Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Fusion of images and point clouds for the semantic segmentation of large-scale 3D scenes based on deep learning

Rui Zhang[a,b,*], Guangyun Li[a], Minglei Li[a], Li Wang[a]

[a] Information Engineering University, 450001 Zhengzhou, China
[b] North China University of Water Resources and Electric Power, 450045 Zhengzhou, China

ARTICLE INFO

Keywords:
3D scene segmentation
2D image
3D point cloud
Large-scale
High-resolution

ABSTRACT

We address the issue of the semantic segmentation of large-scale 3D scenes by fusing 2D images and 3D point clouds. First, a Deeplab-Vgg16 based Large-Scale and High-Resolution model (DVLSHR) based on deep Visual Geometry Group (VGG16) is successfully created and fine-tuned by training seven deep convolutional neural networks with four benchmark datasets. On the *val* set in CityScapes, DVLSHR achieves a 74.98% mean Pixel Accuracy (*mPA*) and a 64.17% mean Intersection over Union (*mIoU*), and can be adapted to segment the captured images (image resolution $2832 * 4256$ pixels). Second, the preliminary segmentation results with 2D images are mapped to 3D point clouds according to the coordinate relationships between the images and the point clouds. Third, based on the mapping results, fine features of buildings are further extracted directly from the 3D point clouds. Our experiments show that the proposed fusion method can segment local and global features efficiently and effectively.

1. Introduction

Compared with object classification, object detection and object recognition, semantic segmentation is a higher-level task that paves the way towards a complete scene understanding in computer vision. It is the pixel-level classification of different objects against a complex background. The importance of semantic segmentation as a core computer vision issue is highlighted by the increasing number of applications that adopt this approach to infer different types of information, including remote-sensing mapping, autonomous driving, indoor navigation, robotics, augmented reality, human-computer interaction, city planning, etc.

Recently, laser scanners have become popular equipment for 3D scene perception due to their stable 3D data capturing capability both at day and night. In combination with digital cameras, 2D images, 2.5D depth images and 3D point clouds can all be captured quickly and efficiently and be used to understand and infer the nature of the 3D world. Meanwhile, they present great challenges for the quick and accurate segmentation of 3D scenes. In addition, with the development of Graphics Processing Units (GPUs), machine learning and the appearance of public 3D point cloud datasets, deep learning has started to be applied to 3D scene segmentation, which breaks through the technical barrier of traditional 3D point cloud segmentation by solving the following problems: (1) the data need to be preprocessed to remove ground points; (2) only one type of object can be extracted at a time; (3) 3D objects need to be extracted using hand-designed features, which depends on the professional knowledge of the researchers; and (4) the processing speed is slow and the combination with Compute Unified Device Architecture (CUDA) is difficult.

The application of deep learning in the survey area has only just begun. 3D objects are extracted, mainly based on 2D projective images. Our study is inspired by the success of deep learning in 2D images. We first tried seven famous semantic segmentation models, compared their performance and evaluated their suitability for different scenes. On this basis, to obtain a Deep Convolution Neural Network (DCNN) suitable for large-scale 3D scenes and high-resolution images ($2832 * 4256$ pixels), we modified and fine-tuned the weights of the publicly available ImageNet-pretrained Deeplabv2-Vgg16 and adapted it to large-scale outdoor scenes and high-resolution images. To ensure the validity of the DVLSHR model, four benchmark datasets (PASCAL VOC12, SIFT-Flow, CamVid and CityScapes) were utilized in the training and validation stages. Two test datasets were captured with a Nikon D700 digital camera and a Riegl VZ-400 laser scanner. Then, the segmentation results of the 2D images were mapped to their corresponding 3D point clouds according to their coordinate transformations. Features of the 3D objects were then coarsely extracted from the 3D point clouds. Until now, not all 3D objects can be segmented well; for example, difficulty in segmenting buildings was encountered. Due to the limitations of image labels, only the outlines of

* Corresponding author at: Information Engineering University, 450001 Zhengzhou, China.
E-mail address: 599154503@qq.com (R. Zhang).

https://doi.org/10.1016/j.isprsjprs.2018.04.022
Received 29 August 2017; Received in revised form 17 April 2018; Accepted 25 April 2018
0924-2716/

buildings were labeled while local structures such as windows, balconies and doors were not. To address this problem, 3D objects with coarse segmentation were further refined with Fuzzy Clustering combined with the Generalized Hough Transformation algorithm (named as the FC-GHT algorithm).

The key contributions of our work are as follows:

(1) 3D point cloud descriptors face many challenges compared with 2D images at present. For semantic segmentation methods based on deep learning, the study of 3D point clouds directly used as input to implement scene segmentation is rare, except for Stanfords PointNet and its extended version PointNet + +. The only publically available 3D dataset is Stanford 2D-3D-S, which consists of indoor scenes rather than urban scenes. However, deep neural network models based on 2D image segmentation are more mature, and there are many more available datasets that can greatly benefit model training. As such, in this work we synchronously acquired 2D images and 3D point clouds. Then, the 2D images were input into the convolutional neural network to get their segmentation results, as discussed in Section 3. According to the style of the 3D point cloud segmentation being assisted by the 2D images, we successfully modified and fine-tuned a DCNN for large-scale scenes and high-resolution images, as discussed in Section 3.1.

