Combining Unmixing and Deep Feature Learning for Hyperspectral Image Classification

Fahim Irfan Alam*, Jun Zhou*, Lei Tong†, Alan Wee-Chung Liew* and Yongsheng Gao‡

*School of Information and Communication Technology, Griffith University, Australia
†Faculty of Information Technology, Beijing University of Technology, China.
‡School of Engineering, Griffith University, Australia

Emails: fahim.alam@griffithuni.edu.au, {jun.zhou, a.liew, yongsheng.gao}@griffith.edu.au, lei_tong@bjut.edu.cn

Abstract—Image classification is one of the critical tasks in hyperspectral remote sensing. In recent years, significant improvement have been achieved by various classification methods. However, mixed spectral responses from different ground materials still create confusions in complex scenes. In this regard, unmixing approaches are being successfully carried out to decompose mixed pixels into a collection of spectral signatures. Considering the usefulness of these techniques, we propose to utilize the unmixing results as an input to classifiers for better classification accuracy. We propose a novel band group based structure preserving nonnegative matrix factorization (NMF) method to estimate the individual spectral responses from different materials within different ranges of wavelengths. Then we train a convolutional neural network (CNN) with the unmixing results to generate powerful features and eventually classify the data. This method is evaluated on a new dataset and compared with several state-of-the-art models, which shows the promising potential of our method.

Index Terms—Image Classification, Unmixing, Nonnegative Matrix Factorization, Convolutional Neural Network.

I. INTRODUCTION

Hyperspectral sensors produce images in contiguous light wavelength indexed bands over a given spectral range [1], which provides accurate and detailed spectral information about the underlying data. This capability has greatly benefited the identification and classification of spectrally similar materials. Along with spectral information, the spatial relationships among various spectral responses in a neighborhood can be explored, which allows development of spectral-spatial models for accurate image classification. These advantages have contributed significantly to a wide range of remote sensing applications in agriculture, mineralogy, surveillance and environmental sciences [2].

One of the critical challenges of hyperspectral image processing is the problem of mixed pixels. Usually, when the spectral resolution increases, the spatial resolution decreases. In case of high altitude sensors covering wide areas, low spatial resolution is a common problem [3]. These limitations significantly affect the performance of methods used to analyse and process hyperspectral data. In particular, classification tasks suffer greatly due to the problem of mixed pixels in which case a pixel may contain more than one material/class. The combination of mixed and pure pixels also happen in high spatial resolution images [4].

The notion of a pure material can be application dependent. Suppose a hyperspectral image contains materials such as bricks, roads, water, plants, soil, cement, which gives a general assumption of the presence of six classes. However, if the percentage of pixels covered by cement is comparatively too small, then we may not want to define an independent class for cement. It also depends on if we are indeed required to estimate the proportion of cement. Similarly, if the application demands us to distinguish between two types of plants, then we must create two plant classes based on their spectral signatures. Fortunately, by applying unmixing process, pixels in the spectra can be decomposed into a collection of spectral signatures, called endmembers, directly from data without much prior knowledge about the presented classes. These endmembers are represented as a set of fractional estimations called abundances. Among different unmixing techniques, methods based on linear mixing model considers each pixel as a linear combination of endmembers.

Linear mixture models can be further divided into three categories: geometry, sparse regression and statistics based. Geometry based methods [5] exploit the geometric relationships among the endmembers and estimate their abundances. Sparsity constraint is often applied to sparse regression methods [6] to select few endmembers with high variance of material reflectance. Spatial regularization term can also be included in the spectral domain using prior information about the endmembers to unmix the data [7]. Statistical unmixing methods such as nonnegative matrix factorization (NMF) [8] decompose the image into nonnegative endmember and abundance matrices. Various constraints such as sparsity [9], combination of spectral and spatial constraints [10] for smoothing of endmembers or maintaining manifold structure of unmixed data [11] have been applied.

Along with pixel based unmixing methods, region based methods [12] have also been developed. It is quite interesting to explore the relationships among the regions in a hyperspectral image because there is a high possibility that the same endmembers may appear in a local neighbourhood or in other homogeneous regions. Therefore, it is important to both consider the consistency of abundance of those endmembers in a homogeneous region and discriminate the contributions of those endmembers from other regions across the image.
To address this property, Tong et al. [13] proposed a novel region based NMF (R-NMF) method that forces consistent abundances within each homogeneous region and also separate the contribution from endmembers among the regions.

Using unmixing results to improve classification performance has also been investigated in the literature where unmixing has been used as a dimensionality reduction process and later classify images [14]. Villa et al. [15] proposed a semi-supervised method which uses linear spectral unmixing method to label data samples for classification. In a supervised classification method [16], probabilistic SVM was used to get preliminary classification map so that mixed pixels can be identified. Then a spectral unmixing method based on fully constrained least squares (FCLS) method was adopted to solve sub-pixel mixing problem in the final classification map. Similar approach was also discussed in [17] where mixture-tuned matched filtering (MTMF) method was used to get the abundance map which is used as the input to classification step. All these approaches treated the whole spectral information together without investigating further into the contribution from specific wavelength ranges or band groups in the unmixing step. Therefore, abundance information from different range of wavelengths have not been sufficiently explored and utilized to benefit the classification step.

