Neurocomputing 174 (2016) 1101–1106

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Detection based object labeling of 3D point cloud for indoor scenes

Wei Liu [a,b,1], Shaozi Li [b,*], Donglin Cao [b], Songzhi Su [b], Rongrong Ji [b]

[a] Department of Information Engineering, East China Jiaotong University, China
[b] Department of Cognitive Science, Xiamen University, China

ARTICLE INFO

Article history:
Received 21 November 2014
Received in revised form
2 October 2015
Accepted 2 October 2015
Communicated by Ke Lu
Available online 19 October 2015

Keywords:
Point cloud
Labeling
Object detection
RGB-D

ABSTRACT

While much exciting progress is being made in 3D reconstruction of scenes, object labeling of 3D point cloud for indoor scenes has been left as a challenge issue. How should we explore the reference images of 3D scene, in aid of scene parsing? In this paper, we propose a framework for 3D indoor scenes labeling, based upon object detection on the RGB-D frames of 3D scene. First, the point cloud is segmented into homogeneous segments. Then, we utilize object detectors to assign class probabilities to pixels in every RGB-D frame. After that, the class probabilities are projected into the segments. Finally, we perform accurate inference on a MRF model over the homogeneous segments, in combination with geometry cues to output the labels. Experiment on the challenging RGB-D Object Dataset demonstrates that our detection based approach produces accurate labeling and improves the robustness of small object detection for indoor scenes.

1. Introduction

Coming with the popularity of Kinect sensors and advanced 3D reconstruction techniques, we are facing an increasing amount of 3D point clouds. Consequently, the demand of scene understanding is emerging [6,7]. Understanding 3D scenes is a fundamental problem in perception and robotics, as knowledge of the environment is a preliminary step for subsequent tasks such as route planning, augmented reality and mobile visual search [1,2].

In the literature, a significant amount of work has been done in semantic labeling for pixels or regions in 2D images. In spite of this, semantic labeling of 3D point clouds remains an open problem. Its solution, however, can bring a breakthrough in a wide variety of computer vision and robotics research, with great potential in human-computer interface, 3D object indexing and retrieval, object manipulation in robotics [4] as well as exciting applications such as self-driving vehicles and semantic-aware augmented reality.

There is a great deal of literature on 3D scene labeling in indoor environments [5,16]. Many of these operate directly on a 3D point cloud, as 3D point clouds contain very important shape and spatial information, which allows to model context and spatial relationships between objects However,one missing information of these methods to explore is the reference images. Existing works [4,3] have shown that multi-view images can be used to recognize objects of 3D scene to a higher degree of accuracy.

In this paper, we focus on detection based object labeling in 3D scene. Specifically, we tackle the problem of object labeling in 3D scene with the help of object detection results from part-of-scene 2D reference image. Our goal is to transfer such reliable 2D labeling results into 3D to enhance the inference accuracy. It is noted that the proposed approach works in the scenario where a single point cloud is merged from multiple images. First, the point cloud is segmented into homogeneous segments. Then, we utilize object detectors to assign class probabilities to pixels in every RGB-D frame. After that, the class probabilities are projected into the segments. Finally, we perform accurate inference on a MRF model over the homogeneous segments, in combination with geometry cues to output the labels.

The outline of this paper is organized as follows. Section 2 surveys related work. The detailed of the proposed method is presented in Section 3. Experimental results and comparison are provided in Section 4. Finally, we draw conclusion of this paper in Section 5.

2. Related work

With the popularity of Kinect camera and 3D laser scanner, recently there is ever increasing research focus on semantic labeling of 3D point cloud [8,9,5]. The application scenarios are

* Corresponding author.
E-mail address: szlig@xmu.edu.cn (S. Li).
[1] This work was done when the author was a Ph.D. candidate at Xiamen University.

http://dx.doi.org/10.1016/j.neucom.2015.10.005
0925-2312/

either indoor or outdoor, with corresponding tasks such as labeling tables, beds, desks, computers, or trees, cars, roads and pedestrians. So far, most robotic scene understanding work has focused on 3D outdoor scenes with related applications such as mapping and autonomous driving [10–13]. The task is to label point clouds obtained from 3D laser scanner data into a few coarse geometric classes (e.g., ground, buildings, trees, vegetation, etc.) [8,9,14,15]. Indoor scenes, in comparison, cover a wider range of objects, scene layout, and scales, which is therefore more challenging. A major challenge of this task arises from the fact that most indoor scenes are cluttered and occluded with each other [16,17]. There is also some recent work in labeling indoor scenes [18,19,4,5,20]. While these methods achieve impressive results on large architectural elements (e.g. walls, chairs), the performance is still far from satisfactory for small objects such as torches and cups.

