MULTI-SCALE 3D DEEP CONVOLUTIONAL NEURAL NETWORK FOR HYPERSPECTRAL IMAGE CLASSIFICATION

Mingyi He, Bo Li, Huahui Chen

Northwestern Polytechnical University, School of Electronics and Information
International Center for Information Acquisition & Processing, Xi'an, Shaanxi, China, 710129

ABSTRACT

Research in deep neural network (DNN) and deep learning has great progress for 1D (speech), 2D (image) and 3D (3D-object) recognition/classification problems. As HSI that with 2D spatial and 1D spectral information is quite different from 3D object image, the existing DNN cannot be directly extended to hyperspectral image (HSI) classification. A Multi-scale 3D deep convolutional neural network (M3D-DCNN) is proposed for HSI classification, which could jointly learn both 2D Multi-scale spatial feature and 1D spectral feature from HSI data in an end-to-end approach, promising to achieve better results with large-scale dataset. Although without any hand-craft features or pre/post-processing like PCA, sparse coding etc, we achieve the state-of-the-art results on the standard datasets, which shows the technical validity and advancement of our method.

***Index Terms*—** Hyperspectral image classification, 3D convolution, multi-scale, end-to-end, deep neural network

1. INTRODUCTION

Hyperspectral image (HSI) is usually composed of hundreds of spectral bands varying from visible light to shortwave, which provides rich information for target detection and classification applications. Different from the RGB CCD image or infrared image, HSI has the spatial and spectral information simultaneously in 3 dimensional data cube, resulting in great difficulties in HSI processing. Recent researches show that, for the purpose of detection and classification of targets from HSI, the contextual information could provide great advantages, leading to a growth of interest in the research of joint spatial-spectral classification approaches [[1],[2]].

Convolutional neural network (CNN) [3] can hierarchically extract more implied and deeper features due to its layer-by-layer structure. Naturally, CNN has attracted considerable attentions in HSI classification and detection [4]. A brief overview of CNN based methods for HSI classification is given below.

This work is partially supported by Natural Science Foundation of China (61420106007, 61671387) and NPU Seed Foundation of Innovation and Creation (Z2016120).

In [5], an unsupervised CNN was proposed for remote sensing image processing, in which they used greedy layer-wise strategy to train their model. Xing *et al.*[6] used stacked autoendcoder to extract features for HSI. Hu *et al.*[7] proposed a two layer 1D CNN architecture to extract spectral features and achieve significant results. Both [6] and [7] only considered the spectral information. For spatial-spectral feature extraction, Chen *et al.*[8] proposed a HSI classification model based on deep belief network (DBN), in which the input is a flattened neighbor region. Yue *et al.*[9] developed a 1D CNN framework for HSI classification and they also utilized some hand-craft features to improve the performance further. Liang *et al.*[10] utilized 2D CNN to extract the spatial-spectral features through sparse representation. He *et al.*[11] proposed a modified deep stacking network (DSN) for HSI classification, in which the coarse spectral features by band selection and course spatial features by PCA were used as inputs to the DSN. Mei *et al.*[12] used 1D CNN to extract spectral features and then fused hand-craft spatial features to improve the final performance. Du *et al.*[13] proposed a 8-layer 3D convolution network named "C3D" for RGB video classification problems, however it cannot be directly used for HSI classification as HSI is quite different from the RGB video in the correlations and resolutions among the 3 dimensional data. More recently, Chen *et al.*[14]proposed a deep 3D CNN for HSI classification.

In this paper, we are trying to design a Multi-scale 3D Deep Convolutional Neural Network, M3D-DCNN, to directly extract both the multi-scale spatial feature and the spectral feature for HSI classification. In our study, finally, a 5-layer M3D-DCNN is carried out for HSI classification, with which we achieve promising results on the standard HSI datasets including India Pines, Pavia Univ., and Salinas.

Compared with the 1D CNN, 3D convolutional kernel could slip between the spatial and spectral dimensions jointly to meet the requirements of multi-scale and multi-resolution requirements. Thus it has the power to extract more complicated spatial-spectral information in a nature and elegant way. Compared with the C3D [13] for video analysis that cannot be directly and effectively used for HSI classification, our M3D-DCNN contains smaller kernel size, which could reduce overfitting due to the HSI dataset is usually small. Compared with

the 3D-DNN [14], M3D-DCNN contains smaller kernel size and deeper layers but less parameters, reducing over-fitting in these small HSI datasets, without virtual samples, thus more closed to the practical. In addition, with the multi-scale structure, our M3D-DCNN could effectively extract HSI features, contributing to improving HSI classification accuracy.

