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

A multi-scale fully convolutional network for semantic labeling of 3D point clouds

Mohammed Yousefhussien [a,*], David J. Kelbe [b], Emmett J. Ientilucci [a], Carl Salvaggio [a]

[a] Rochester Institute of Technology, Chester F. Carlson Center for Imaging Science, Rochester, NY, USA
[b] Oak Ridge National Laboratory, Geographic Information Science and Technology Group, Oak Ridge, TN, USA

ARTICLE INFO

Article history:
Received 2 October 2017
Received in revised form 9 March 2018
Accepted 16 March 2018
Available online xxxx

Keywords:
LiDAR
3D-labeling contest
Deep learning

ABSTRACT

When classifying point clouds, a large amount of time is devoted to the process of engineering a reliable set of features which are then passed to a classifier of choice. Generally, such features – usually derived from the 3D-covariance matrix – are computed using the surrounding neighborhood of points. While these features capture local information, the process is usually time-consuming and requires the application at multiple scales combined with contextual methods in order to adequately describe the diversity of objects within a scene. In this paper we present a novel 1D-fully convolutional network that consumes terrain-normalized points directly with the corresponding spectral data (if available) to generate point-wise labeling while implicitly learning contextual features in an end-to-end fashion. This unique approach allows us to operate on unordered point sets with varying densities, without relying on expensive hand-crafted features; thus reducing the time needed for testing by an order of magnitude over existing approaches. Our method uses only the 3D-coordinates and three corresponding spectral features for each point. Spectral features may either be extracted from 2D-georeferenced images, as shown here for Light Detection and Ranging (LiDAR) point clouds, or extracted directly for passive-derived point clouds, *i.e.* from multiple-view imagery. We train our network by splitting the data into square regions and use a pooling layer that respects the permutation-invariance of the input points. Evaluated using the ISPRS 3D Semantic Labeling Contest, our method scored second place with an overall accuracy of 81.6%. We ranked third place with a mean F1-score of 63.32%, surpassing the F1-score of the method with highest accuracy by 1.69%. In addition to labeling 3D-point clouds, we also show that our method can be easily extended to 2D-semantic segmentation tasks, with promising initial results.

1. Introduction

The past decade of computer/machine vision research and remote sensing hardware development has broadened the availability of 3D point cloud data through innovations in Light Detection and Ranging (LiDAR), Synthetic Aperture Radar (SAR), dense stereo- or multiview-photogrammetry and structure from motion (SfM). Despite the prevalence of 3D-point cloud data, automated interpretation and knowledge discovery from 3D-data remains challenging due to the irregular structure of raw point clouds. As such, exploitation has typically been limited to simple visualization and basic mensuration (Hackel et al., 2016). Or, some authors rasterized the point cloud onto a more tracTable 2.5D- Digital Surface Model (DSM) from which conventional image processing techniques are applied, *e.g.* (Hug and Wehr, 1997; Haala et al., 1998).

In order to generate exploitation-ready data products directly from the point cloud, semantic classification is desired. Similar to per-pixel image labeling, 3D-semantic labeling seeks to attribute a semantic classification label to each 3D-point. Classification labels, *e.g.* vegetation, building, road, etc., can subsequently be used to inform derivative processing efforts such as surface fitting (Xing et al., 2017), 3D modeling (Moussa and El-Sheimy, 2010), object detection (Jochem et al., 2009), and bare-earth extraction (Yunfei et al., 2008). However, the task of labeling every data point in the irregularly distributed point cloud captured by aerial platforms is challenging, especially in urban scenes with different object types and various scales ranging from very small spatial neighborhoods (power lines) to very large spatial neighborhoods (buildings). Moreover, point clouds are unstructured and unordered data with variable spatial densities. In order to scale the

* Corresponding author.
E-mail address: myhussien@gmail.com (M. Yousefhussien).

https://doi.org/10.1016/j.isprsjprs.2018.03.018
0924-2716/

Please cite this article in press as: Yousefhussien, M., et al. A multi-scale fully convolutional network for semantic labeling of 3D point clouds. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.03.018

semantic classification task to meet the demands of emerging data volumes potentially at sub-meter resolution and global in coverage an efficient, streamlined, and robust model that directly operates on 3D point clouds is needed. The goal of this research is to introduce a flexible and simple multi-scale deep learning framework for direct semantic labeling of 3D aerial point clouds, thus eliminating the need for calculating costly, handcrafted features. The algorithm respects the permutation-invariance of input points and therefore avoids the need to transform the points to images or volumes.

2. Related work

Point cloud labeling algorithms can generally be grouped into two main categories. Section 2.1 describes "Direct Methods", which operate immediately on the point clouds themselves and do not change the 3D-nature of the data. Section 2.2 describes "Indirect Methods", which transform the input point cloud into an image or a volume as a preconditioning step to more traditional (raster-based) segmentation approaches. Considering the relative trade-offs of these techniques, Section 2.3 proposes a novel approach with 7 specific contributions for semantic classification of point clouds.

