Crafting a multi-task CNN for viewpoint estimation

Francisco Massa
http://imagine.enpc.fr/~suzano-f/
Renaud Marlet
http://imagine.enpc.fr/~marletr/
Mathieu Aubry
http://imagine.enpc.fr/~aubrym/

LIGM, UMR 814, Imagine,
Ecole des Ponts ParisTech, UPEM,ESIEE Paris, CNRS, UPE
Champs-sur-Marne, France

Convolutional Neural Networks (CNNs) were recently shown to provide state-of-the-art results for object category viewpoint estimation. However different ways of formulating this problem have been proposed and the competing approaches have been explored with very different design choices. This paper presents a comparison of these approaches in a unified setting as well as a detailed analysis of the key factors that impact performance. Followingly, we present a new joint training method with the detection task and demonstrate its benefit. We also highlight the superiority of classification approaches over regression approaches, quantify the benefits of deeper architectures and extended training data, and demonstrate that synthetic data is beneficial even when using ImageNet training data. By combining all these elements, we demonstrate a consistent improvement of approximately 5% mAVP over previous state-of-the-art results on the Pascal3D+ dataset [4].

Contributions: In this paper, we study several factors that affect performance for the task of joint object detection and pose estimation with CNNs and introduce a new approach for the joint training. Using the best design options, we rationally define an effective method to integrate detection and viewpoint estimation, quantify its benefits, as well as the boost given by deeper networks and more training data, including data from ImageNet and synthetic data. The relative benefits of each of these elements as well as a comparison with baseline is summarized in table 1. We demonstrate that the combination of all these elements leads to an important improvement over state-of-the-art results on Pascal3D+, going for example from 31.1% to 36.1% AVP in the case of the most challenging 24 viewpoints classification. While several of the elements that we employ have been used in previous work [2, 3], we know of no systematic study of their respective and combined effect, resulting in an absence of clear good practices for viewpoint estimation and sub-optimal performances.

[1] B. Pepik, M. Stark, P. Gehler, and B. Schiele. Teaching 3D geometry to deformable part models. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[2] H. Su, C. R. Qi, Y. Li, and L. J. Guibas. Render for CNN: Viewpoint estimation in images using CNNs trained with rendered 3D model views. In *Proceedings of the IEEE International Conference on Computer Vision*.

[3] S. Tulsiani and J. Malik. Viewpoints and keypoints. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*.

[4] Y. Xiang, R. Mottaghi, and S. Savarese. Beyond Pascal: A benchmark for 3D object detection in the wild. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2014.

Table 1: Summary of results and comparison with baselines using AVP24

Method	aero	bike	boat	bus	car	chair	table	mbike	sofa	train	tv	mAVP24
DPM-VOC+VP [1]	9.7	16.7	2.2	42.1	24.6	4.2	2.1	10.5	4.1	20.7	12.9	13.6
Render For CNN [2]	21.5	22.0	4.1	38.6	25.5	7.4	11.0	24.4	15.0	28.0	19.8	19.8
Viewpoints & Keypoints [3]	37.0	33.4	10.0	54.1	40.0	**17.5**	19.9	34.3	28.9	43.9	22.7	31.1
Classif. approach & AlexNet	21.6	15.4	5.6	41.2	26.4	7.3	9.3	15.3	13.5	32.9	24.3	19.3
+ our joint training	24.4	16.2	4.7	49.2	25.1	7.7	10.3	17.7	14.8	36.6	25.6	21.1
+ VGG16 instead of AlexNet	26.3	29.0	8.2	56.4	36.3	13.9	14.9	27.7	20.2	41.5	26.2	27.3
+ ImageNet data	42.4	37.0	**18.0**	59.6	43.3	7.6	25.1	**39.3**	29.4	**48.1**	28.4	34.4
+ synthetic data	**43.2**	**39.4**	16.8	**61.0**	**44.2**	13.5	**29.4**	37.5	**33.5**	46.6	**32.5**	**36.1**