ISPRS Journal of Photogrammetry and Remote Sensing xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Deep fusion of multi-view and multimodal representation of ALS point cloud for 3D terrain scene recognition

Nannan Qin [a], Xiangyun Hu [a,b,*], Hengming Dai [a]

[a] School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China
[b] Collaborative Innovation Center of Geospatial Technology, Wuhan University, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 2 September 2017
Received in revised form 12 December 2017
Accepted 12 March 2018
Available online xxxx

Keywords:
Deep learning
3D scene recognition
ALS
Multi-view representation
Fusion network

ABSTRACT

Terrain scene category is useful not only for some geographical or environmental researches, but also for choosing suitable algorithms or proper parameters of the algorithms for several point cloud processing tasks to achieve better performance. However, there are few studies in point cloud processing focusing on terrain scene classification at present. In this paper, a novel deep learning framework for 3D terrain scene recognition using 2D representation of sparse point cloud is proposed. The framework has two key components. (1) Initially, several suitable discriminative low-level local features are extracted from airborne laser scanning point cloud, and 3D terrain scene is encoded into multi-view and multimodal 2D representation. (2) A two-level fusion network embedded with feature- and decision-level fusion strategy is designed to fully exploit the 2D representation of 3D terrain scene, which can be trained end-to-end. Experiment results show that our method achieves an overall accuracy of 96.70% and a kappa coefficient of 0.96 in recognizing nine categories of terrain scene point clouds. Extensive design choices of the underlying framework are tested, and other typical methods from literature for related research are compared.

1. Introduction

In recent decades, with the development of airborne laser scanning (ALS) in various applications, several studies were conducted to exploit ALS point cloud in several classification tasks, including DEM filtering (Sithole and Vosselman, 2004; Hu and Yuan, 2016), building extraction (Kim and Shan, 2011; Wu et al., 2016), and point cloud classification (Huang and You, 2016; Li et al., 2017). Due to varying scene complexity, it is difficult to develop general algorithms of point cloud filtering and classification fitting to any type of scenes. 3D terrain scene recognition can provide these tasks with context information, which is important for achieving more accurate classification results by using suitable algorithms or setting of the algorithms in such context. Moreover, 3D terrain scene recognition can be used for classifying landforms, which is important in several geography research areas such as digital terrain analysis (Deng, 2007), and ecological environment study (Hood, 2007). In this regard, developing effective methods to accurately classify 3D terrain scene is important.

Over the last several years, convolutional neural networks (CNNs) (Cun et al., 1990) have become popular in computer vision, due to a series of successful applications in many recognition tasks, such as image classification (He et al., 2016; Huang et al., 2016), object detection (Girshick et al., 2014; He et al., 2017), and semantic segmentation (Long et al., 2015; Ghiasi and Fowlkes, 2016). Naturally, several works are currently aimed at the adaptation of CNN to 3D data. In the 3D shape recognition domain, many CNN-based approaches have been proposed based on different representations. Voxel-based methods (Wu et al., 2015; Maturana and Scherer, 2015b; Garcia-Garcia et al., 2016; Riegler et al., 2017; Wang et al., 2017) represent the 3D shape with regular 3D grid or octree and apply 3D CNN over the voxels for shape recognition. Image-based approaches (Su et al., 2015; Shi et al., 2015; Sinha et al., 2016; Johns et al., 2016) render the 3D shape into a set of 2D images or convert the 3D shape into geometric/panoramic image and feed the images into a 2D CNN. In the ALS point cloud processing domain, works have attempted to apply CNN in different classification tasks. Further, Maturana and Scherer (2015a) performs 3D CNN operations over a volumetric occupancy map built from point cloud for landing zone detection. Huang and You (2016) applies 3D CNN over a volumetric representation of point cloud for semantic labeling. Hu and Yuan (2016) represents each point of the point cloud as a feature map and feeds them to 2D

* Corresponding author at: School of Remote Sensing and Information Engineering, Wuhan University, Wuhan 430079, China.
E-mail address: huxy@whu.edu.cn (X. Hu).

https://doi.org/10.1016/j.isprsjprs.2018.03.011
0924-2716/

CNN for DEM extraction. However, to the best of our knowledge, relevant studies that recognize 3D terrain scene from the ALS point cloud rarely exist, which may be attributed to lack of public 3D scene datasets, especially for the 3D terrain scene dataset from ALS.