(2) The segmentation results of the 2D images were then mapped to 3D point clouds according to their coordinate transformation relationships. We deduced the mapping process suitable for our proposed segmentation method, discussed in Section 3.2.

(3) For the mapping results, only the outlines of each class were segmented, instead of local features. For buildings, the main structures in 3D urban scenes, the extraction was insufficient, and we merely obtained building outlines. Therefore, based on the mapping results, we further extracted the physical planes of the buildings with 3D point clouds using the FC-GHT algorithm, as discussed in Section 3.3. The segmentation results and plane extraction validated the effectiveness of our proposed method.

2. Related works

2.1. Point cloud descriptors

In what manner should the 3D point clouds be represented? A large corpus of shape descriptors has been developed for drawing inferences about 3D objects in deep learning. These descriptors can be classified into four broad categories: methods based on hand-extracted features, 2D projection maps, voxel-based representation and raw point clouds Guo (2017).

Previously, 3D point cloud descriptors were largely “hand-designed” according to the particular geometric properties of a shape’s surface or volume, such as length, width, height, area, reflected intensity, normal vector or curvature (Berthold, 1984; Bu et al., 2014). Hand-designed features need to first be extracted to be input into a Deep Neural Network (DNN) to learn high-layer features accordingly, which still depend on the selected, hand-designed features and parameter optimization; thus, the advantages of deep learning were lost to some extent, and the problem of automatic learning could not be solved.

The second category is view-based descriptors, which describe the shape of a 3D object by “how it looks” in a collection of 2D projections. Murase and Nayar (1995) recognized objects by matching their appearances in parametric Eigen-spaces formed by large sets of 2D renderings of 3D models under varying poses and illuminations. Su et al. (2015) rendered a 3D shape from 12 different views and passed these views through a Convolutional Neural Network (CNN) to extract view-based features. Shi et al. (2015) converted 3D shapes into a panoramic view, namely, a cylindrical projection around its principle axis. Sinha et al. (2016) created geometric images using authalic parametrization in a spherical domain. Kalogerakis et al. (2016) obtained shaded and depth images from different viewpoints and different scales. The flaws of these methods are that the local and global structures are changed, which will reduce their identification performance for various scene features.

The third category is voxel-based representation. Wu et al. (2015) represented a geometric 3D shape as a probability distribution of binary variables on a 3D voxel grid. Xu et al. (2016) created binary images based on different rotations along the *x*-, *y*-, and *z*-axes. Li et al. (2016b) represented 3D shapes as volumetric fields. Qi et al. (2016b) compared CNNs based upon volumetric representations with those based on multi-view representations. Wu et al. (2016) generated 3D objects from a probability space by leveraging recent advances in volumetric convolutional networks and generative adversarial nets. Although these methods completely preserve the 3D shape information, they face some new challenges: (1) the 3D voxel resolution cannot be too high to ensure that the training is not too complex, where a resolution of $30 * 30 * 30$ pixels is usually used. However, too low of a resolution will limit the segmentation performance. (2) The voxel proportion of 3D surfaces is not high, which results in the voxel result being sparse. Therefore, it is necessary to design a reasonable network structure that avoids zero or void operations.

The methods in the fourth category adopt the raw point clouds directly (Vinyals et al., 2015; Qi et al., 2016a). According to the scattered and unstructured characteristics of 3D point clouds, these methods design special network input layers. However, developing classifiers and other supervised machine learning algorithms on top of such 3D shape descriptors poses a number of challenges (Su et al., 2015). First, the number of organized databases with annotated 3D models is rather limited compared with image datasets because generating 3D datasets for segmentation is costly and difficult. Second, 3D shape descriptors (including voxel-based representation) tend to be very high-dimensional, and few deep learning methods can process such data directly. Thus, 3D point cloud datasets are unpopular at present. Third, real point clouds are big, and a single laser scan possesses tens of millions of unstructured points, which constitutes a large computational burden. The main bottleneck is the large number of 3D nearest-neighbor queries, which significantly slows down processing. Instead of computing exact neighborhoods for each point, Hackel et al. (2016) downsampled the entire point clouds to generate a multi-scale pyramid and computed a separate search structure per scale level.

From the above summary, we can see that 3D point cloud descriptors face many challenges compared with 2D images until now. The first category requires more prior knowledge, while a considerable amount of 3D information will be lost or distorted in the second category. The third category bears higher complex and computational burden of data preprocessing, while the fourth category contains a major bottleneck, i.e. the large number of 3D nearest-neighbor queries, which significantly slows down the processing. However, there is no need for image-based DNN to project data into another dimensional space; besides, data preprocessing is much simpler, and there are many more shared datasets, which together greatly benefit model training. In this work, a novel large-scale point cloud segmentation method is proposed, in which 2D images synchronously acquired with 3D point clouds are input into a CNN to obtain preliminary segmentation results.

2.2. Deep CNNs

Until now, to apply deep learning to the semantic segmentation of images, three main parts have been included in the generic framework: (1) 2D images are input; (2) a Fully Convolutional Network (FCN) is used at the front-end of the model to coarsely extract features; and (3) the output from the front-end is optimized by the back-end with a Conditional Random Field/Markov Random Field (CRF/MRF) to obtain the segmentation results. At present, many of the classical semantic segmentation methods for urban complex scenes based on deep learning adopt this frame, such as FCN (Long et al., 2015), SegNet