Our main contributions in this paper can be summarized as three aspects: (1). We proposed a 3D deep CNN (3D-DCNN) approach and explored the power of 3D convolution for HSI classification and compared it with currently used CNN based methods for the problem. (2). A multi-scale 3D convolution block is proposed. With which, we proposed a multi-scale 3D deep CNN (M3D-DCNN) for HSI classification to meet the multi-scale targets in spatial domain. The experimental results show that M3D-DCNN can synchronously extract spatial & spectral features in a nature and elegant way. (3). Without any hand-craft features and pre/post-processing like PCA, sparse presentation etc, our proposed M3D-DCNN achieves the state-of-the-art result on the standard datasets. More importantly, our method is totally an end-to-end approach, promising to achieve better results with large-scale dataset in the future.

2. MULTI-SCALE 3D DEEP CONVOLUTIONAL NEURAL NETWORK

2.1. 1D, 2D, 3D Convolution For HSI Data

In hyperspectral image processing field, researchers usually use 1D CNN to extract spectral features separately in the spectral domain. When applied to HSI classification problems, it is crucial to capture joint features both in spatial dimensions and spectral dimension. Inspiring from Ji *et al.*'s work [15] on human action recognition using 3D convolution to extract the spatial-temporal features, we explore 3D kernel for jointly mining of the spatial and spectral features from HSI data.

An illustration of 1D, 2D, 3D convolution is presented in Figure 1. The formulations of 1D, 2D, 3D convolution are given below:

$$v_{ij}^{z} = f(r_{ij} + \sum_{m=0}^{M_i-1} \sum_{b=0}^{B_i-1} k_{ijm}^{b} v_{(i-1)m}^{(z+b)}) \tag{1}$$

$$v_{ij}^{xy} = f(r_{ij} + \sum_{m=0}^{M_i-1} \sum_{h=0}^{H_i-1} \sum_{w=0}^{W_i-1} k_{ijm}^{hw} v_{(i-1)m}^{(x+h)(y+w)}) \tag{2}$$

$$v_{ij}^{xyz} = f(r_{ij} + \sum_{m=0}^{M_i-1} \sum_{b=0}^{B_i-1} \sum_{h=0}^{H_i-1} \sum_{w=0}^{W_i-1} k_{ijm}^{hwb} v_{(i-1)m}^{(x+h)(y+w)(z+b)}) \tag{3}$$

where, v means the output variable in the feature map. B, H, W represent the size of kernel along spectral and spatial dimensions respectively. (b, h, w) are the indexes of kernel and

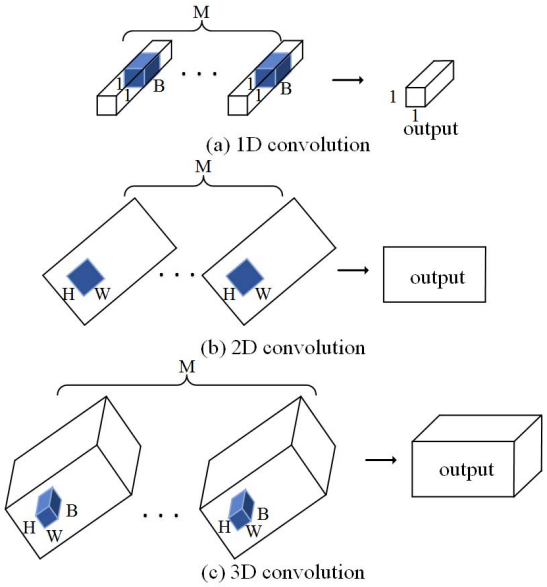

Fig. 1. Illustration of 1D, 2D, 3D convolution in HSI data. B, H, W represent the size of kernel along spectral and spatial dimension respectively. M is the number of feature maps.

(z, x, y) are the indexes of feature map, respectively corresponding to the 2 spatial and 1 spectral dimensions. k means the kernel parameters. i, j, m are the indexes of input layer, output layer and feature map respectively. M is the number of feature maps, thus M_i means the number of feature maps in the ith layer. r is the bias term. Rectified linear unit(ReLU) as the activation function is selected in this work, which is

$$f(x) = \max(0, x) \tag{4}$$

2.2. Multi-scale 3D Convolution Block

Multi-scale information has been proved for classification of related problems [16]. This is partly because multi-scale structure contains abundant context information. While it is still not well studied in the HSI classification field. In this paper, we propose a multi-scale 3D convolution block, which could be utilized as a basic structure and to construct more powerful CNN model for HSI detection and classification.

