Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)

Auto-Annotation of 3D Objects via ImageNet

Huan Luo, Cheng Wang, Jonathan Li
Fujian Key Laboratory of Sensing and Computing for Smart Cities
School of Information Science and Engineering
Xiamen University, Xiamen, FJ 361005
scholar.luo@gmail.com, {cwang, junli}@xmu.edu.cn

Abstract

Automatic annotation of 3D objects in cluttered scenes shows its great importance to a variety of applications. Nowadays, 3D point clouds, a new 3D representation of real-world objects, can be easily and rapidly collected by mobile LiDAR systems, e.g. RIEGL VMX-450 system. Moreover, the mobile LiDAR system can also provide a series of consecutive multi-view images which are calibrated with 3D point clouds. This paper proposes to automatically annotate 3D objects of interest in point clouds of road scenes by exploiting a multitude of annotated images in image databases, such as LabelMe and ImageNet. In the proposed method, an object detector trained on the annotated images is used to locate the object regions in acquired multi-view images. Then, based on the correspondences between multi-view images and 3D point clouds, a probabilistic graphical model is used to model the temporal, spatial and geometric constraints to extract the 3D objects automatically. A new dataset was built for evaluation and the experimental results demonstrate a satisfied performance on 3D object extraction.

Introduction

Nowadays, along with the booming of deep learning used in 3D applications, such as 3D object classification and detection, the need for large amounts of annotated 3D data becomes increasingly urgent for model learning. Commonly, the acquisition of such annotations is time-consuming and labor-intensive. Fortunately, web image databases, such as LableMe and ImageNet, provide abundant annotations for a multitude of digital images which cover a variety of scenarios. Exploiting these tagged image databases to accomplish automatic annotation of 3D objects in cluttered scenes shows a bright prospect in many applications.

In recent years, with the rapid development of Light Detection and Ranging (LiDAR) technologies, large-scale road scenes are now depicted by large volumes of highly dense and accurate 3D point clouds that are collected by mobile LiDAR systems. By smoothly integrating laser scanners with position and orientation systems, mobile LiDAR systems can rapidly capture undistorted 3D point clouds with real-world coordinates (See Fig. 1). Moreover, with complementary onboard multi-view high-resolution digital cameras, a series of consecutive images (calibrated with 3D point clouds) are captured to provide textual/color information of objects in a scene from different viewpoints and distinctive points in time.

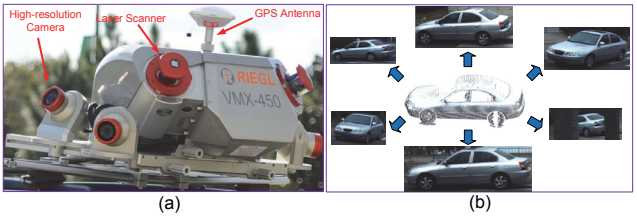

Figure 1: (a) mobile LiDAR system with four high-resolution cameras and two laser scanners; (b) multi-view images taken from distinctive angles and distances.

This paper presents a method to automatically annotate 3D objects from point clouds by building a bridge to connect the image databases with 3D point clouds. More specifically, we firstly use Faster-RCNN (Ren et al. 2015) to train an object-specific detector on ImageNet for transferring category labels from ImageNet to reference images (the multi-view images registered with 3D point clouds). Then, the trained object detector is used to locate the object regions in the reference images. After that, through correspondences between superpixels from reference images and the supervoxel from 3D point clouds, we impose the temporal, spatial and geometric constraints on labels of 3D objects by exploiting a probabilistic graphical model. Finally, the graphical model is effectively solved to generate the annotated 3D objects.

Method

The whole processing flow of our proposed method is illustrated in Fig. 2. The remaining section focuses on building a probabilistic graphical model among the superpixels and supervoxels. Let $\mathcal{P}$ and $\mathcal{L}$ denote the set of superpixels generated from reference images and the set of supervoxels generated from point clouds scenes, respectively. For each superpixel $i \in \mathcal{P}$ and each supervoxel $l \in \mathcal{L}$, we denote variables s_i and s_l taking values from the category labels $\mathcal{S} = \{1, 0\}$. Here, 1 and 0 represent the category of object of interest and background, respectively. Let $\mathbf{s} = \{s_i | i \in \mathcal{P}\} \bigcup \{s_l | l \in \mathcal{L}\}$,

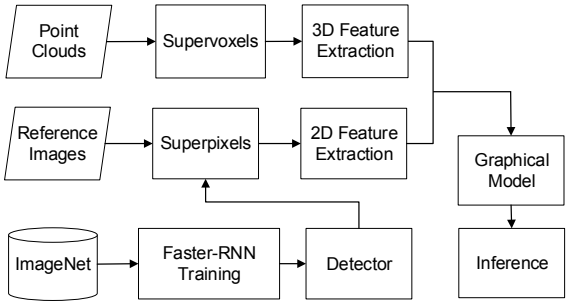

Figure 2: Processing flow of our proposed methods.

we model the temporal, spatial and geometric constraints by a Markov Random Field (MRF) model and its energy function is formalized as follows:

$$E(\mathbf{s}) = \sum_{i\in\mathcal{P}} \varphi_i^{\mathcal{P}}(s_i) + \alpha \sum_{i,j\in\mathcal{P}} \phi_{i,j}^{\mathcal{P},\mathcal{P}}(s_i, s_j) + \beta \sum_{l,k\in\mathcal{L}} \phi_{l,k}^{\mathcal{L},\mathcal{L}}(s_l, s_k) + \gamma \sum_{i\in\mathcal{P},l\in\mathcal{L}} \phi_{i,l}^{\mathcal{P},\mathcal{L}}(s_i, s_l) \quad (1)$$

where $\varphi(\cdot)$ and $\phi(\cdot)$ represent the unary and pairwise potentials, respectively. α, β, and γ are the parameters controlling the weight of pairwise potentials.

