SCIENCE CHINA
Information Sciences

• MOOP •

July 2019, Vol. 62 074101:1–074101:3
https://doi.org/10.1007/s11432-018-9688-4

Localizing object parts in 3D from a single image

Shen YIN, Bin ZHOU*, Mingjia YANG & Yu ZHANG

State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing 100191, *China*

Received 14 September 2018/Accepted 4 December 2018/Published online 23 April 2019

Citation Yin S, Zhou B, Yang M J, et al. Localizing object parts in 3D from a single image. Sci China Inf Sci, 2019, 62(7): 074101, https://doi.org/10.1007/s11432-018-9688-4

Object localization in 3D from 2D images is an important computer vision problem that enables modern robotic vision systems to interact properly with the objects present in the real world. Owing to its significance, techniques for recovering the 6-DOF pose and dimensions of objects from images has received increasing attention in recent years [1–3]. However, to the best of our knowledge, the existing approaches all focus on object-level inference instead of analyzing object parts. Most of the time we expect a robot to interact with a semantic part of an object (e.g., holding the bar of a cup), part-level 3D object localization is considerably preferred in practical applications.

Compared with object-level case, localizing object parts in 3D is a significantly challenging problem. First, it requires recovering the orientations and dimensions of each semantic part separately. This problem is difficult due to the inherently large variance of viewpoints, sizes, and shapes of the semantic parts within an object category. Challenges are greater if self-occlusions among object parts are considered. Second, part-level 3D object localization is still a novel problem, thus there is lack methodologies and datasets regarding it. Training data is particularly scarce since annotating parts in 3D is inherently difficult and time-consuming.

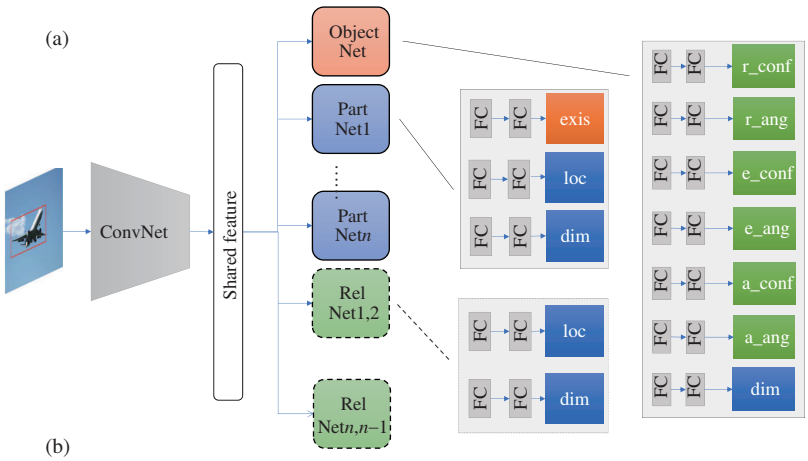
We address the aforementioned two problems and make the first attempt to solve part-level 3D object localization. First, a baseline network and its training strategies are proposed regarding this task, which achieves impressive performance without manual efforts. Second, we propose an improved network that explores the spatial context among object parts, which further improves the performance. Finally, we provide high-quality manual annotations for large-scale dataset to make it possible for this task to be quantitatively evaluated, thereby helping to push the research further along this direction.

The 3D part localization network. Given an input RGB image, we assume that the target object is cropped at first and its category is known, such as [1], and the target object contains several semantic parts, which are defined previously. Our objective is to estimate the 3D bounding boxes of all the semantic parts.

With regard to a certain object part, we parametrize its 3D bounding box using a 4-tuple $(e, \boldsymbol{l}, \boldsymbol{d}, \boldsymbol{R})$ in the object coordinate system, which will be explained later. Here, $e \in \{0, 1\}$ represents the possibility of existence of the part, which solves the problem of the part category not being fixed in number. The center location of the bounding box and its dimension are denoted with $\boldsymbol{l} = [x(\boldsymbol{l}), y(\boldsymbol{l}), z(\boldsymbol{l})]$ and $\boldsymbol{d} = [x(\boldsymbol{d}), y(\boldsymbol{d}), z(\boldsymbol{d})]^{\mathrm{T}}$, respectively. The orientations are encoded by the matrix $\boldsymbol{R}(\theta, \phi, \psi)$, parametrized by the azimuth, elevation and roll angles [4]. For simplicity, we constrain the part bounding box to be axis-aligned, which means that it has the same orientations as that of the entire object.

