Accurate Camera Pose Estimation from Learned Point Features: A Case Study*

Tomasz Nowak[0000−0002−2635−7732]

Institute of Robotics and Machine Intelligence
Poznan University of Technology, ul. Piotrowo 3A, 60-965 Poznań, Poland
tomasz.nowak@doctorate.put.poznan.pl

Abstract. This paper provides a case study of using a recent deep learning based approach to keypoint detection on images in the task of camera pose estimation. The application context is assisted docking to a charging station with an electric bus using monocular vision. We examined the influence of three factors on the achieved results: the backbone network, the size of the final activation maps generated by the network, and the number of convolutional layers in the keypoint head. The proposed configurations were evaluated to find the best trade-off between pose estimation accuracy (2D translation and the yaw angle estimation error were measured) and the computational complexity. The evaluation dataset was gathered using a real bus, during different weather conditions, and the ground truth data was provided by a Differential GPS. The result presented in this paper shows that proposed by us modifications of architecture can improve the accuracy of the whole processing system.

Keywords: Camera pose estimation · keypoints · deep learning.

1 Introduction

We consider a scenario of using monocular vision to guide the driver of an electric bus while docking to a charging station. An important prerequisite for successful planning of such a maneuver is to localize the bus accurately with respect to the charger's head and its supporting pylon. The bus has only a monocular camera mounted to the roof, which has to be localized with respect to some predefined points of the charger's structure. Therefore, once the charger gets detected [2], the task reduces to estimation of the pose of the camera (2D position and the yaw angle – orientation) with respect to the detected keypoints. Unlike the direct approach to camera pose estimation [7], we proposed in [3] to solve this problem using a two-stage procedure. Firstly, the keypoints are detected by a modified Faster R-CNN [4] neural network, which is also used to detect the entire charger. Secondly, a transformation between the camera frame and the charger's coordinate frame is estimated by solving the optimization problem which minimizes reprojection error upon the known locations of the keypoints

* This work was supported by PUT internal grant 0214/SBAD/0235

on the image, the real 3-D positions of those points obtained from a 3-D model of the charger, and the calibrated camera parameters [1].

This paper provides a case study of employing the recent deep neural architecture for keypoint detection: High Resolution Network (HRNet, [5]) for estimation of the keypoints in the considered task. HRNet was proposed mainly within the application context of human body pose estimation. We use this architecture in an entirely different application, where the accuracy of keypoints location is crucial. The aim of our study is to choose the HRNet configuration that yields the most accurate keypoints, maintaining also a reasonable computation complexity of the neural model.

2 Structure of the Proposed Solution

The aim of our deep learning model is to extract the keypoints selected on the charger's head and pylon. The selected points must be located in an appropriate way to provide good conditioning for the camera pose estimation problem. Hence, we selected four corners of the charger, which are harder to detect than the fiducials used in [3], but there is no need to modify the charger's appearance. The exact locations of the selected keypoints are shown in Fig.1A.

The neural network for the detection of keypoints consists of a backbone block that extracts feature maps from the image and the keypoint head which generates heatmaps from feature maps. The investigated HRNet architecture's backbone is designed to maintain a high-resolution representation of features through the whole network. Additionally, the unbiased data processing methods described in [6] were used to provide an accurate estimation of subpixel keypoints locations. The keypoint head in the default implementation has only one convolutional layer which outputs n heatmaps, where n is the number of points to be predicted. By adding a transposed convolutional layer with stride=2 (deconv) before the

