ISPRS Journal of Photogrammetry and Remote Sensing 130 (2017) 139–149

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Simultaneous extraction of roads and buildings in remote sensing imagery with convolutional neural networks

Rasha Alshehhi [a,*], Prashanth Reddy Marpu [a], Wei Lee Woon [a], Mauro Dalla Mura [b]

[a] Institute Center for Smart and Sustainable Systems, Masdar Institute of Science and Technology, Abu Dhabi, United Arab Emirates
[b] GIPSA-lab, Grenoble Institute of Technology, Grenoble, France

ARTICLE INFO

Article history:
Received 3 January 2017
Received in revised form 29 April 2017
Accepted 2 May 2017

Keywords:
Convolutional neural network
Low-level features
Adjacent regions
Extraction

ABSTRACT

Extraction of man-made objects (e.g., roads and buildings) from remotely sensed imagery plays an important role in many urban applications (e.g., urban land use and land cover assessment, updating geographical databases, change detection, etc). This task is normally difficult due to complex data in the form of heterogeneous appearance with large intra-class and lower inter-class variations. In this work, we propose a single patch-based Convolutional Neural Network (CNN) architecture for extraction of roads and buildings from high-resolution remote sensing data. Low-level features of roads and buildings (e.g., asymmetry and compactness) of adjacent regions are integrated with Convolutional Neural Network (CNN) features during the post-processing stage to improve the performance. Experiments are conducted on two challenging datasets of high-resolution images to demonstrate the performance of the proposed network architecture and the results are compared with other patch-based network architectures. The results demonstrate the validity and superior performance of the proposed network architecture for extracting roads and buildings in urban areas.

1. Introduction

Automated extraction of multiple urban objects from aerial and satellite imagery is an important step in many applications such as infrastructure planning, updating geographical databases, land use analysis and change detection. Classifying pixels into semantic objects in aerial and satellite images of urban areas is one of the most challenging and important problems. This is because the remote sensing images are usually characterized by complex data in the form of heterogeneous regions with large intra-class variations and often lower inter-class variations. This is even more prominent with urban objects such as buildings and roads. Such heterogeneity in remote sensing images restricts most of the existing methods that often depend on a set of predefined features, which in turn are extracted using tunable parameters. As a result, it is highly difficult to design a method which could achieve high accuracy, especially with increasing spatial resolutions. Recently, deep Convolutional Neural Network (CNN) architectures (Krizhevsky et al., 2012; Lin et al., 2014; Zeiler and Fergus, 2014; Szegedy et al., 2015; Liu and Deng, 2015; Zhou et al., 2014) have been achieving impressive state-of-the art performance for semantic classification, not only in remote sensing applications (Firat et al., 2014; Makantasis et al., 2015; Paisitkriangkrai et al., 2015; Kampffmeyer Michael and Jenssen, 2015; Jiang et al., 2015; Vakalopoulou et al., 2015; Castelluccio et al., 2015; Sherrah, 2016; Nogueira et al., 2016), but also in some networks that have been proposed in computer vision, e.g., fully connected network (Long et al., 2015), SegNet (Badrinarayanan et al., 2015), ReSeg (Visin et al., 2015), DeepLab (Chen et al., 2015; Papandreou et al., 2015), Deconvolutional network (Noh et al., 2015), Decoupled Network (Hong et al., 2015), Patch Network (Brust et al., 2015; Brust et al., 2015), Deep Parsing Network (Liu et al., 2015), integrated CNN with CRFs (Chen et al., 2015; Zheng et al., 2015; Lin et al., 2016) and combined CNN with segmentation (Zhao et al., 2015; Kim et al., 2015).

Deep CNN architecture is quickly becoming prominent in remote sensing applications since it has the ability to effectively encode spectral and spatial information based on the input image data, without any prepossessing step. It consists of multiple interconnected layers and learns a hierarchical feature representation from raw pixel-data. It discovers features in multiple levels of

* Corresponding author.
E-mail addresses: ralshehhi@masdar.ac.ae (R. Alshehhi), pmarpu@masdar.ac.ae (P.R. Marpu), wwoon@masdar.ac.ae (W.L. Woon), mauro.dalla-mura@gipsa-lab.grenoble-inp.fr (M.D. Mura).

http://dx.doi.org/10.1016/j.isprsjprs.2017.05.002

representations. The lowest level is depicted by the primitive features of pixels (e.g., spectral properties) and the higher level involves transforming from raw pixel representation into gradually more abstract representations that are invariant to small geometric variations (e.g., edges and corners), and further transforming them gradually to make them invariant to contrast changes and contrast inversion (e.g., object parts). At the end, the most frequent patterns related to more abstract categories associated with whole objects are identified.

There have been several methods of CNN architectures in remote sensing. Paisitkriangkrai et al. Paisitkriangkrai et al., 2015 combined simple features (e.g., Digital Surface Model (DSM) and Normalized DSM) with multi-resolution CNN features to detect multiple classes using multiple binary classifiers. They applied multi-class concatenation classifier on CNN features and then applied pixel-based Conditional Random Field (CRF) classifier as the post-processing stage to smoothen the final pixel-based classification. In Sherrah, 2016, all convolution layers in CNNs are replaced with fully connected layers, and down-sampling pooling is replaced with no down-sampling pooling. Kampffmeyer Michael and Jenssen (2015) applied similar deep CNN architecture to extract small objects (e.g., cars), which have lower class distribution by combination with deconvolution layers. In Jiang et al. (2015), graph-based segmentation (Felzenszwalb and Huttenlocher, 2004) is integrated with CNNs to localize image patches,[1] which help in localizing vehicles effectively. In Lngkvist et al. (2016), CNNs are integrated with spectral features of Simple Linear Iterative Clustering (SLIC) segmentation (Achanta et al., 2012) in the post processing stage to improve the performance of CNNs.