2.3. M3D-DCNN Model

With our multi-scale 3D convolution block, we construct a multi-scale 3D convolutional neural network model, which is illustrated in Figure 3. It is consisted of 10 convolution layers and 1 fully connect layer, and the depth of this network is 5. We utilize the dropout [16] layer to prevent over-fitting. And the ratio of dropout is 0.6 in our experiments. Considering the limitation of the labeled data in HSI field, the depth of our model is beyond most other CNN models [4,5,6] for HSI

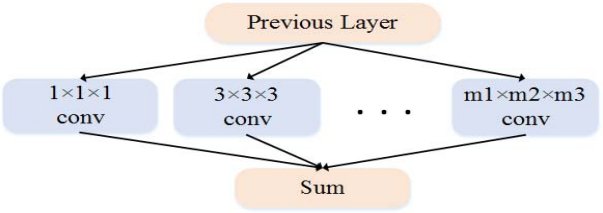

Fig. 2. Illustration of multi-scale 3D convolution block. m_1, m_2 and m_3 denote the kernel sizes in the 2 spatial and 1 spectral dimensions, respectively.

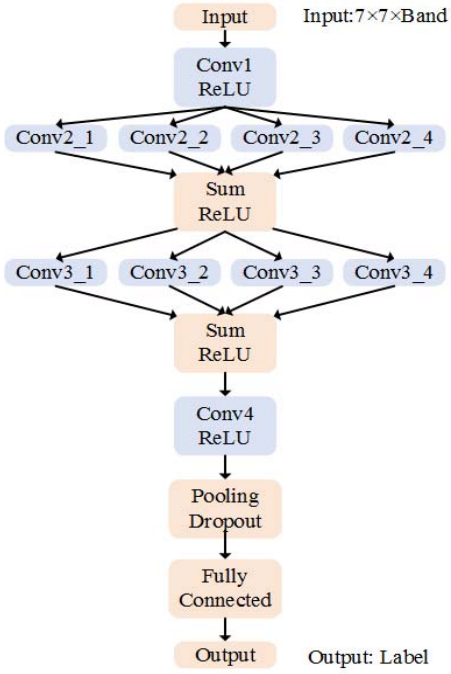

Fig. 3. Our proposed M3D-DCNN model. The size of the input patch is $7 \times 7 \times Band$. The details of other hyper-parameters are presented in Table 1.

classification. Considering the spatial resolution of the data and the target sizes for each classes to be classified, a relative small kernel size in the 2 spatial dimensions is suit for our experiments.

The detailed hyper-parameter setting of this model is presented in Table 1. The hyper-parameters are chosen for validation on the training data. In other words, we used 80% of the training samples to learn weights and the remaining 20% to choose the proper hyper-parameters. We used the same model setting for all the three datasets. In other words, we don't deliberately tune the hyper-parameters to pursue a higher performance.

We train our network with the multinomial logistic loss:

Table 1. Parameters of convolutional layers

kernel name	kernel number	kernel size H, W, B	kernel stride $\Delta(H, W, B)$
conv1	16	3,3,11	1,1,3
conv2_1	16	1,1,1	1,1,1
conv2_2		1,1,3	
conv2_3		1,1,5	
conv2_4		1,1,11	
conv3_1	16	1,1,1	1,1,1
conv3_2		1,1,3	
conv3_3		1,1,5	
conv3_4		1,1,11	
conv4	16	2,2,3	1,1,1
pooling	–	2,2,3	2,2,3

$$E = -\frac{1}{N}\sum_{n=1}^{N}\log(p_k^n) \tag{5}$$

p is the output of the softmax layer:

$$p_i = \frac{\exp x_i}{\sum_{i'=1}^{m}\exp x_{i'}} \tag{6}$$

were, N is the number of training samples, k the correspondent label of sample n, m is the number of classes. x is the input of the softmax layer.

3. EXPERIMENTS AND RESULTS

We conduct our experiments on the widely used datasets of Indian Pines, Pavia Univ. and Salina Valley. All the programs are implemented in Caffe [17] which is a widely-used deep learning framework. We train our model by AdaGrad [18] algorithm, in which the base learning rate is 0.01. In addition, we set the batch as 40, weight decay as 0.01 for all the layers.

