Journal of Robotics, Networking and Artificial Life
Vol. 7(4); March (2021), pp. 249–252
DOI: https://doi.org/10.2991/jrnal.k.201215.008; ISSN 2405-9021; eISSN 2352-6386
https://www.atlantis-press.com/journals/jrnal

Research Article

Object 6 Degrees of Freedom Pose Estimation with Mask-R-CNN and Virtual Training

Victor Pujolle*, Eiji Hayashi

Computer Science and System Engineering, Kyushu Institute of Technology, 680-4 Kawazu, Iizuka 820-0053, Japan

ARTICLE INFO

Article History

Received 07 November 2019
Accepted 25 June 2020

Keywords

Pose estimation
deep-learning
keypoints localization
instance segmentation
virtual training
factory automation

ABSTRACT

Pose estimation algorithms' goal is to find the position and the orientation of an object in space, given only an image. This task may be complex, especially in an uncontrolled environment with several parameters that can vary, like the object texture, background or the lightning conditions. Most algorithms performing pose estimation use deep learning methods. However, it may be difficult to create dataset to train such kind of models. In this paper we developed a new algorithm robust to a high variability of conditions using instance segmentation of the image and trainable on a virtual dataset. This system performs semantic keypoints based pose estimation without considering background, lighting or texture changes on the object.

1. INTRODUCTION

The goal of the paper is to present a new framework to handle estimation of the 6 Degree of Freedom (6-DoF) pose (translation and rotation in 3D) of an object from a single image. It has many applications, especially in robotic manipulation. However, despite the need of a general and accurate framework, this task tends to be treated on a case by case basis. For example, approaches tend to differ depending of the object's texture.

Because 3D pose dataset are difficult and time consuming to make, training a network for general pose estimation in an uncontrolled environment with multiple object's instances is challenging. To tackle this problem, our approach combined previously existing pose estimation approach with a classic instance segmentation network. The algorithm can be divided in three stages. The first stage is the instance segmentation where we use a mask-R-CNN network [1] to provide semantic masks of objects' instances on an image. Because instance segmentation is already a well-studied subject this paper does not focus on this part. For this step, any state-of-the-art instance segmentation network would work, the better the accuracy, the better the final result. The second stage is the keypoints localization network. For this task we use a heavy convolutional network to predict of set of semantic keypoints [2]. This network takes as input the semantic masks extracted by the mask-R-CNN. Because the input is only a mask of the object, it is possible to train this network using virtual images generated randomly with a simple mesh of the object. This way, the creation of the virtual dataset does not require skills in 3D modelling to beat the reality gap [3].

The last stage is to solve the pose to perspective problem. Because it is already a well-documented subject, this paper does not focus on this step. While this work focuses only of RGB input, it could be adapted fairly easily to Red Green Blue Depth (RGBD) inputs with some light modification in the network's architecture and for the virtual dataset generation.

2. RELATED WORK

6 degree of freedom pose estimation is extremely useful for many robotic tasks and several approaches of the subject have already been studied. But these methods typically address the problem for highly textured objects. This leads to failure for textureless objects and reduces generalization capability. However, they need instance specific 3D model of every object, limiting their real applications. Methods using 2D landmarks localization on the image can works for textureless objects. Finding the pose of a rigid object given the position of n landmarks on the image is commonly referred as pose to perspective problem.

3. TECHNICAL APPROACH

3.1. Keypoints Localization

The keypoints localization network uses stacked convolutional encoder decoder that has been proven effective for human pose estimation using keypoints [4]. It seems natural to think that the same architecture could be used for object keypoints localization. The stacked encoder decoder are plugged directly at the end of an already trained mask-R-CNN network. Its task is to reduces objects to masks, removing background, textures and specific details.

*Corresponding author. Email: victor.pujolle5@gmail.com

Only the form of the object projected on the camera captor is left. This enables the keypoints localization network to consider only the form of the object and helps the generalization.

3.1.1. Network architecture

The network architecture (Figure 1) is mostly inspired of the architecture presented in Kang et al. [2], Tremblay et al. [3] and Hinterstoisser et al. [5] with few minor modifications. The network input is a mask extracted by the mask-R-CNN from an RGB image and the output is a set of heatmaps, one for each keypoint. The heatmap intensity at a given pixel indicates the confidence of the respective keypoint to be located at this pixel. The network is made of stacked hourglasses. Each hourglass consists of two processing stage. In the first stage, a series of convolution and max-pooling layers are applied the input image. Each max-pooling layer divides by two the resolution of the feature map. This process continues until the resolution reach a minimal resolution set by the user (4×4 in our model). After this down-sampling process, series of deconvolution layers are applied to the feature map. A residual layer is also applied to increase the accuracy of the network. This process continues until the feature map reaches the input resolution. Another hourglass can be stacked to the end of the first one to refine the output heatmaps. The ground truth labels used for the training are heatmaps made by applying 2D gaussian centered at every keypoint. The loss function used is the l2 norm. It is possible the use intermediate supervision at the end of each hourglass module the increase the accuracy and fasten the training by providing more signal for the gradient back propagation. The output of the last module is used as the output of the whole network. The maximum of each heatmap indicates the localization of the respective keypoint.