Correspondingly, the high-level view of our 3D part localization network is shown in Figure 1(a). The network is category specific, based on the number of parts of the object category. The cropped image is input into a set of convolutional

* Corresponding author (email: zhoubin@buaa.edu.cn)



(b)

CATG	Method	Existence (precision)	Viewpoint		3D IoU average
			MedErr	$\text{Acc}_{\frac{\pi}{6}}$	
Aero	Baseline	0.93	0.29	0.68	0.332
	Baseline+SC	0.93	0.35	0.62	0.343
Chair	Baseline	0.91	0.13	0.95	0.221
	Baseline+SC	0.90	0.14	0.91	0.263
Car	Baseline	1	0.11	0.99	0.539
	Baseline+SC	1	0.13	0.97	0.559
Motor	Baseline	0.85	0.24	0.89	0.488
	Baseline+SC	0.86	0.27	0.86	0.490
Table	Baseline	0.86	0.30	0.58	0.119
	Baseline+SC	0.82	0.86	0.43	0.082

Figure 1 (Color online) (a) Proposed network architecture; (b) three parts of accuracy on 5 categories.

layers to produce a shared high-level feature representation. This feature is then entered into a set of part-specific sub-networks and an object-specific sub-network. For the part-net, it produces parameters w.r.t. the existence e, dimensions $\boldsymbol{d}$, and locations $\boldsymbol{l}$ of the parts. For object-net, it involves the dimensions $\boldsymbol{d}_0$ and azimuth, elevation and roll angles that define the rotations $\boldsymbol{R}$.

Exploring spatial context. Note that the proposed network estimates the 3D parameters individually for each part. This would be suboptimal in several cases, especially when a part is severely occluded. Thus, we explore the context information via a simple modification of the initial network. As shown in Figure 1(a), we add a sequence of relation sub-networks into the original network, illustrated as the dashed part. These sub-networks encode the mutual relationships among each pair of the parts. Since the part bounding boxes are all axis-aligned, we primarily consider their mutual relative locations and dimensions. Thus, $\text{RelNet}_{i,j}$ outputs a $\boldsymbol{l}_{i,j}$ and a $\boldsymbol{d}_{i,j}$, where

$$\boldsymbol{l}_{i,j} = \big[x(\boldsymbol{l}_j) - x(\boldsymbol{l}_i), y(\boldsymbol{l}_j) - y(\boldsymbol{l}_i), z(\boldsymbol{l}_j) - z(\boldsymbol{l}_i)\big]^{\mathrm{T}}, \quad (1)$$

and $\boldsymbol{d}_{i,j}$ is defined similarly. To utilize these pairwise estimates to solve individual estimation errors, we let the final estimates meet both the individual and pairwise predictions. Formally, denote the concatenated location and dimension estimates for the i-th part as $\boldsymbol{\alpha}_i = [\boldsymbol{l}_i, \boldsymbol{d}_i]^{\mathrm{T}}$, and the final estimates to optimize as $\boldsymbol{\alpha}_i^*$. Moreover, let $\boldsymbol{\alpha}_{i,j} = [\boldsymbol{l}_{i,j}, \boldsymbol{d}_{i,j}]^{\mathrm{T}}$ be the pairwise estimates output by the network. To obtain $\boldsymbol{\alpha}_i^*$, we solve the following problem:

$$\min_{\{\boldsymbol{\alpha}_i^*\}_i} \sum_i (\boldsymbol{\alpha}_i^* - \boldsymbol{\alpha}_i)^2 + \lambda \sum_{(i,j)} [\varphi(\boldsymbol{\alpha}_i^*, \boldsymbol{\alpha}_j^*) - \boldsymbol{\alpha}_{ij}]^2, \quad (2)$$

where $\varphi(\cdot)$ computes the relative spatial offsets following (1) and λ is a positive constant controlling the relative weights. The resulting problem is a least square system, which can be solved efficiently using linear programming. Finally, we use the solved location and dimension estimates to substitute the individually estimated ones for further processing.

Data collection. Since 3D object part localization from images is still a novel task, there lacks mature datasets for training and evaluation. To solve this problem, we implement an annotation tool. This tool provides a set of segmented 3D objects sampled from the ShapeNet dataset [5]. First, users are asked to select a 3D object that can align several possible parts. For those parts that cannot be aligned properly at this time, users could select another 3D model. As for groundtruth, for each object we record the existence status of all parts, three viewpoint angles, and the coordinate of eight vertices of 3D bounding boxes in object coordinate frame, which could be consistent since the coordinates of each category object are oriented in the same direction.