In this paper, we first present a deep learning framework for 3D terrain scene recognition using multi-view and multimodal representation of 3D ALS point cloud. This approach is inspired by the success of the preceding image-based approaches purposed for 3D shape recognition. Compared to other kinds of representations, the major benefit of 2D image representation is the direct exploit of well-engineered image-based CNNs (e.g., AlexNet (Krizhevsky et al., 2012)) and the corresponding models pre-trained on large-scale image datasets, such as ImageNet (Deng et al., 2009). Hence, our approach can effectively tackle the problem of limited labeled 3D data at present. The key idea of our method is to represent the irregular 3D LiDAR point cloud as a series of regularly sampled 2D feature maps and apply an image-based CNN to solve the classification problem. To this end, we propose a multi-view and multimodal encoding scheme to obtain a compact and effective representation for sparse 3D point cloud. Then, a two-level fusion network (TLFnet) is designed to fully exploit the multi-view and multimodal representation of 3D terrain scenes. The TLFnet is a unified CNN architecture embedded with feature- and decision-level fusion strategy, which can be trained in an end-to-end manner. We demonstrate the utility of our method on a measured point cloud dataset containing nine categories of terrain scenes. Experimental results demonstrate that our method outperforms three typical methods from the literature. In addition, we evaluate the performance of the proposed framework with extensive underlying ablation experiments. The main contributions of our work are threefold:

(1) We first propose a deep learning framework for 3D terrain scene recognition, which to the best of our knowledge, has not been previously conducted.
(2) We present a multi-view and multimodal representation for ALS point cloud to better encode 3D terrain scenes.
(3) We design a TLFnet, which performs feature- and decision-level fusion over multiple inputs, to fully exploit the power of multi-view and multimodal representation.

The remainder of this paper is organized as follows. Section 2 describes our proposed method. Section 3 presents the experiments, including framework design analysis and comparison with other methods, as well as their results and corresponding analysis. Section 4 draws conclusions and provides several aspects for future research.

2. 3D terrain scene recognition based on deep fusion of multi-view and multimodal representation

Our framework takes a set of feature maps extracted from ALS point cloud as the input of a 2D CNN for 3D terrain scene recognition. As illustrated in Fig. 1, the framework first generates multi-view and multimodal representation of a 3D scene point cloud and then fully exploits the representation via a TLFnet.

2.1. Input: Multi-view and multimodal representation

By representing 3D point cloud with a set of feature maps, we can leverage well-engineered image-based CNNs and public large labeled image datasets to tackle the problem of limited 3D labeled data in 3D terrain scene recognition. Based on this scenario, we propose a multi-view and multimodal representation to encode the spatial sparse ALS point cloud. The entire algorithm can be described as follows.

Algorithm 1. Multi-view and multimodal representation generation.

Input: ALS point cloud
Output: multi-view and multimodal representation
1: Extract a few discriminative low-level features from the point cloud
2: **repeat**
3: Project the point cloud onto an image plane along one of the preset viewing directions
4: Perform grid interpolation operations over the features of the projected point cloud to generate multimodal representation
5: **until** multimodal representations in all preset viewing angles are generated

2.1.1. Feature extraction from point cloud

Our multimodal representation is encoded by three generic and discriminative low-level features: elevation, slope angle, and intensity. Among these features, elevation is the fundamental information. Slope angle is used to enhance the lost geometric information detail of elevation due to normalization, while intensity is used to provide extra spectral feature information. Given a 3D LiDAR point $p = (x, y, z, i)$, we directly assign value z and the intensity value i of point p to elevation feature and intensity feature, respectively. Then, the value of the slope angle feature at point p is calculated as below:

$$slope\ angle = \frac{\sum_{i=1}^{k} \arctan s_i}{k}, \tag{1}$$

where k denotes the number of the neighborhood of the current point, and s_i is the slope between the current point and its i-th neighborhood point. s_i is calculated as follows:

$$s_i = \frac{|z - z_i|}{\sqrt{(x - x_i)^2 + (y - y_i)^2}}, \tag{2}$$

where x_i, y_i, and z_i denote the spatial coordinates of the i-th neighborhood point. An 8-neighborhood is used in this paper.

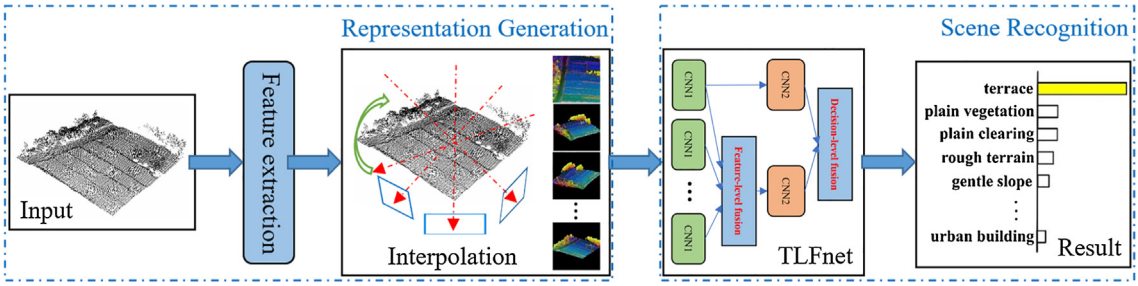

Fig. 1. Framework flow of 3D terrain scene recognition.

Please cite this article in press as: Qin, N., et al. Deep fusion of multi-view and multimodal representation of ALS point cloud for 3D terrain scene recognition. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.03.011