3.1.2. Design explanation

The main goal of this architecture is to provide a heatmaps with the same resolution of the input. The combination of down sampling and up sampling enable the network to use local and global features of the input. This is especially useful given the large variability of the object and the relatively low dimension of the output. The addition of intermediate supervision has been proven useful to increase gradient signal and therefore fasten the training process while reducing issues such as vanishing gradient. Finally, residual layers than save the signal during the down sampling are used to help the up sampling process.

3.2. Training

3.2.1. Virtual training

Training may become very challenging when no dataset is available for the task. In the case of pose estimation, most datasets focus on very few objects with limited number of instances in a very controlled environment. Because labelling data for pose estimation is extremely time consuming, virtual training has been proven efficient to train state-of-the-art pose estimation algorithms. However, creating realistic scenes and 3D models of objects can also be very challenging. In this paper we use a virtual dataset created using only a simple mesh of every objects. This approach makes possible to create important dataset in a limited amount of time. To do so we compute mask projection of the object with random poses as illustrated in Figure 2. This mask projection is made using OpenCV functions for projection and drawing. We also apply different kinds of blur to simulate the imperfection of the mask-R-CNN's outputs. These images are similar to the output of an instance segmentation network; therefore, they can be used to train the keypoints localization network. Because the mask-R-CNN is easy to train with a good accuracy, we consider only the training of the hourglass network.

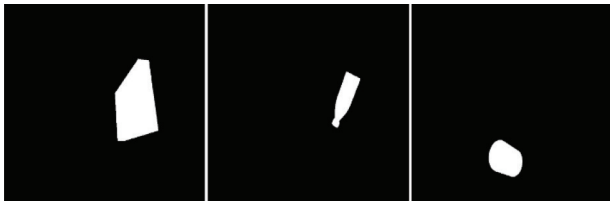

Figure 2 | Examples of virtual masks for three objects: a book, a bottle and a cup.

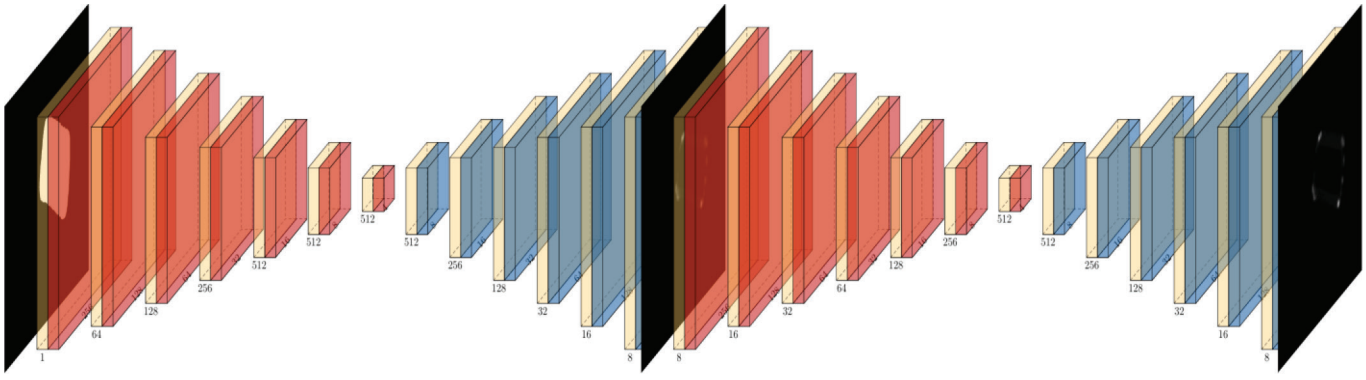

Figure 1 | Architecture of the keypoints localization network with intermediate supervision. Max-pooling layers are drawn in red, deconvolution layers in blue. The symmetric nature of this architecture makes possible the use of residual layers, which are not represented. The size of the feature maps and number of channels are indicated for each layer.