The superpixel unary potential $\varphi_i^{\mathcal{P}}(s_i)$ encodes the possibility of superpixel i taking the category label s_i

$$\varphi_i^{\mathcal{P}}(s_i) = -\log p_i^{\mathcal{P}}(s_i) \quad (2)$$

where $p_i^{\mathcal{P}}(s_i)$, obtained by applying the object detector, represents the probability whether the supervixel i belongs to the object of interest.

The superpixel pairwise potential $\phi_{i,j}^{\mathcal{P},\mathcal{P}}(s_i, s_j)$ encodes category relations between superpixels i and j. We adpot the Potts model (Kohli, Kumar, and Torr 2007) to encourage the neighboring superpixels taking the identical category if they own the similar texture

$$\phi_{i,j}^{\mathcal{P},\mathcal{P}}(s_i, s_j) = [s_i \neq s_j] \cdot (1 - D_{2d}(i, j)) \quad (3)$$

where $[\cdot]$ is indicator function whose value is 1 if its argument is true, otherwise 0. $D_{2d}(i, j)$ determines the similarity of texture of two superixels in 2D images.

Similarly, we formulate the supvoxel pairwise potential $\phi_{l,k}^{\mathcal{L},\mathcal{L}}(s_l, s_k)$ to encourage category consistency of spatially neighboring supervoxels based on the geometric features

$$\phi_{l,k}^{\mathcal{L},\mathcal{L}}(s_l, s_k) = [s_l \neq s_k] \cdot (1 - D_{3d}(l, k)) \quad (4)$$

where $D_{3d}(l, k)$ measures the geometrical similarity between supervoxel l and k.

The 2D/3D pairwise potential $\phi_{i,l}^{\mathcal{P},\mathcal{L}}(s_i, s_l)$ encourages that supervoxel should take the identical category with its corresponding superpixel. The corresponding relation can be obtained by projecting the points of supervoxel onto the reference image plane. we formulate $\phi_{i,l}^{\mathcal{P},\mathcal{L}}$ as follows:

$$\phi_{i,l}^{\mathcal{P},\mathcal{L}}(s_i, s_l) = \exp(-\frac{Dist(l, i)}{\sigma}) \cdot [OD(l)] \quad (5)$$

where $Dist(l, i)$ computes the spatial distance between the positions of supervoxel, l, and the viewpoint where the mobile LiDAR system records the image which the superpixel, i belongs to. Therefore, Eq. (5) encourages that the superpixel should own more confidence to take the same category with the supervoxel if the position of its viewpoint is near to the supervoxel. Here, σ is a scale factor which makes the potential comparable. The $OD(l)$ obtained by calculating the visibility of point clouds (Katz and Tal 2015), determines whether supervoxel, l, is occluded by the other 3D points. In the proposed method, no penalty will be imposed on the energy function (1), if the supervoxel is occluded. The energy function (1) meets the semimetric condition and can be efficiently minimized by the Graph Cuts algorithm (Boykov, Veksler, and Zabih 2001).

Experiments

Our proposed method was evaluated on a dataset collected by a RIEGL VMX-450 mobile LIDAR system on Xiamen Island, China. This dataset contains the point clouds and high-resolution optical images which is calibrated with the point clouds. In this dataset, the vehicles in point clouds were manually annotated for evaluating the proposed method. Here, the amount of vehicles is 113.

The parameters α, β, and γ were determined by implementing a grid search on a small validation set. The similarity metrics were defined by computing χ^2 distance in feature spaces. In images, the mean RGB value was used to describe a superpixel. In point clouds, the FPFH and spectral features are used to describe the geometric feature for a supervoxel.

To verify the validity of our proposed method, we design a Unary method by removing the pairwise potentials $\phi_{i,j}^{\mathcal{P},\mathcal{P}}$ and $\phi_{l,k}^{\mathcal{L},\mathcal{L}}$ from energy function (1). As the experimental results exhibits in Table 1, our proposed method achieves a satisfying performance on annotation of 3D vehicles in the evaluation.

Table 1: Annotation results on vehicle dataset

	Precision	**Recall**	**F1-score**
Unary	0.851	0.628	0.722
Full Model	0.964	0.721	0.825

References

Boykov, Y.; Veksler, O.; and Zabih, R. 2001. Fast approximate energy minimization via graph cuts. *PAMI* 23(11):1222–1239.

Katz, S., and Tal, A. 2015. On the visibility of point clouds. In *ICCV*, 1350–1358.

Kohli, P.; Kumar, M. P.; and Torr, P. H. 2007. P3 & beyond: Solving energies with higher order cliques. In *CVPR*, 1–8. IEEE.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 91–99.