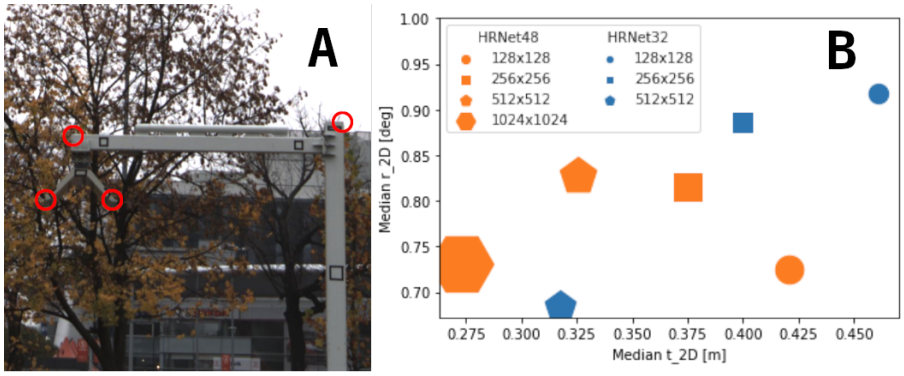

Fig. 1. Keypoints used for pose estimation (A). Translation and rotation errors for HRNet32 and HRNet48 with different heatmap size configurations. The size of marker corresponds to the number of operations required to inference single image (B).

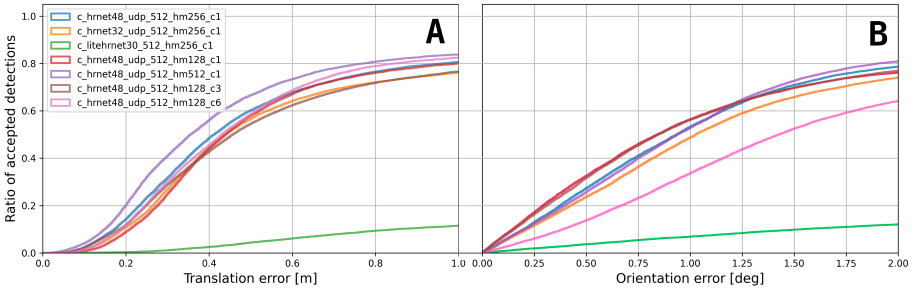

Fig. 2. Cumulative distribution functions of 2D translation error (A) and orientation error (B) for the evaluated architectures.

final convolutional layer it is possible to increase the resolution of the returned heatmap, which by default has four times smaller resolution than the input image. Extra convolutional layers followed by batch normalization and ReLU activation can be added between deconv layers and the final convolutional one to enhance the network's ability to accurately estimate the location of keypoints.

3 Experiments

The purpose of the presented experiments was to examine the influence of the neural network architecture on the 2-D pose estimation accuracy and the computational complexity. All experiments were performed off-line on a custom dataset recorded using a real electric bus and charger with the ground truth poses obtained using DGPS. The dataset consists of 81 sequences gathered over 5 days. The diversity of data was achieved by different maneuver starting points, different bus trajectories, and weather conditions. To consider the detection of a keypoint as accepted, the RMSE of the 3-D point projected on the image should be less than 10 pixels. We use three metrics to compare neural network architectures in the evaluation procedure: median of the 2-D translation error, the median of the yaw angle estimation, and the percentage of accepted detections. We selected three areas for the potential improvements in the HRNet model: the selection of an appropriate backbone network for the feature extraction, the size of the returned heatmaps, and the number of the convolutional layers in the keypoint heatmap head. For all experiments, the image input size was set to 512×512 pixels.

3.1 Backbone

In the following experiments, three backbone networks from the family of the HRNet were evaluated: HRNet48, HRNet32, and LiteHRNet. For all configurations, the heatmap size was 256×256 pixels and no extra convolutional layers were added to the keypoint head. The results show that using a bigger backbone head reduces both translation and rotation error. Moreover, it improves also the network's capability to identify all keypoints on the image (Tab. 1).

