ISPRS Journal of Photogrammetry and Remote Sensing 113 (2016) 155–165

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

Learning multiscale and deep representations for classifying remotely sensed imagery

Wenzhi Zhao, Shihong Du *

Institute of Remote Sensing and GIS, Peking University, Beijing 100871, China

ARTICLE INFO

Article history:
Received 23 July 2015
Received in revised form 11 January 2016
Accepted 12 January 2016
Available online 1 February 2016

Keywords:
Multiscale convolutional neural network (MCNN)
Deep learning
Feature extraction
Remote sensing image classification

ABSTRACT

It is widely agreed that spatial features can be combined with spectral properties for improving interpretation performances on very-high-resolution (VHR) images in urban areas. However, many existing methods for extracting spatial features can only generate low-level features and consider limited scales, leading to unpleasant classification results. In this study, multiscale convolutional neural network (MCNN) algorithm was presented to learn spatial-related deep features for hyperspectral remote imagery classification. Unlike traditional methods for extracting spatial features, the MCNN first transforms the original data sets into a pyramid structure containing spatial information at multiple scales, and then automatically extracts high-level spatial features using multiscale training data sets. Specifically, the MCNN has two merits: (1) high-level spatial features can be effectively learned by using the hierarchical learning structure and (2) multiscale learning scheme can capture contextual information at different scales. To evaluate the effectiveness of the proposed approach, the MCNN was applied to classify the well-known hyperspectral data sets and compared with traditional methods. The experimental results shown a significant increase in classification accuracies especially for urban areas.

1. Introduction

Recently, the highly developed remote sensing techniques can provide very-high-resolution (VHR), accurate images in both spatial and spectral domains. Such detailed ground information can now be used for urban planning, environmental protection and crop monitoring, etc. These applications often require very fine accuracy of classification. However, due to the complex properties in both spatial and spectral domains, it is commonly recognized that higher resolutions do not naturally result in higher classification accuracies (Huang and Zhang, 2013), especially when dealing with urban area structures. Therefore, many applications remain to be explored and novel algorithms are required to handle the complex properties of VHR images (Fauvel et al., 2012). For remote sensing image classification, two major issues increase the difficulties in image interpretation. On the one hand, with such high spatial resolution, many small objects and materials emerge in images. This increases the intra-class variation, while decreases the inter-class separability. On the other hand, the high-dimension of spectral information provided in such detailed images (especially, for hyperspectral images) leads to "Hughes phenomenon" (Hughes, 1968), making classification more difficult. Moreover, various structures in urban areas are made of the same materials, leading to that classifiers are ineffective if spectral information is used alone. Based on the two difficulties, it is widely desired to explore effective spatial features and integrate them with spectral information for improving the performance of image interpretation.

Intensive studies have been reported on spectral–spatial feature exploration and accurate classification of hyperspectral images. Bayesian models (Landgrebe, 2005), feature extraction and feature reduction techniques, neural networks (Ratle et al., 2010) and kernel methods (Schölkopf and Smola, 2002) have been investigated for the classification of hyperspectral images. In particular, the kernel-based methods use the kernel trick to separate different classes in a high dimensional space by means of nonlinear transformation and show remarkable performance in hyperspectral image classification. Instead of using a single kernel, the composite kernel (CK)-based (Camps-Valls et al., 2006) methods combining different kernels (e.g., spectral kernel and spatial kernel) become a new trend to exploring the spectral–spatial conjoint properties in hyperspectral image classification. Following a different strategy, several statistical classifiers that simultaneously use contextual and spectral information have been proposed. For instance,

* Corresponding author.
E-mail address: dshgis@hotmail.com (S. Du).

http://dx.doi.org/10.1016/j.isprsjprs.2016.01.004

Jackson and Landgrebe (2002) proposed an iterative statistical classifier on Markov random field (MRF) to exploit the continuity of the neighboring labels. However, the observed spectral vectors are assumed to be conditional independent in MRF models which neglect the contextual information in the observed data. To handle this problem, Zhong and Wang (2010) proposed a discriminative model by using the conditional random field (CRF) to incorporate the contextual information with spectral ones. However, both MRF and CRF-based methods need to define a large neighbor system which imposes intractable computational problems, thereby limiting the benefits of such statistical methods. Rather than defining a crisp neighbor set for every pixel, morphological filters (Benediktsson et al., 2003) were used to adaptively define the neighborhoods of pixels. Then, SVM and CK methods were used to combine the spatial and spectral information during the classification process. Li et al. (2012) combined spectral and spatial information to hyperspectral image segmentation. In this work, the multinomial logistic regression (MLR) and a subspace projection method were combined to learn posterior probability distributions. Recently, multiple feature learning (Li et al., 2015) have been proved to be more effective than CK-based methods as it can exploit the specific physical or acquisition conditions of each feature.

