordinates.

3.4.2 HandNet

Based on the hand bounding box, the image is cropped to get two hand images relatively and they are fed into a hand estimation network to get the 3D and 2D results. Graph convolutional neural network (Ge et al., 2019) has been selected as the method. Firstly, the image passes through a two-stacked hourglass network to extract the feature maps and 2D heat maps, which are then processed and encoded as a latent feature vector by a residual network. Secondly, the latent feature is put into a Graph CNN to predict the 3D coordinates of mech vertices. Finally, the 3D hand pose is linearly regressed from the 3D hand mesh. The pose loss function is defined as follows:

$$L_J = \sum_{j=1}^{J}\left\|\varphi_j^{3D} - \hat{\varphi}_j^{3D}\right\|_2^2 \tag{3}$$

Where φ_j^{3D} denotes the ground-truth 3D joint locations while $\hat{\varphi}_j^{3D}$ is the estimated 3D joint locations.

3.5 2D and 3D Integration

The 2D or 3D skeleton diagram of the body and hands are generated by pose and hand estimation models respectively. For the 2D integration, the 2D outputs can be directly achieved by overlapping the 2D body and 2D hands results based on common joint points. For 3D integration, the hand coordinates system can be transformed into the body coordinate system through a linear transformation with at least three sets of coordinates. Two common key points, the root of the hand and middle finger can be utilized as the first two sets and the root of the index finger is chosen to be the third set of coordinates, which has been contained in the hand coordinate system. The information of the root of the index finger in the body coordinate system can be predicted by rotating the coordinate of the root of the middle finger by 15° counterclockwise or clockwise on the palm plane. The rotation matrix is shown in Eq. 4, where $R_i(\theta_i)$ represents rotation matrix with θ_i rotation angle around axis (x,y,z).

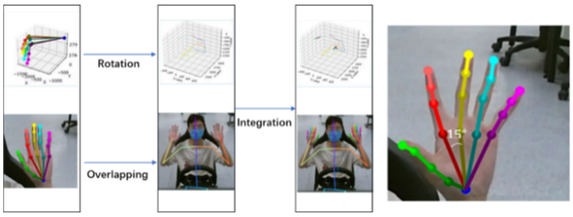

Figure 2: Transformation pipeline and hand estimation.

$$R_x(\theta_x) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & cos\theta_x & -sin\theta_x \\ 0 & sin\theta_x & cos\theta_x \end{bmatrix}$$

$$R_y(\theta_y) = \begin{bmatrix} cos\theta_y & 0 & sin\theta_y \\ 0 & 1 & 0 \\ -sin\theta_y & 0 & cos\theta_y \end{bmatrix} \qquad (4)$$

$$R_z(\theta_z) = \begin{bmatrix} cos\theta_z & -sin\theta_z & 0 \\ sin\theta_z & cos\theta_z & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

The coordinate of the root of the index finger (I) can be computed by Eq. 5, where R_i is rotation transformations (x,y,z) and M means coordinate of the root of the middle finger.

$$I = R_i * M \qquad (5)$$

Then, the transformation matrix can be calculated by the three sets of 3D coordinates. Assume that the three sets of coordinates in hand and body coordinate systems are represented as X_i ($X = A, B, C; i = hand\ or\ body\ coordinate\ system$). Suppose that all these coordinates are raw matrix, and the inverse matrix is as Eq. 6 while the transformed coordinates of the hand coordinate system can be derived by Eq. 7, where J_i represent the joint coordinates of the hand or body coordinate system.

$$Inv\ matrix \cdot \begin{bmatrix} A_{hand\ coordinate\ system} \\ B_{hand\ coordinate\ system} \\ C_{hand\ coordinate\ system} \end{bmatrix} \qquad (6)$$
$$= \begin{bmatrix} A_{body\ coordinate\ system} \\ B_{body\ coordinate\ system} \\ C_{body\ coordinate\ system} \end{bmatrix}$$

4 EXPERIMENT

4.1 Equipment

An experiment platform has been utilized to simulate the driving process, which contains three display screens, two monitors, a seat, and a steering wheel (Fig.3). The steering wheel (Logitech G29) can generate realistic force feedback, making the details of the driver's hand movements much more realistically when driving while the monitors are used to record the driver posture.