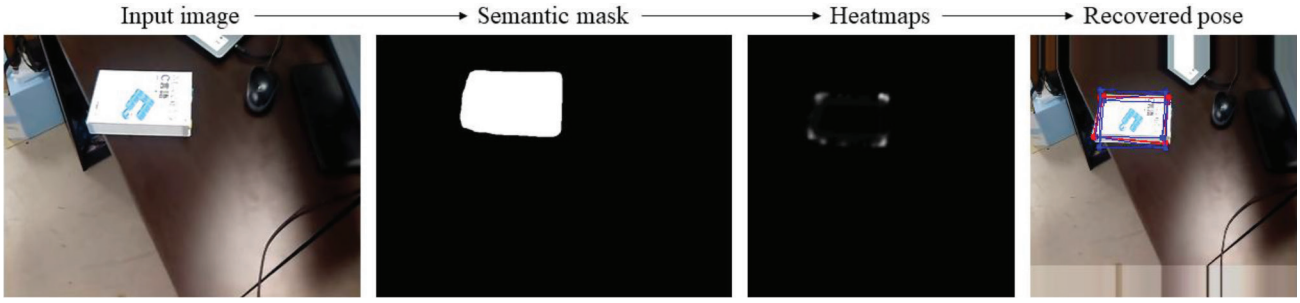

Figure 3 | Full workflow of the algorithm. The semantic mask is inferred by a mask-R-CNN. Then the heatmaps are created by the stacked hourglass network. Finally, the pose is recovered from the keypoints with the solvePnPRansac from the OpenCV library. On the final image, 2D estimated keypoints are drawn in red and the projection of the bounding box with the recovered 6-DoF pose is drawn in blue.

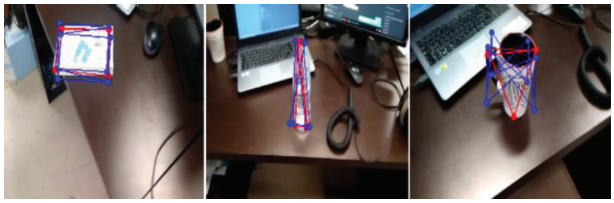

Figure 4 | Illustration of the pose recovery. The output of the network is drawn in red, the projected bounding box in blue.

However, the virtual dataset is not a perfect representation of the mask-R-CNN's output and some issues with the reality gap may appear.

4. EXPERIMENTS

In this section we focus on solving the pose to perspective problem for specific objects. This task has application in robotic tasks like pick and place and can be useful for factory automation. But as explained in Subsection 3.2, the creation of a dataset of real images is very time consuming and we did not create one. Because of that the only evaluation possible for real images in a visual one. The do this we compute the pose of the object for every frame of a video. Then we project the bounding box of the object back one the image and we evaluate visually the output of the algorithm. This evaluation cannot give a good measure of the accuracy but can still prove that the idea of the algorithm is valid at least in some cases. However we could not perform quantitative evaluation of the algorithm due to a lack of dataset, and we can give no more that general recommendation.

Our proposed method achieved acceptable result on virtually generated images. However, when using masks outputted by the mask-R-CNN it does not achieve good accuracy neither stability. We believe that the masks are not sharp enough, but more importantly it shows the importance of texture information on pose estimation task. We advise to use texture information for this kind of task if the object projected geometry change when rotating.

5. SUMMARY

In this paper we proposed an original method to infer the 6-DoF of an object from a single RBG image. This keypoints localization network can be trained very easily using a dataset made of virtual images and can be plugged over any instance segmentation network with only few adjustments. With its easy training process and the good generalization capability, our framework can be useful for many robotic applications where the creation of a specific dataset of real annotated images cannot be done. We also investigated the role of texture information in the pose estimation and shown that it is probably not possible to recover the pose of an object only using texture-less images.

CONFLICTS OF INTEREST

The authors declare they have no conflicts of interest.

REFERENCES

[1] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, Venice, Italy, 2017, pp. 2980–2988.

[2] J. Kang, D. Kim, V. Pujolle, J. Lee, D. Lee, Y.J. Heo, et al., 3D object pose recognition framework for robot task based on RBG image, Postech Domestic Conference, Korea, 2019.

[3] J. Tremblay, T. To, B. Sundaralingam, Y. Xiang, D. Fox, S. Birchfield, Deep object pose estimation for semantic robotic grasping of household objects, Conference on Robot Learning (CoRL), Zurich, 2018.

[4] A. Toshev, C. Szegedy, DeepPose: human pose estimation via deep neural networks, 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Columbus, OH, USA, 2014, pp. 1653–1660.

[5] S. Hinterstoisser, C. Cagniart, S. Ilic, P.F. Sturm, N. Navab, P. Fua, et al., Gradient response maps for real-time detection of texture-less objects, IEEE Trans. Pattern Anal. Mach. Intell. 34 (2012), 876–888.

AUTHORS INTRODUCTION

Mr. Victor Pujolle

He graduated his masters from Kyushu Institute of Technology and the Ecole des Mines de Saint Etienne in 2020. His research interests include Deep learning, computer vision, virtual training and information processing.

Prof. Eiji Hayashi

He is a professor in the Department of Intelligent and Control Systems at Kyushu Institute of Technology. He received the PhD (Dr. Eng.) degree from Waseda University in 1996. His research interests include Intelligent mechanics, Mechanical systems and Perceptual information processing. He is a member of The Institute of Electrical and Electronics Engineers (IEEE) and The Japan Society of Mechanical Engineers (JSME).