Unfortunately, the above methods either using spectral information of neighboring pixels or morphological properties as spatial features which are low-level ones and require to be specified with empirical parameters (e.g., neighbor size and orientation), leading to that the detected spatial features strongly depend on the experience and parameter setting. Therefore, it is difficult to find appropriate parameters to generate appropriate features for each type of objects. As conformed in neuroscience, the reason for well performance of human brain in object recognition is the ability of hierarchical feature generalization at high levels (DiCarlo et al., 2012). Specifically, high-level spatial features generated by hierarchical abstraction is similar to the principle of human visual system (Malach et al., 1995), thus they show impressive stability and effectiveness in image classification. Deep learning (Lee et al., 2009; Lin and Nevatia, 1998) probably is the best way to extract such robust and high-level spatial features because of its hierarchical learning framework. It can learn non-linear spatial filters automatically and generalize the high-level features from low-level ones, layer by layer. In the field of remote sensing imagery classification, stacked auto-encoder (Hinton and Salakhutdinov, 2006) framework was firstly introduced into hyperspectral image classification (Yushi et al., 2014). However, the auto-encoder framework can only handle 1-dimension input features (spectral features), but overlooks spatial patterns lying in images. To overcome this problem, convolutional neural networks (CNN) was introduced into vehicle detection in high resolution remote sensing imagery (Chen et al., 2014) by considering contextual and structural information in spatial domain. In this work, only one specific type of vehicles was extracted by CNN. Furthermore, scale variation is quite common for objects detection in remote sensing images (e.g., roofs with different sizes), requiring to consider multiscale contextual and structural information in spatial domain. To solve this problem, Zhao et al. (2015) explored deep features at pixel-level through multiscale convolutional auto-encoder in an unsupervised way. However, the auto-learned multiscale features are lack of class-specific meaning and not always effective for the classification of complex data sets. Therefore, to classify complex objects in VHR images, class-specific deep features should be explored at different scales.

Unlike traditional convolutional neural networks (CNN) (Krizhevsky et al., 2012), the multiscale convolutional neural network (MCNN) is proposed in this study to extract high-level spatial features at multiple scales for classifying remote sensing images. In the MCNN, image pyramid was constructed to capture spatial features across scales. Moreover, with such hierarchical property of the MCNN, high-level features can be generated layer by layer. Finally, high-level spatial features were combined with spectral features to train a logistic regression (LR) classifier for original images classification. The main contributions of this study lie in the following three aspects: (1) traditional CNN was extended to the MCNN for learning multiscale spatial features; (2) high-level and abstract spatial features were generated by the MCNN; and (3) an effective voting scheme was proposed to produce final results. The MCNN-based classification was tested on two well-known hyperspectral data sets with high spatial resolution. The results indicated that the proposed MCNN is more effective than existing methods for hyperspectral image classification. Learning new spatial features is very important to improve classification accuracy of VHR images because the available features related to shapes, spectrums and textures are not always effective for recognizing objects from VHR images, especially in urban areas. The presented learning mechanism of spatial features outperforms existing methods including sparse and hierarchical solutions (Tuia et al., 2015) and multi-index learning approach (Huang et al., 2014) as the latter can only learn features from some limited and user-predefined filters or parameters, while cannot find multiscale features. However, our method can automatically learn filters and obtain effective scales, thus can produce more robust multiscale features without prior knowledge.

The proposed spectral and multiscale spatial feature extraction is described in Section 2.1. LR classifier and voting algorithms are presented in Section 2.2. The experimental data sets are described in Section 3. In Section 4, parameters of the MCNN are analyzed and discussed. The last section shows the conclusions of this paper.

2. The proposed method

The proposed approach includes three main components: dimension reduction, feature extraction and classification. In dimension reduction step, the original data sets are projected into a low dimensional space by using principle component analysis and three PC bands are chosen (more than 95% spectral variance preserved). In feature extraction step, image pyramid is built for each PC band, and multiscale training samples are chosen for extracting spatial features across scales with the deep learning framework. In classification step, with the combination of spatial and spectral features, LR classifier is trained for image classification of each PC band. To integrate the classification results of different PC bands into the final results, an effective voting scheme is presented. Fig. 1 illustrates the flowchart of this framework.

2.1. Conventional CNNs

Typically CNNs are multilayer neural networks that can hierarchically extract deep features. Commonly, a standard CNN framework contains two kinds of layers: convolutional layer and sub-sampling layer. The convolutional layer offers filter-like function to generate convoluted feature maps, while the sub-sampling layer generalizes the convoluted features into higher levels which makes features more abstract and robust. The two kinds of layers are interspersed in the CNN, which is similar to the structure of layer-wise stacked simple and complex cells in the primary visual cortex. For a large image, receptive fields with fixed sizes (Hubel and Wiesel, 1959) (i.e., a window), are fed into the CNN for extracting features (Fig. 2).

2.1.1. Convolutional layer

At lth convolution layer, the feature maps of $(l-1)$th layer are first convolved with learnable filters $\boldsymbol{k}$, and then put through the