Efficient 3D Pose Estimation and 3D Model Retrieval

Alexander Grabner[1], Peter M. Roth[1], and Vincent Lepetit[2,1]
{alexander.grabner,pmroth,lepetit}@icg.tugraz.at

I. Problem Statement and Motivation

Retrieving 3D models for objects in 2D images is an increasingly important problem, driven by the recent emergence of large databases of 3D models such as ShapeNet [1]. However, this task is challenging for two main reasons: (1) 2D images and 3D models have considerably different representations and characteristics, making it hard to compare them. (2) The appearance of objects can significantly vary with the pose, but it is in general unknown and, thus, multiple poses have to be considered, which is very inefficient. To overcome these problems, in [2] we proposed first to predict the object's pose and then to use the estimated pose as a prior to retrieve 3D models from a database. In the following, we give a short summary of the approach in Sec. II and a sketch of results in Sec. III. For more details, we refer to [2].

II. Overview of the Approach

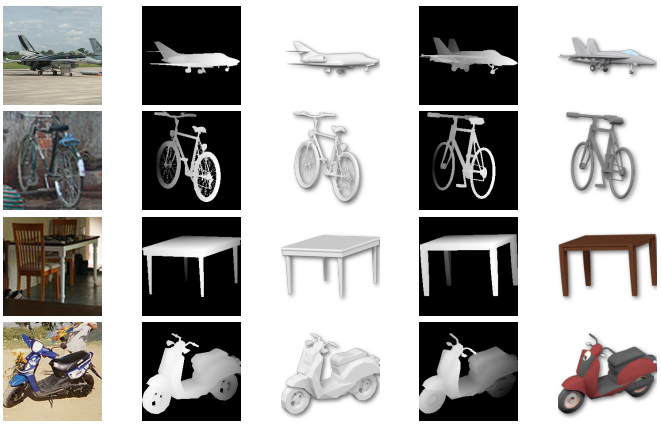
Pose Estimation: To robustly compute the 3D pose of the objects of interest, similar to [3], we predict the 2D image locations of virtual control points using a CNN. In particular, we compute the 2D image locations of the projections of the object's eight 3D bounding box corners. The actual 3D pose is then estimated by solving a perspective-n-point (PnP) problem. As this requires the 3D coordinates of the virtual control points to be known, we predict the spatial dimensions of the object's 3D bounding box and use these to scale a unit cube, which approximates the ground truth 3D coordinates. For this purpose, we introduce a CNN architecture which jointly predicts the 2D image locations of the projections of the eight 3D bounding box corners (16 values) as well as the 3D bounding box dimensions (3 values).

Model Retrieval: The actual 3D model retrieval is realized via descriptor matching between RGB images and depth images rendered under the estimated objects' pose. Using this prior significantly reduces the computational complexity compared to methods which need to process multiple renderings per 3D model. In addition, using depth instead of RGB images avoids problems with texture, different material properties, and illumination. However, RGB images and depth images have considerably different characteristics. Thus, we introduce a multi-view metric learning approach based on triplet loss optimization [4], which maps images from both domains to a common representation.

[1]Institute of Computer Graphics and Vision, Graz University of Technology, Austria
[2]Laboratoire Bordelais de Recherche en Informatique, University of Bordeaux, France

III. Discussion and Illustrative Results

In this way, we are the first to report quantitative results for 3D model retrieval on Pascal3D+ [5] and show that our method, which was trained purely on Pascal3D+, retrieves rich and accurate 3D models from ShapeNet given RGB images of objects in the wild. In addition, we significantly outperform the state-of-the-art in 3D viewpoint estimation on Pascal3D+. A few illustrative 3D model retrieval results are sketched in Fig. 1. For more results, we refer to [2].



(a) (b) (c) (d) (e)

Fig. 1: 3D pose estimation and 3D model retrieval from ShapeNet given unseen images from Pascal3D+: (a) query RGB image, (b) depth image and (c) RGB rendering illustrating the ground truth pose and 3D model from Pascal3D+, (d) depth image and (e) RGB rendering illustrating our predicted pose and retrieved 3D model from ShapeNet.

Acknowledgment

This work was supported by the Christian Doppler Laboratory for Semantic 3D Computer Vision, funded in part by Qualcomm Inc.

References

[1] A. Chang, T. Funkhouser, L. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu, "Shapenet: An Information-Rich 3D Model Repository," Stanford University — Princeton University — Toyota Technological Institute at Chicago, Tech. Rep., 2015.
[2] A. Grabner, P. M. Roth, and V. Lepetit, "3D Pose Estimation and 3D Model Retrieval for Objects in the Wild," in *Proc. CVPR*, 2018.
[3] M. Rad and V. Lepetit, "BB8: A Scalable, Accurate, Robust to Partial Occlusion Method for Predicting the 3D Poses of Challenging Objects Without Using Depth," in *Proc. ICCV*, 2017.
[4] K. Weinberger and L. Saul, "Distance Metric Learning for Large Margin Nearest Neighbor Classification," *Journal of Machine Learning Research*, vol. 10, pp. 207–244, 2009.
[5] Y. Xiang, R. Mottaghi, and S. Savarese, "Beyond Pascal: A Benchmark for 3D Object Detection in the Wild," in *Proc. WACV*, 2014.