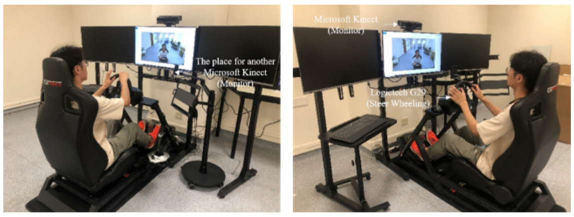

Figure 3: Experiment environment.

4.2 Implementation Environment

This project was conducted mainly in Python based on the Ubuntu system. The following main open-source libraries were installed in a virtual environment: Python3.8, OpenCV, Scikit-image, tqdm, fire, pprint, Pillow, Keras, Pytorch, torchvision, cpython, ipdb, numpy, scipy, yacs, Matplotlib.

4.3 Result and Discussion

Unlike the hand estimation and body estimation models which have enough datasets to evaluate their performance, we have not found public datasets that evaluate the performance of the model estimating hand and body simultaneously has been created.

Hence, the integrated model utilises 1126 images extracted from a video collected by the Microsoft Kinect and the model performance is evaluated by the results observation of these images. The performance is divided into 3 levels. Level 1, both 2D and 3D estimation are perfect, which means that the predicted 2D joint locations are matched with the origin 2D image, and 3D output is evaluated manually as reasonable without considering accurate error. Level 2, the 2D estimation is perfect but the 3D estimation has some problems. For example, the hand is squeezed into a line, the hand size is problematic, and the pose or gesture is unreasonable. Level 3, the 2D estimation and 3D estimation are both unexpected, which means the predicted joint points of 2D outputs are not fitted to the actual joint points correctly. Finally, 88% of results are in level 1 and level 2, which is acceptable accuracy and performance. The output of the integrated model contains estimations of hand, body pose, and the combination of hand and body pose in 2D and 3D formats. Only the combination of hand and body pose is shown because of our research focus. Some perfect visualization results are shown in Fig. 4 while some problematic results are shown in Fig. 5-7.

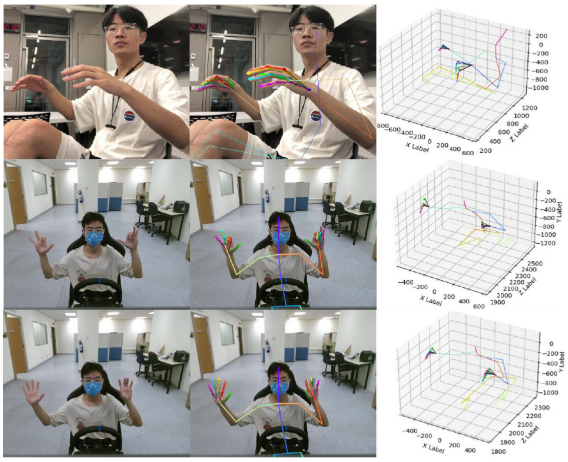

Figure 4: The visualization of results.

In Fig. 5 the 2D outputs are perfect, where the joints of both hand and body are predicted exactly. However, the right hand of 3D outputs is squeezed into a line. One possible reason is that the created joints have an error due to the incorrect rotation direction and angle. Another factor is that the root of the hand of these two models is not matched. From 2D outputs, it's clear that the key points of the hand root failed to match, which means the same issues in 3D outputs.

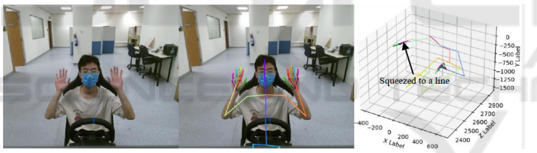

Figure 5: The result with the hand squeezed into a line.

Fig. 6 illustrates the unreasonable size of the 3D output, especially the human hands. The 3D outputs of hand parts are nearly shrunk to one point. As the output of the hand estimation model is correct, the potential problem is from the rotation transformation. The essential factors of this problem are similar to the estimated hand squeezed into a line.

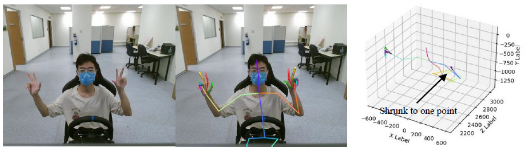

Figure 6: The result with the hands shrunk to one point.