In general, CNN architectures for semantic pixel-based classification use two main approaches: patch-based and pixel to pixel based (end to end). Patch-based methods commonly start with training of CNN classifier on small image patches and then predict the class of each pixel, using a sliding window approach. Alternatively, the fully connected layers can be converted to convolution layers, avoiding overlapping computations required for each pixel (Paisitkriangkrai et al., 2015; Sherrah, 2016). This approach is usually used to detect large urban objects. Pixel-based methods use an end to end CNN, where usually Fully Convolutional Network (FCN) or encoder-decoder architectures are used by applying upsampling, interpolation, etc. (Jiang et al., 2015; Lngkvist et al., 2016). This approach is important to detect fine detail of the input images.

In this work, we propose a modified patch-based CNN architecture to simultaneously extract roads and buildings from satellite imagery by replacing fully connected layers with Global Average Pooling (GAP) (Lin et al., 2014; Szegedy et al., 2015; Zhou et al., 2015), which considers an average of all feature maps from the last convolution layer of the CNN. We concentrate on roads and buildings because these classes make up a large portion of urban fabric. Moreover, they exhibit a significant amount of urban structure that can be exploited to improve classification in noisy data. As a post-processing step, Simple Linear Iterative Clustering (SLIC) segmentation (Achanta et al., 2012) is applied on the CNN probability map. The shape features of adjacent SLIC regions of roads and buildings are used to link discontinuous road segments and to merge misclassified regions of buildings.

The remainder of this paper is organized as follows. Section 2 introduces some of the related works that used CNNs in extracting roads and buildings and other works which are more related to the proposed CNN architecture. An overview of the proposed CNN is presented in Section 3. Section 4 presents the experimental results and Section 5 summarizes the most important findings.

[1] Image windows with predefined dimensions.

2. Related works and contribution

In this section, some promising CNN approaches for extracting roads and buildings from aerial imagery are discussed, highlighting their main contributions. Some related CNN architectures in computer vision applications are also summarized and finally contributions of this paper are outlined.

There is a significant amount of literature on semantic pixel-based classification for extraction of roads and buildings in remote sensing imagery. Mnih (2013) proposed a road extraction method based on patch-based CNN. The CNN input is extracted from Principal Component Analysis (PCA) features. Then the PCA vectors are trained by the Restricted Boltzmann Machine (RBM) and refined by a post-processing network to incorporate structure such as road connectivity into the final road network. Shu (2014) illustrated the main differences between the performance of the CNN architecture and image segmentation on the same dataset. Saito and Aoki (2015), Saito et al. (2016) used a single CNN architecture for extracting roads and buildings on the Mnih imagery dataset (Mnih, 2013), where each image consists of RGB channels. The CNN predicts a multi-class probability output of roads, buildings and background simultaneously. They also applied Channel-wise Inhibited Softmax (CIS) function to suppress the effect of the background.

Maggiori et al. (0000) suggested a similar architecture as (Shu, 2014) to detect buildings. However, they used a pixel-based approach by applying deconvolution operators, which uses upsampling into the initial resolution to produce dense pixel-based classification. To detect buildings, Marcu and Leordeanu (0000) proposed a dual-stream deep network model to extract roads and buildings separately based on Alex-Net (Liu and Deng, 2015)[2] and VGG-Net (Liu and Deng, 2015).[3] Alex-Net considers information from large areas around the object of interest due to the larger filter size. VGG-Net network focuses on local and object level information due to the smaller filter size. Both networks are combined into final subnet, composed of three Fully Connected (FC) layers.

In computer vision, CNN architectures such as Network in Network (NIN) (Lin et al., 2014) and GoogLetNet (Szegedy et al., 2015) proposed avoiding the use of Fully connected layers to minimize the number of parameters while maintaining the high performance. In Lin et al. (2014) and Szegedy et al. (2015), global average pooling is used to act as a structural regularizer, preventing overfitting during training. Zhou et al. (2015) revisited (Lin et al., 2014) and showed that convolution units have the ability to localize objects in convolution layers; however, this ability is lost when fully connected layers are used. Therefore, they use global average pooling and are able to achieve lower error for object localization. Another similar approach is based on global maximum pooling by Oquab et al. (2015). They applied global maximum pooling to localize a point lying on object boundaries, rather than the complete extent of the objects.

In this work, we follow the same approach as (Lin et al., 2014; Szegedy et al., 2015; Zhou et al., 2015; Oquab et al., 2015) for extracting roads and buildings from two challenging datasets with different spatial image resolutions and different conditions. This work proposes a multi-class prediction method with a single CNN architecture by predicting three different classes simultaneously

[2] Alex-Net, proposed by Krizhevsky et al. (2012), was the winner of ImageNet Large Scale Visual Recognition Challenge (ILSVRC) Deng et al., 2009. It consists of five convolution layers, some of which are followed by max-pooling layers, and three fully connected layers with a final softmax.

[3] VGG-Net, presented in (Liu and Deng, 2015), won the localization and classification tracks of the ILSVRC-2014 competition. It has thirteen convolution layers, five pooling ones and three fully connected one with a final softmax.