Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Developing a multi-filter convolutional neural network for semantic segmentation using high-resolution aerial imagery and LiDAR data

Ying Sun[a], Xinchang Zhang[b,*], Qinchuan Xin[a,c,*], Jianfeng Huang[a]

[a] Department of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China
[b] School of Geographical Sciences, Guangzhou University, Guangzhou 510006, China
[c] Guangdong Key Laboratory for Urbanization and Geo-simulation, Guangzhou 510275, China

ARTICLE INFO

Keywords:
LiDAR
High-resolution imagery
Multi-modal fusion
Multi-resolution segmentation
Semantic segmentation

ABSTRACT

Semantic segmentation of LiDAR and high-resolution aerial imagery is one of the most challenging topics in the remote sensing domain. Deep convolutional neural network (CNN) and its derivatives have recently shown the abilities in pixel-wise prediction of remote sensing data. Many existing deep learning methods fuse LiDAR and high-resolution aerial imagery towards an inter-modal mode and thus overlook the intra-modal statistical characteristics. Additionally, the patch-based CNNs could generate the salt-and-pepper artifacts as characterized by isolated and spurious pixels on the object boundaries and patch edges leading to unsatisfied labelling results. This paper presents a semantic segmentation scheme that combines multi-filter CNN and multi-resolution segmentation (MRS). The multi-filter CNN aggregates LiDAR data and high-resolution optical imagery by multi-modal data fusion for semantic labelling, and the MRS is further used to delineate object boundaries for reducing the salt-and-pepper artifacts. The proposed method is validated against two datasets: the ISPRS 2D semantic labelling contest of Potsdam and an area of Guangzhou in China labelled based on existing geodatabases. Various designs of data fusion strategy, CNN architecture and MRS scale are analyzed and discussed. Compared with other classification methods, our method improves the overall accuracies. Experiment results show that our combined method is an efficient solution for the semantic segmentation of LiDAR and high-resolution imagery.

1. Introduction

With the rapid development of remote sensing technology, advanced airborne sensors such as optical sensors and Laser Detection and Ranging (LiDAR) could provide high-resolution remote sensing (HRRS) images at the sub-meter and even centimeter spatial resolution. As a research frontier in the field of remote sensing, classification of high-resolution images as well as LiDAR point cloud data plays an important role in a wide range of applications such as urban planning, environmental monitoring, forestry and agriculture management and land inventory. Among various classification approaches, semantic segmentation is a common one that interprets the remote sensing images at the pixel level (i.e., making a prediction for each individual pixel). However, the complexity of high-resolution images and LiDAR data in spatial and spectral patterns makes semantic segmentation a challenging task.

Machine learning approaches that use support vector machine (SVM, Vapnik and Vapnik, 1998), AdaBoost (Freund and Schapire, 1995), random forest (RF, Ho, 1998), and artificial neural network (ANN, Fukushima et al., 1983) have been widely developed for semantic segmentation. In these approaches, LiDAR point cloud data are commonly converted to range images and then concatenated with raw images to obtain images that are more informative than either of the data sources (Zhang, 2010). However, accurate feature representation for the concatenated images is essential for pixel-wise prediction in the above-mentioned machine learning approaches (Zhou et al., 2016). To address this issue, many hand-crafted intra-modal features based on spectral signals, geometry, height and texture have been extracted (Stumpf and Kerle, 2011), such as spectral indices that highlight certain objects and geometry features extracted by the morphological profile (Liao et al., 2015), scale-invariant feature transform (Lowe, 2004), local binary patterns (Ojala et al., 2002), histogram of oriented gradients (Dalal and Triggs, 2005), the bag of visual words model (Yang and Newsam, 2010) and sparse representation (Han et al., 2014). These intra-modal features extracted from the two data sources, i.e., feature level fusion, are superior to the raw data (Wang et al., 2007). However, due to the heterogeneous appearance and large intra-class variance characteristics of the high-resolution imagery and LiDAR data, the above studies only extracted shallow features that were low-level or middle-level, which are not representative enough. A higher

[*] Corresponding authors at: Department of Geography and Planning, Sun Yat-Sen University, Guangzhou 510275, China (Q. Xin).
E-mail addresses: eeszxc@mail.sysu.edu.cn (X. Zhang), xinqinchuan@gmail.com (Q. Xin).

https://doi.org/10.1016/j.isprsjprs.2018.06.005
Received 27 September 2017; Received in revised form 18 February 2018; Accepted 6 June 2018

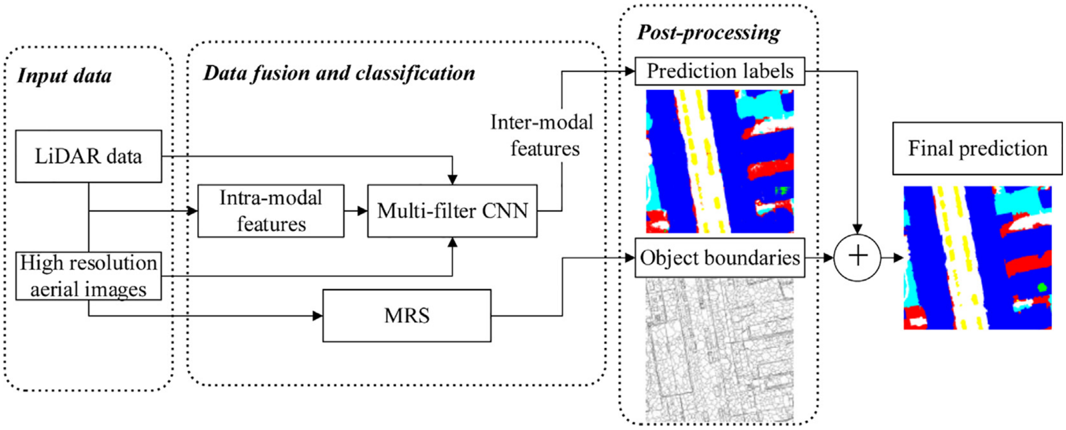

Fig. 1. The flowchart of the semantic segmentation method that consists of multi-filter CNN and multi-resolution segmentation.

level of abstract features is more discriminating and will be helpful for the improvement of semantic segmentation (Zhang et al., 2016).

