When 3D Reconstruction Meets Ubiquitous RGB-D Images (extended abstract)

Quanshi Zhang†, Xuan Song†, Xiaowei Shao†, Huijing Zhao†, and Ryosuke Shibasaki†
†University of Tokyo, ‡Peking University

1. Extended abstract

3D reconstruction from a single image is a classical problem in computer vision. However, it still poses great challenges for single-view reconstruction of daily-use objects with irregular† shapes. In this paper, we re-consider the problem of single-view 3D reconstruction in terms of two CV areas: category modeling and knowledge mining from big visual data.

Category modeling & task output: For the bottleneck in single-view 3D reconstruction, i.e., the reconstruction of objects with irregular† structures, we have to return to the concept of “category modeling”.

Therefore, the objective is to train a model to detect objects in the target category in large images, while simultaneously projecting their 2D shapes into the 3D space at the pixel level. The category model encodes the knowledge of how intra-category structure deformation and object rotations affect 2D object shapes.

Mining from big visual data & task input: Another bottleneck lies in efficiently learning the category-specific knowledge of 3D reconstruction for a huge number of categories. Ideally, we would need to train a model for each object category in daily use, so as to construct a knowledge base to provide a 3D reconstruction service for arbitrary RGB images. Therefore, we hope to learn from big visual data to avoid the labor of manually preparing training samples (e.g., well-built 3D models) for each category, and thus ensure a high learning efficiency.

†The word “irregular” is used to indicate that our approach focuses on general object categories without setting strong assumptions for object shapes. In contrast, Cheeger-set-based methods focus on ball-like surfaces, and perspective-based methods require vanishing points.

In this paper, we train category models directly from informally captured and “unaligned” RGB-D images. We use the phrase “informally captured” to describe loose requirements for ubiquitous images. They are typical of what can be directly collected by search engines (Figures 1 and 2).

The informally captured images are not manually aligned, and they consist of small objects that are randomly positioned. In particular, these daily-use objects are usually designed with texture variations, various rotations, and some structure deformation for commercial purposes. Technically speaking, these images can be loosely regarded as a kind of big visual data.

Proposed method: The flowchart of our method is shown in Fig. 1. We train both the category detector and the category model for 3D reconstruction using ubiquitous images without labeling “what is where” to ensure high efficiency in model learning. We first mine the category detector from informally captured RGB-D images to collect RGB-D objects, and then use these object samples to train the category model. For testing, the category detector and model are used in sequence to localize target objects and estimate their 3D structures in cluttered 2D images.

The use of ubiquitous images raises a number of challenges, as shown in Fig. 2. The category detectors collect RGB-D (or RGB) object samples from the informally captured RGB-D (or RGB) images, which serves as the basis for further training (or testing) of 3D reconstruction. The training of category detectors should overcome challenges...
of the small sizes of objects and the variations in texture, rotation, and scale. We use unsupervised category modeling based on graph mining [7] to train the category detector. This algorithm automatically discovers a set of key object parts for each category. Thus, the category detector achieves a part-level detection. We then train the category model to use the relative 2D positions between the key parts to determine the object’s pose and 3D structure. In particular, the category model is trained to use the 2D object shape to simultaneously identify and remove the effects of viewpoint changes, 3D structure deformation, and depth measurement errors.

Comparison with pioneering studies: We limit our discussion to 3D reconstruction from a single image without any human interactions. Many methods for single-view 3D reconstruction have strong assumptions for the image environment. For example, “shapes from shading” methods [5] had special requirements for lighting conditions and textures. Many studies [6] used the perspective principle for 3D reconstruction, which typically relied on the vanishing points in images. Fouhey et al. [2] proposed to learn 3D primitives from RGB-D images for indoor environments. The Cheeger Set approach [4] assumed that the object structure in question had smooth surfaces.

However, such assumptions for object structure are usually not valid for the reconstruction of irregular shapes. From this perspective, the basic idea is to train a model for 3D reconstruction of each category. However, [3, 1] usually required a comprehensive dataset of 3D object samples that were captured in all poses from different viewpoints or well built 3D object models for training. Therefore, in order to reduce the great cost of human labeling, we propose directly learning from informally captured RGB-D images without “what is where”.

Summary: The contributions of this paper can be summarized as follows. We regard 3D reconstruction from a single image as a category modeling problem to overcome the difficulties in the reconstruction of irregular shapes. To ensure a high learning efficiency and a wide application, this is the first attempt to learn category models from informally captured RGB-D images, and then apply the category model back to the informally captured RGB images for 3D reconstruction. We explore a new set of challenges raised by learning from ubiquitous images, and provide a solution.

The original paper is published as [8].

References