3.1. Data Sets

The Indian Pines dataset gathered by Airborne Visible/Infrared Imaging Spectrometer (AVIRIS) sensor in North-western Indiana consists of 145×145 pixels with a ground resolution of 17 m and 220 spectral reflectance bands in the wavelength rang 0.4-2.5 μm. We reduce the number of bands to 200 by removing bands covering the region of water absorption. It includes 16 classes, and we select 8 classes due to some classes having too few labeled samples.

The University of Pavia dataset acquired by the Reflective Optics System Imaging Spectrometer (ROSIS) sensor during a flight campaign over Pavia University consists of 610×340 pixels with ground resolution of 1.3 m and 103 bands.

The Salinas dataset was collected by the 224-band AVIRIS sensor over Salinas Valley, California, comprising 512×217

pixels with a ground resolution of 3.7 m. We reduce the number of bands to 204 by removing bands covering the region of water absorption.

In all the 3 datasets, we randomly select 200 labeled pixels per class for training and the rest for testing. The input of our network is the HSI 3D patch in the size of $7 \times 7 \times Band$, where $Band$ denotes the total number of spectral bands. The size of the output is the number of the classes. For paper space limit, only the training number and test number for each lass of the Indian Pines dataset are presented in Table 2.

Table 2. Number of training and test data used in the Indian Pines dataset.

	Class Name	Training #	Test #
1	Corn-notill	200	1228
2	Corn-mintill	200	630
3	Grass-pasture	200	283
4	Hay-windrowed	200	278
5	Soybean-notill	200	772
6	Soybean-mintill	200	2255
7	Soybean-clean	200	393
8	Woods	200	1065
	Total	1600	6904

Due to the limitation of the training samples in HSI field, we augment the dataset by adding Gaussian noise in the spectral domain. At last, the augmented training data is twice as large as the original one.

3.2. Result Analysis

Firstly, we compare our M3D-DCNN method with other state-of-the-art methods like RBF-SVM [7], Hu's CNN [7], and Mei's CNN [12]. All the methods are compared under the same experiment settings like the number of training samples and patch size etc. The results are listed in Table 3. As we can see, our M3D-DCNN method has better or comparable performance than the other three methods, even the method in [12] utilized the hand-craft spatial features in their network.

For visual comparison, the experimental results with the Indian Pines dataset are drawn in Pseudo-color in Figure 4. It is obvious that our M3D-DCNN achieves the best performance.

The extensive experiments with the three public HSI datasets have proven the technology validity and advancement of our method. The results prove that (M)3D-DCNN is also an elegant way to jointly extract spatial-spectral features for HSI data.

In addition, we verify the effectiveness of our multi-scale design. To this end, we replace the multi-scale block with normal 3D convolution layer. The correspondent results are presented in Table 4. As we can see, the multi-scale design improves the performance significantly.

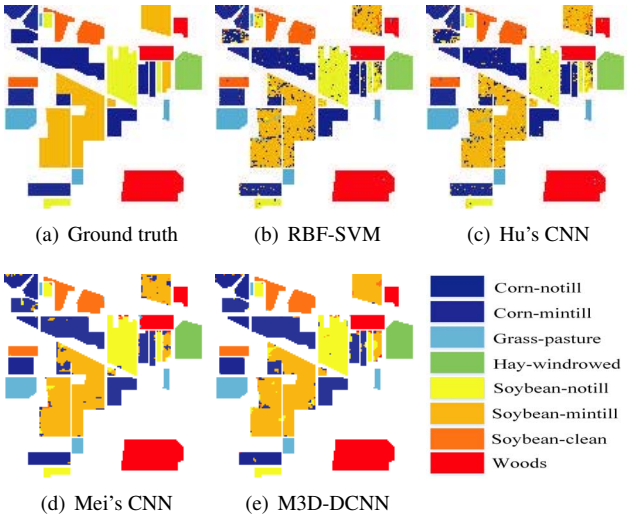

Fig. 4. Classification maps for the Indian Pines dataset. From left to right: (a) ground truth, (b) RBF-SVM[7], (c) Hu's CNN[7], (d) Mei's CNN[12], and (e) our M3D-DCNN

Table 3. Comparisons with 3 state-of-the-art methods: RBF-SVM [7], Hu's CNN [7], Mei's CNN [12] and M3D-DCNN