With the emerging 3D reconstruction techniques such as PTAM [21,22], DTAM [23], and Kinectfusion [24,25], it is feasible to robustly merge RGB-D videos of a scene into constant and detailed 3D point cloud [26]. To label the 3D scene in this scenario, one potential solution is to directly work with the merged point cloud, which has achieve high performance for outdoor scenes, where laser data is classified and modeled into a few coarse geometric classes [8,9,27]. The recent researches of [5,20] demonstrate the feasibility to label indoor point clouds. The other solutions conduct segmentation and/or detection on individual frames respectively, whose outputs are then merged into point cloud and undergo a joint optimization procedure [12,4]. Most of these approaches label individual 3D laser points or voxels using features describing local shape and appearance, and the joint optimization is typical accomplished using spatial and temporal smoothing via graphical model inference

The closest work to ours is presented in [3]. The authors utilized sliding window detectors trained from object views to assign class probabilities to pixels in every RGB-D frame. Then, these probabilities are projected into the reconstructed 3D scene and integrated using a voxel representation. Finally, inference is performed by using a Markov Random Field (MRF) model over the voxels, in combination with cues from view-based detection and 3D shape. The downside of their detection based approach is the spoil of object boundaries, leading to misclassification of objects.

To maintain accurate object boundaries, we present a detection based scheme in combination with clustering of 3D points for labeling indoor scenes. We first introduce a normal and color based segmentation algorithm to oversegment the reconstructed 3D scene into homogeneous segments. Then, we utilize sliding window detectors trained from object views to assign class probabilities to pixels in every RGB-D frame. After that, the class probabilities are projected into the segments. Finally, we perform accurate inference on a MRF model over the homogeneous segments, in combination with geometry cues to output the labels (semantic segments) of the point cloud. We term the proposed approach as DetSMRF, which refers to 3D indoor scene labeling based on detection, segmentation and MRF.

DetSMRF can be considered as a generalization of [3]. The main difference between the proposed framework and [3] is that the nodes in our MRF model are homogeneous segments rather than voxels, which is key to preserve object boundaries. Moreover, it is worthwhile to note that the geometry cues integrated into MRF model are simply the difference of normals between two nearby segments which does not require any kinect pose parameters, as compared to [1]. Experiment on the challenging RGB-D Object Dataset demonstrates that our DetSMRF approach produces accurate labeling and improve the robustness of object detection for indoor scenes.

3. DetSMRF

We now describe the proposed DetSMRF model, as shown in Fig. 1. The 3D point clouds we consider are captured from a set of RGB-D video frames. To reconstruct a 3D scene, RGB-D mapping [26] is employed to globally align and merge each frame with the scene under a rigid transform. The goal of this paper is to label small everyday objects of interest, which may comprise only a small part of the whole 3D point cloud.

DetSMRF is defined over a graph $\mathcal{G}=(\mathcal{V},\mathcal{E})$, where $\mathcal{V}$ are vertices representing homogenous segments obtained by oversegmentation of 3D point cloud, $\mathcal{E}$ are edges connecting adjacent vertices. Each segment is associated with a label $y_v \in \{1,\ldots,C,c_B\}$, where $\{1,\ldots,C\}$ are object classes of interest and c_B is the background class. The joint distribution of segments are modeled via a MRF with pairwise interactions. The optimal labeling of the scene minimizes the following energy function defined over the field of homogenous segments $\{1,\ldots,C\}$:

$$E(y_1,\ldots,y_{|\mathcal{V}|})=\sum_{v\in\mathcal{V}}\Psi_v(y_v)+\lambda\sum_{\{i,j\}\in\mathcal{P}}\Phi_{i,j}(y_i,y_j), \tag{1}$$

where $\mathcal{P}$ is the set of all pairs of neighboring segments. The data term Ψ_v models the observation cost for individual segments, and the pairwise term $\Phi_{i,j}$ measures interactions between adjacent segments. λ is the smoothing constant, penalizing adjacent elements do not share the same label.

3.1. Oversegmentation of 3D point cloud

Given the 3D point cloud merged by RGB-D mapping from RGB-D video we first reduce the complexity from millions of points on the 3D surface to a few thousand segments. That is to say, we want to group points from 3D point cloud into perceptually meaningful segments which conform to object boundaries. We assume that the boundaries between objects are represented by discontinuities in color and surface orientation. The grouping is done by oversegmentation. To this end, a normal and color based

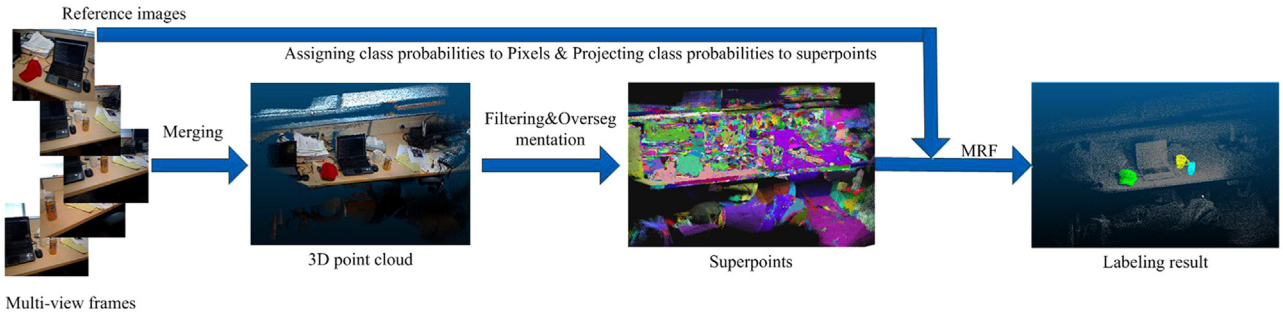

Fig. 1. Flow chart of our method.