The third main problem is the incorrect estimation. From Fig. 7, the hand joints of 2D output are unacceptable and the key point representing the hand root is located outside of the hand in 3D outputs. The main factor is that the selected hand estimation model failed to predict the joints, which means that the robustness of the model is not perfect.

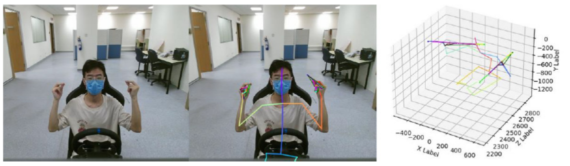

Figure 7: The result with incorrect hand estimation.

In conclusion, despite the high proportion of feasible outputs, it does exist some unexpected results, especially 3D output. There are three main reasons: 1. The robustness of the selected model is not perfect, which means some postures cannot be recognized successfully. 2. Some errors may exist in the coordination of hand estimation and body estimation. 3. The rotation transformation is not precise enough, especially for some complex gestures.

5 CONCLUSION

This paper proposed an integrated method based on five existing models to achieve the estimation of body and hands simultaneously and the model performance and potential problems are analysed based on the experiment. Besides, human body poses, and hand estimation-related techniques and models have been reviewed. Drive estimation is truly one of the most important topics in autonomous driving, and an important problem is that there are no publicly available datasets for the whole-body including details of hand, body pose and face, which means that there is no authoritative and recognized evaluation method to measure the performance of the integrated model. Despite this study having defined a simple evaluation standard, it is based on manual observation which is not strict and persuading enough. Hence, the public and recognized evaluation criteria are necessary.

In the future, more accurate rotation transformation methods or other approaches should be developed to avoid the inconsistency between the key points of hand estimation and body pose estimation. Besides, more models should be integrated to create a new model to estimate the human body, hands, feet, face, and other parts of the body simultaneously. More importantly, the publicly available datasets catering for the whole-body estimation and an evaluation method should be created.

REFERENCES

Ballan, L., Taneja, A., Gall, J., Van Gool, L., & Pollefeys, M. (2012). Motion capture of hands in action using discriminative salient points. *Proceedings of European Conference on Computer Vision*, *7577 LNCS*(PART 6), 640–653. https://doi.org/10.1007/978-3-642-33783-3_46

Barsoum, E. (2016). *Articulated Hand Pose Estimation Review*. 1–50. http://arxiv.org/abs/1604.06195

Erol, A., Bebis, G., Nicolescu, M., Boyle, R. D., & Twombly, X. (2007). Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, *108*(1–2), 52–73. https://doi.org/10.1016/j.cviu.2006.10.012

Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., & Yuan, J. (2019). 3D hand shape and pose estimation from a single RGB image. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 10825–10834. https://doi.org/10.1109/CVPR.2019.01109

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 580–587. https://doi.org/10.1109/CVPR.2014.81

Gkioxari, G., Girshick, R., Dollár, P., & He, K. (2018). Detecting and Recognizing Human-Object Interactions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *1*(c), 8359–8367. https://doi.org/10.1109/CVPR.2018.00872

Hamer, H., Gall, J., Weise, T., & Van Gool, L. (2010). An object-dependent hand pose prior from sparse training data. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 671–678. https://doi.org/10.1109/CVPR.2010.5540150

He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask R-CNN. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 2980–2988. https://doi.org/10.1109/ICCV.2017.322

He, K., Zhang, X., Ren, S., & Sun, J. (2014). Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *8691 LNCS*(PART 3), 346–361. https://doi.org/10.1007/978-3-319-10578-9_23

Hei Law, Yun Teng, Olga Russakovsky, J. D. (2019). *CornerNet-Lite : Efficient Keypoint-Based Object Detection*.

Iasonas Oikonomidis, Nikolaos Kyriazis, and A. A. A. (2011). Markerless and Efficient 26-DOF Hand Pose Recovery. *Proceedings of the 10th Asian Conference on Computer Vision*, *6978 LNCS*(PART 1), 365–373. https://doi.org/10.1007/978-3-642-24085-0_38

Law, H., & Deng, J. (2018). CornerNet. *European Conference on Computer Vision(ECCV)*, 765–781.