Table 1. Comparison of pose estimation errors and size of the network depending on the used backbone network, heatmap size and number of covolution layers in the head. HS means heatmap size and C depicts number of convolutional layers in the head

Configuration	Parameters [M]	Operations [GFLOPs]	Median t_2D [m]	Median r_2D [deg]	Percent of accepted detections
HRNet48 HS256 C1	63.79	87.44	0.3751	0.8148	0.944 %
HRNet32 HS256 C1	28.67	43.34	0.3998	0.8865	0.926 %
LiteHRNet HS256 C1	5.05	1.93	1.3306	2.3700	0.284 %
HRNet48 HS128 C1	63.60	84.10	0.4207	0.7242	0.937 %
HRNet48 HS256 C1	63.79	87.44	0.3751	0.8148	0.944 %
HRNet48 HS512 C1	64.84	156.56	0.3253	0.8279	0.941 %
HRNet48 HS128 C1	63.60	84.10	0.4207	0.7242	0.937 %
HRNet48 HS128 C3	63.60	84.18	0.4095	0.7199	0.937 %
HRNet48 HS128 C6	63.61	84.31	0.4702	1.3081	0.942 %

3.2 Heatmap Size

The second aspect which influences the pose estimation accuracy is the size of the output heatmaps. The default implementation of the keypoint detector based on HRNet returns heatmaps which are downsampled four times compared to the input image size, so using an image of 512×512 pixels results in 128×128 pixels heatmaps. The upsampling of the heatmaps is achieved using Transposed Convolutional layers. A single transposed convolutional layer increases the width and height of the heatmap twice. The compared configurations use HRNet48 as the backbone without extra convolutional layers in the head. The influence on the percentage of accepted detections is marginal, but increasing the heatmap size significantly reduces translation error. Further increasing the size of the heatmap significantly increases computational cost and makes the network prone to overfitting (Fig. 1B and Tab. 1).

3.3 Keypoint Head Depth

For the comparison of the influence of the number of convolutional layers in the head, we used a network with HRNet48 as the backbone and 128×128 pixels heatmaps. Increasing the number of convolutional layers to 3 improves the pose estimation accuracy with respect to both translation and rotation. However, a network with 6 layers performs worse than the smaller versions. This accuracy loss is attributed to the overfitting of the network (Fig. 2 and Tab. 1).

4 Conclusion

We demonstrated that it is possible to adopt the HRNet architecture to the task of keypoints detection for camera pose estimation, finally achieving better results than in [3], in spite of the lack of fiducials on the charger. In the considered application, the HRNet48 HS512 C1 configuration is the best choice as we do not have tight constraints on the computing resources while achieving the highest possible accuracy reduces the risk of failed maneuver. Further research will concern modifications to the loss function that should introduce geometric priors stemming from the known pattern of keypoints.

References

1. Hartley, R. I. and Zisserman, A.: Multiple view geometry in computer vision, Cambridge University Press (2004).
2. Nowak, T., Nowicki, M., Cwian, K., Skrzypczyński, P.: How to improve object detection in a driver assistance system applying explainable deep learning. IEEE Intelligent Vehicles Symposium, Paris, 226–231, (2019).
3. Nowak, T., Nowicki, M., Cwian, K., Skrzypczyński, P.: Leveraging object recognition in reliable vehicle localization from monocular images. Automation 2020: Towards Industry of the Future, AISC, Vol. 1140, Springer, 195–205, (2020).
4. Ren, S., He, K., Girshick, R. Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. NeurIPS, 91–99, (2015).
5. Wang, J., Sun, K., Cheng, T., *et al.*: Deep high-resolution representation learning for visual recognition. IEEE Trans. on PAMI, 43, 3349–3364, (2021).
6. Huang, J., and Zhu, Z., Guo, F., *et al.*: The Devil is in the Details: Delving into Unbiased Data Processing for Human Pose Estimation. arXiv:2008.07139, (2020).
7. Xiang, Y., Schmidt, T., Narayanan, V. Fox, D.: PoseCNN: a convolutional neural network for 6D object pose estimation in cluttered scenes. Proc. Robotics: Science and Systems, Pittsburgh, (2018).