Besides, data augmentation is also explored by synthesizing images using the off-the-shelf part annotations in modern 3D model datasets. Although manually annotated 3D part bounding boxes provide accurate training data, they are still expensive to obtain. Moreover, training deep neural networks heavily relies on the availability of large-scale training data. Inspired by recent approaches on rendering for CNN [6], we make use of existing 3D shape datasets to rapidly generate a large number of synthesized training images.

Experiments and discussions. Since this task has not been addressed previously, making direct comparisons is difficult. Thus, the primary experiments is that we conduct evaluations to analyze different parts of the proposed approach, to provide baseline results for further research and indicate the underlying challenges and scopes for improvement.

We train the network using the synthesized images, while perform evaluations on the ObjectNet3D images. In this setting, we test the generalization ability of the network trained from synthesized data on realistic situations. The results on the 5 categories are summarized in Figure 1(b).

Following the conventions of 3D object detection and viewpoint estimation, we apply four metrics for evaluation. For part existence, a simple accuracy is applied here. Subsequently, we calculate 3D IoU (intersection over union) in object coordinate frame for evaluation of part location and dimension. As for viewpoint estimation, two metrics [4] are used here: MedErr measures the median error of the rotation matrix, $\text{Acc}_{\frac{\pi}{6}}$ measures the accuracy of viewpoint error within $\frac{\pi}{6}$. From the results, we can observe that our baseline network can achieve promising performance. After incorporating the spatial context information, the localization IoUs are improved consistently regarding all the categories. It can be demonstrated that such improvements are more significant regarding the challenging small parts.

Conclusion. We focus on the task of part-level 3D detection from a single image, to our knowledge, no relevant research has been conducted before. In an attempt to solve this problem, we propose a baseline network and improve it by considering spatial context. We also enhance ObjectNet3D [7] and PASCAL3D+ [8] datasets by annotating 3D bounding box of all parts of object to evaluate our model. Regarding the evaluation experiments we achieve a significant result and considerable studies require to be conducted. Future work should deal with a considerably effective network concerning both viewpoint and 3D box; larger and more accurate datasets would also be helpful.

Limitation. We estimate the positions and dimensions of part 3D bounding boxes. Instead of estimating the complex orientations, we assume that part bounding boxes share the same orientations with object bounding boxes. In this case, the bounding box we obtain is not the most fitting one. Since the actual orientations may be different among the parts, such as the two wings of an airplane, the assumption that different parts share a same viewpoint in an image is only reasonable to obtain an axis-aligned bounding box (AABB) in our study. For motion parts, such as a car's door, although it is still possible to obtain its AABB according to previous assumptions, we remove the images that the part has moved. It seems that estimating an oriented bounding box (OBB) in 3D would be more preferable although challenging for future research.

Acknowledgements This work was partly supported by National Natural Science Foundation of China (Grant Nos. U1736217, 61502023).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Mousavian A, Anguelov D, Flynn J, et al. 3D bounding box estimation using deep learning and geometry. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017
2 Fidler S, Dickinson S J, Urtasun R. 3D object detection and viewpoint estimation with a deformable 3D cuboid model. In: Proceedings of the 26th Annual Conference on Neural Information Processing Systems, 2012
3 Chen X Z, Kundu K, Zhang Z Y. Monocular 3D object detection for autonomous driving. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2016
4 Tulsiani S, Malik J. Viewpoints and keypoints. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2015
5 Chang A X, Funkhouser T A, Guibas L J. Shapenet: an information-rich 3D model repository. 2015. ArXiv:1512.03012
6 Su H, Qi C R, Li Y Y, et al. Render for CNN: viewpoint estimation in images using cnns trained with rendered 3D model views. In: Proceedings of IEEE International Conference on Computer Vision, 2015
7 Xiang Y, Kim W, Chen W. Objectnet3D: a large scale database for 3D object recognition. In: Proceedings of European Conference on Computer Vision, 2016
8 Xiang Y, Mottaghi R, Savarese S. Beyond PASCAL: a benchmark for 3D object detection in the wild. In: Proceedings of IEEE Winter Conference on Applications of Computer Vision, 2014