Mehta, D., Rhodin, H., Casas, D., Fua, P., Sotnychenko, O., Xu, W., & Theobalt, C. (2018). Monocular 3D human pose estimation in the wild using improved CNN supervision. *Proceedings - 2017 International Conference on 3D Vision, 3DV 2017*, 506–516. https://doi.org/10.1109/3DV.2017.00064

Mehta, D., Sotnychenko, O., Mueller, F., Xu, W., Elgharib, M., Fua, P., Seidel, H. P., Rhodin, H., Pons-Moll, G., & Theobalt, C. (2020). XNect: Real-time Multi-Person 3D Motion Capture with a Single RGB Camera. *ACM Transactions on Graphics*, *39*(4), 1–24. https://doi.org/10.1145/3386569.3392410

Moon, G., Chang, J. Y., & Lee, K. M. (2019). Camera distance-aware top-down approach for 3D multi-person pose estimation from a single RGB image. *Proceedings of the IEEE International Conference on Computer Vision*, *2019-Octob*, 10132–10141. https://doi.org/10.1109/ICCV.2019.01023

Mueller, F., Mehta, D., Sotnychenko, O., Sridhar, S., Casas, D., & Theobalt, C. (2017). Real-Time Hand Tracking under Occlusion from an Egocentric RGB-D Sensor. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 1163–1172. https://doi.org/10.1109/ICCV.2017.131

Oikonomidis, I., Kyriazis, N., & Argyros, A. (2011). *Efficient model-based 3D tracking of hand articulations using Kinect*. *June 2014*, 101.1-101.11. https://doi.org/10.5244/c.25.101

Pavllo, D., Feichtenhofer, C., Grangier, D., & Auli, M. (2019). 3D human pose estimation in video with temporal convolutions and semi-supervised training. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2019-June*, 7745–7754. https://doi.org/10.1109/CVPR.2019.00794

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016-Decem*, 779–788. https://doi.org/10.1109/CVPR.2016.91

Redmon, J., & Farhadi, A. (2018). *YOLOv3: An Incremental Improvement*. http://arxiv.org/abs/1804.02767

Ren, S., He, K., Girshick, R., & Sun, J. (2017). Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *39*(6), 1137–1149. https://doi.org/10.1109/TPAMI.2016.2577031

Sharp, T., Keskin, C., Robertson, D., Taylor, J., Shotton, J., Kim, D., Rhemann, C., Leichter, I., Vinnikov, A., Wei, Y., Freedman, D., Kohli, P., Krupka, E., Fitzgibbon, A., & Izadi, S. (2015). Accurate, robust, and flexible realtime hand tracking. *Conference on Human Factors in Computing Systems - Proceedings*, *2015-April*, 3633–3642. https://doi.org/10.1145/2702123.2702179

Sridhar, S., Mueller, F., Zollhöfer, M., Casas, D., Oulasvirta, A., & Theobalt, C. (2016). Real-time joint tracking of a hand manipulating an object from RGB-D input. *International Journal of Computer Vision*, *9906*

LNCS, 294–310. https://doi.org/10.1007/978-3-319-46475-6_19

Sun, X., Xiao, B., Wei, F., Liang, S., & Wei, Y. (2018). Integral human pose regression. *Proceedings of the European Conference on Computer Vision (ECCV)*, *11210 LNCS*, 536–553. https://doi.org/10.1007/978-3-030-01231-1_33

Tompson, J., Stein, M., Lecun, Y., & Perlin, K. (2014). Real-time continuous pose recovery of human hands using convolutional networks. *ACM Transactions on Graphics*, *33*(5). https://doi.org/10.1145/2629500

Tzionas, D., Ballan, L., Srikantha, A., Aponte, P., Pollefeys, M., & Gall, J. (2016). Capturing Hands in Action Using Discriminative Salient Points and Physics Simulation. *International Journal of Computer Vision*, *118*(2), 172–193. https://doi.org/10.1007/s11263-016-0895-4

von Marcard, T., Henschel, R., Black, M. J., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering accurate 3D human pose in the wild using IMUs and a moving camera. *Proceedings of the European Conference on Computer Vision (ECCV), 2018*, *11214 LNCS*, 614–631. https://doi.org/10.1007/978-3-030-01249-6_37

Zhou, X., Sun, X., Zhang, W., Liang, S., & Wei, Y. (2016). Deep kinematic pose regression. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *9915 LNCS*(March 2017), 186–201. https://doi.org/10.1007/978-3-319-49409-8_17