Deep learning such as convolutional neural networks (CNN) has received an increasing amount of attention (Gupta et al., 2014; Long and Jin, 2008; Simonyan and Zisserman, 2014; Girshick et al., 2014), and has been applied in semantic labelling of remote sensing imagery (Längkvist et al., 2016; Islam et al., 2017; Lin et al., 2017; Zhao et al., 2017). CNN could fuse the high-resolution imagery and LiDAR data in the inter-modal way and extract high-level features that outperform hand-crafted intra-modal features. Volpi and Tuia (2017) presented an architecture with full patch labeling by learned upsampling (CNN-FPL) using the concatenated imagery and LiDAR data. Sherrah (2016) use the 3-band image data as input, and proposed a hybrid network that combines the pre-trained image features with DSM. Liu et al. (2017) proposed a decision-level scheme for data fusion of image and LiDAR, in which a CNN is trained based on the CIR images; and the inter-modal trained features and the hand-crafted features are combined in the final CRF framework. However, CNN with a fixed scale often limits the receptive field and makes feature extraction difficult. Unlike fixed CNNs, multi-scale CNNs consider multiple scales to capture different information for HRRS classification. They come in three flavors: (i) methods that use the same resolution input images with different patch sizes (Paisitkriangkrai et al., 2016); (ii) methods that use different resolution input images of the same geographical area (Zhao and Du, 2016); and (iii) methods that use CNN with different kernel sizes (Audebert et al., 2016). For the first two approaches, different kinds of input data should be prepared that cannot be directly used in the encoder-decoder CNN architecture as the input images and the corresponding labelled images are of different resolutions. For the third method, multi-scale CNN with different kernels is trained separately for earth observation data classification, and the losses of the three CNNs are averaged for error propagation. In general, there remain two defects that need improvement: (1) Loss averaging may introduce errors from single-kernel CNN and thus influence the correct weights updating; and (2) the existing multi-scale CNNs only use the inter-modal features extracted based on CNN, while intra-modal structures inferred preciously can often help the higher-level features to be mined more accurately. In addition, although the encoder-decoder CNN architecture up-samples the low-resolution features derived from pooling layers to the input resolution (Badrinarayanan et al., 2015; Chen et al., 2016; Long et al., 2015), the object boundaries are blurred irreversibly because the upsampling layers reconstruct the appearance of the object rather than the shape. Whereas CNNs often use patch-based images for classification owing to the computational ability, there is a lack of contextual information for pixels near the patch edges, resulting in the salt-and-pepper artifacts near the patch edges when mosaicking images (Mnih, 2013).

To overcome these problems, we develop a method that combines multi-filter CNN and MRS post-processing for semantic segmentation of high-resolution aerial imagery and LiDAR data. The multi-filter CNN employs three parallel CNNs with filters of different spatial context size, and a two-route loss function is employed for weight updating. LiDAR data and imagery are fused in the multi-filter CNN for multi-modal features extraction and classification. MRS is then used to delineate boundaries of objects and eliminate the salt-and-pepper artifacts. Two datasets, i.e., the Potsdam dataset in the ISPRS 2D labelling contest and the Guangzhou dataset in China, are used for method assessment.

2. Methods

The proposed method of semantic segmentation, as shown in Fig. 1, mainly consists of multi-filter CNN and multi-resolution segmentation. An end-to-end multi-filter CNN is first built for the classification of fused high-resolution aerial imagery and LiDAR data, and the method of multi-resolution segmentation is then applied to delineate object boundaries and refine the results.

2.1. A brief description of the convolution neural network

The convolutional neural network (CNN) is typically comprised of several convolutional stages. Each convolutional stage consists of multiple layers such as the convolutional layer, the activation function layer (usually a rectified linear unit, ReLU), the pooling layer, and the optional layer of batch normalization (BN). In this paper, we use a SegNet-like CNN which has a convolutional-deconvolutional structure, in which the deconvolutional process is to up-sample the input feature maps that are down-sampled by the pooling layers in the convolutional stage. Each deconvolutional stage is usually composed of the upsampling layer, the convolutional layer, and the optional layer of batch normalization.

2.2. A multi-filter convolutional neural network

The size of the receptive field determines the observation scale greatly and affects the prediction results accordingly. As traditional CNN adopts fixed filter sizes and hence limited observation scales (Zhao and Du, 2016), ensemble methods that apply the multi-scale technique are favorable in practice. To explore the multi-resolution features of local and global contexts, we develop a multi-filter CNN that consists of three different filters (i.e., 3 × 3, 5 × 5, and 7 × 7 in parallel) for data fusion and semantic segmentation.

As both high-resolution aerial images and LiDAR data are involved, the method first learns the intra-modal features from each individual data source and then extracts the inter-modal features using the multi-