Dataset	RBF-SVM	Hu's CNN	Mei's CNN	M3D-DCNN
Indian Pines	87.45%	90.07%	95.70%	**97.61%**
Pavia Univ	90.59%	92.74%	98.00%	**98.49%**
Salinas	91.34%	92.52%	94.60%	**97.24%**

Table 4. Influence of with or without multi-scale

Dataset	3D-DCNN	M3D-DCNN
Indian Pines	95.45%	**97.61%**
Pavia Univ	98.04%	**98.49%**
Salinas	96.72%	**97.24%**

4. CONCLUSION

In this paper, a novel multi-scale 3-dimension deep convolutional neural network (M3D-DCNN) is proposed, which could jointly learn both 2D Multi-scale spatial feature and 1D spectral feature from HSI data in an end-to-end approach. Compared with other state-of-the-art methods, we achieved better or comparable performance in the standard datasets.

In future work, we will explore more effective data augmentation methods to overcome data limitation. Furthermore, more powerful network architecture design is also deserved the attention.

5. REFERENCES

[1] Xiuping Jia, Bor-Chen Kou, and Melba Crawford, "Feature mining for hyperspectral image classification," *Proceedings of IEEE*, vol. 101, no. 3, pp. 676–697, 2013.

[2] Mingyi He, Wenjuan Chang, and Shaohui Mei, "Advance in feature mining from hyperspectral remote sensing data," *Spacecraft Recovery & Remote Sensing*, vol. 34, no. 1, pp. 1–12, 2013.

[3] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[4] Suraj Srinivas and R. Venkatesh Babu, "Deep learning in neural networks: An overview," *Computer Science*, 2015.

[5] A. Romero, C. Gatta, and G. Camps-Valls, "Unsupervised deep feature extraction for remote sensing image classification," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 54, no. 3, pp. 1349–1362, 2015.

[6] Chen Xing, Li Ma, and Xiaoquan Yang, "Stacked denoise autoencoder based feature extraction and classification for hyperspectral images," *Journal of Sensors*, vol. 2016, pp. 1–10, 2016.

[7] Wei Hu, Yangyu Huang, Li Wei, Fan Zhang, and Hengchao Li, "Deep convolutional neural networks for hyperspectral image classification," *Journal of Sensors*, vol. 2015, no. 2, pp. 1–12, 2015.

[8] Yushi Chen, Xing Zhao, and Xiuping Jia, "Spectralc-spatial classification of hyperspectral data based on deep belief network," *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing*, vol. 8, no. 6, pp. 1–12, 2015.

[9] Jun Yue, Wenzhi Zhao, Shanjun Mao, and Hui Liu, "Spectralcspatial classification of hyperspectral images using deep convolutional neural networks," *Remote Sensing Letters*, vol. 6, no. 6, pp. 468–477, 2015.

[10] Heming Liang and Qi Li, "Hyperspectral imagery classification using sparse representations of convolutional neural network features," *Remote Sensing*, vol. 8, no. 2, 2016.

[11] Mingyi He and Xiaohui Li, "Deep stacking network with coarse features for hyperspectral image classification," in *WHISPERS'16*, Aug 2016.

[12] S. Mei, J. Ji, Q. Bi, J. Hou, and Q. Du, "Integrating spectral and spatial information into deep convolutional neural network for hyperspectral classification," in *IGARSS*, July 2016, pp. 5067–5070.

[13] Tran Du, Bourdev Lubomir, Fergus Rob, Torresani Lorenzo, and Paluri Manohar, "Learning spatiotemporal features with 3d convolutional networks," *arxiv.org/abs/1412.0767*, 2014.

[14] Yushi Chen, Hanlu Jiang, Chunyang Li, and Xiuping Jia, "Deep feature extraction and classification of hyperspectral images based on convolutional neural networks," *IEEE Transactions on Geoscience & Remote Sensing*, vol. 54, no. 10, pp. 1–20, 2016.

[15] S. Ji, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 35, no. 1, pp. 221–31, 2013.

[16] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[17] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, and Jonathan Long, "Caffe: Convolutional architecture for fast feature embedding," *Eprint Arxiv*, pp. 675–678, 2014.

[18] John Duchi, Elad Hazan, and Yoram Singer, "Adaptive subgradient methods for online learning and stochastic optimization," *Journal of Machine Learning Research*, vol. 12, no. 7, pp. 257–269, 2011.