2.1. Direct methods

Direct methods assign semantic labels to each element in the point cloud based on a simple point-wise discriminative model operating on point features. Such features, known as "eigen-features", are derived from the covariance matrix of a local neighborhood and provide information on the local geometry of the sampled surface, *e.g.* planarity, sphericity, linearity (Lin et al., 2014a). To improve classification, contextual information can explicitly be incorporated into the model. For example, Blomley et al. (2016) used covariance features at multiple scales found using the eigenentropy-based scale selection method (Weinmann et al., 2014) and evaluated four different classifiers using the ISPRS 3D Semantic Labeling Contest.[1] Their best-performing model used a Linear Discriminant Analysis (LDA) classifier in conjunction with various local geometric features. However, scalability of this model was limited due to the dependence upon various handcrafted features and the need to experiment with various models that don't incorporate contextual features and require effort to tune.

Motivated by the frequent availability of coincident 3D data and optical imagery, Ramiya et al. (2014) proposed the use of point coordinates and spectral data directly, forming a per-point vector of (X, Y, Z, R, G, B) components. Labeling was achieved by filtering the scene into ground and non-ground points according to Axelsson (2000), then applying a 3D-region-growing segmentation to both sets to generate object proposals. Like Blomley et al. (2016), several geometric features were also derived, although specific details were not published. Without incorporating contextual features, each proposed segment was then classified according to the five classes from the ISPRS 3D Semantic Labeling Contest.

Alternatively, Mallet (2010) classified full-waveform LiDAR data using a point-wise multiclass support vector machine (SVM). And (Chehata et al., 2009) used random forests (RF) for feature detection and classification of urban scenes collected by airborne LiDAR. The reader is referred to Grilli et al. (2017) for a more complete review of discriminative classification models. While simple discriminative models are well-established, they are unable to consider interactions between 3D points.

To allow for spatial dependencies between object classes by considering labels of the local neighborhood, Niemeyer et al. (2014) proposed a contextual classification method based on Conditional Random Field (CRF). A linear and a random forest model were compared when used for both the unary and the pairwise potentials. By considering complex interactions between points, promising results were achieved, despite the added cost of computation speed: 3.4 min for testing using an RF model, and 81 min using the linear model. This computation time excludes the additional time needed to estimate the per-point, 131-dimensional feature vector prior to testing.

This contextual classification model was later extended to use a two-layer, hierarchical, high-order CRF, which incorporates spatial and semantic context (Niemeyer et al., 2016). The first layer operates on the point level, utilizing higher-order cliques and geometric features (Weinmann et al., 2014) to generate segments. The second layer operates on the generated segments, and therefore incorporates a larger spatial scale. Features included geometric- and intensity-based descriptors, in addition to distance and orientation to road features (Golovinskiy et al., 2009). By iteratively propagating context between layers, incorrect classifications can be revised at later stages; this resulted in good performance on a 2.25 million point dataset of Hannover, Germany. However, this method employed multiple algorithms, each designed separately, which would make simultaneously optimization challenging. Also, the use of computationally-intensive inference methods limits the run-time performance. In contrast to relying on multiple individually-trained components, an end-to-end learning mechanism is desired.

2.2. Indirect methods

Indirect methods – which mostly rely on deep learning – offer the potential to learn local and global features in a streamlined, end-to-end fashion (Yosinski et al., 2015). Driven by the reintroduction and improvement of Convolutional Neural Networks (CNNs) (LeCun et al., 1989; He et al., 2016), the availability of large-scale datasets (Deng et al., 2009), and the affordability of high-performance computing resources such as graphics processing units (GPUs), deep learning has enjoyed unprecedented popularity in recent years. This success in computer vision domains such as image labeling (Krizhevsky et al., 2012), object detection (Girshick et al., 2014), semantic segmentation (Badrinarayanan et al., 2017; Long et al., 2015), and target tracking (Wang and Yeung, 2013; Yousefhussien et al., 2016), has generated an interest in applying these frameworks for 3D classification.

However, the nonuniform and irregular nature of 3D-point clouds prevents a straightforward extension of 2D-CNNs, which were originally designed for raster imagery. Hence, initial deep learning approaches have relied on *transforming* the 3D data into more tracTable 2D images. For example, Su et al. (2015) rendered multiple synthetic "views" by placing a virtual camera around the 3D object. Rendered views were passed though replicas of the trained CNN, aggregated using a view-pooling layer, and then passed to another CNN to learn classification labels. Several other methods use the multiview approach with various modifications to the rendered views. For example, Bai et al. (2016) generated depth images as the 2D views, while other methods accumulated a unique signature from multiple view features. Still other methods projected the 3D information into 36 channels, modifying AlexNet (Krizhevsky et al., 2012) to handle such input. For further details, the reader is referred to (Savva et al., 2016).

Similar multiview approaches have also been applied to ground-based LiDAR point clouds. For example, Boulch et al. (2017) generated a mesh from the Semantic3D Large-scale Point Cloud Classification Benchmark (Hackel et al., 2017); this allowed for the generation of synthetic 2D views based on both RGB information and a 3-channel depth composite. A two-stream Seg-

[1] https://goo.gl/fSK6Fy.

Please cite this article in press as: Yousefhussien, M., et al. A multi-scale fully convolutional network for semantic labeling of 3D point clouds. ISPRS J. Photogram. Remote Sensing (2018), https://doi.org/10.1016/j.isprsjprs.2018.03.018