Recent advances in training multi-layer neural networks have contributed much in a wide variety of machine learning problems including classification or regression tasks. The “deep” architecture such as convolutional neural network (CNN) can extract more abstract and invariant features of data, and thus have the ability to produce higher classification accuracy than the traditional classifiers [18]. Although most of the CNN based classification methods have demonstrated improved performance by automatically discovering underlying features of the data, it would be interesting to investigate the potential of providing additional information to the network to better characterize the data.

In this regard, we propose to utilize unmixing results in the form of abundance matrices as the input to the CNN for hyperspectral image classification. We use a region based NMF method for structure consistency and preservation during the unmixing step [13]. We argue that the derived abundance matrices contain significant information about the underlying endmembers in the regions for classification. Furthermore, feature dimension can be reduced by mapping number of bands to number of endmembers. Also, instead of providing only one instance of abundance matrix for each endmember, we intent to provide multiple instances of abundance matrices which will be obtained from different band groups along the spectral channel. Such representation is very useful for deep learning framework as the model will be able to receive sufficient information to identify interesting patterns within the data. In this way, the problem of limited training samples of hyperspectral data can be resolved to a significant extent which can facilitate the powerful classification ability of deep learning models. To the best of our knowledge, this is the first work that combines abundance maps across different wavelength groups as an unmixing output with deep feature learning for classifying hyperspectral images. Our framework is illustrated in Fig. 1.

II. UNMIXING FOR HYPERSPECTRAL IMAGE CLASSIFICATION

In this section, we briefly explain the unmixing procedure presented in [13] that is used in our method to generate abundance matrices for each endmember. We extended this approach by considering abundance information from different groups of wavelengths across the spectral channels to capture distinct spectral-spatial estimations of each endmember in the image. Later, we illustrate how we provide the unmixing results as the inputs to the CNN for generating features to classify hyperspectral images.

A. Region Based NMF for Structure Consistency and Preservation

At first, a hyperspectral image \( I \) is segmented into a set of \( N \) homogeneous regions \( R = \{R_1, R_2, \ldots, R_N\} \) using a graph-based method [19]. In this method, each pixel in \( I \) is represented as a vertex \( v_i \in V, i = 1, 2, \ldots, P \) and neighbouring pixels \( v_i, v_j \) are connected by an edge whose weight represents the distance between the connecting pixels. This algorithm merges the pixels to construct homogeneous regions by using two important criteria: maximum internal difference \( D_1 \) of a region and minimum connecting weight \( D_2 \) between regions. \( D_1 \) is the largest weight in a minimum spanning tree \( T \) [20] of a homogeneous region \( R \) defined for all \( v_i, v_j \in R, (v_i, v_j) \in T \) as:

\[
D_1 = W_{\text{max}}(v_i, v_j) + \frac{\alpha}{R} \tag{1}
\]

\( \alpha \) is a parameter that controls the contributions from small regions. \( D_2 \) between regions \( R_1 \) and \( R_2 \) is defined for all \( v_i \in R_1, v_j \in R_2, (v_i, v_j) \in T \) as:

\[
D_2 = W_{\text{min}}(v_i, v_j) \tag{2}
\]

If for two homogeneous regions \( D_1 > D_2 \), these two regions are merged. Otherwise, they are not merged.

After we obtain the homogeneous regions \( R \), the mean abundance spectral response \( t_n \) is estimated from the mean values of spectral responses for each region \( R_n \). Within homogeneous region, the spectral responses for each pixel should be similar and as a result, the abundance of every pixel should be similar as well. A constraint between mean abundance \( t_n \) and estimated abundance vector \( a_i \in A \) for each pixel of the region is set to apply the structure consistency in the region. With this constraint, the objective function to minimise for region \( R_n \) is defined as [13]:

\[
F(M_n, A_n) = \frac{1}{2}\|I_n - M_n A_n\|_F^2 + \lambda\|A_n\|_1 + \mu \sum_{a_i \in R_n} \|a_i - t_n\|_2^2 \tag{3}
\]

where \( I_n \) contains the raw spectral responses of pixels in the region, \( M_n \) and \( A_n \) are the estimated endmember matrix and
abundance matrix, respectively. The second term controls the sparsity of abundance in \( R_a \) and \( \lambda \) controls the contribution of this sparsity. The third term is responsible for controlling the consistency of the estimated abundance in \( R_a \) in which \( \mu \) controls the contribution from structure consistency.

Next, structure preserving constraint is applied to discover the relationship between the homogeneous regions to separate their contribution across the image. This constraint helps in preserving the local affinity of data distribution both before and after the matrix factorization. Also, it avoids the effect of distant repulsion [21] which is a distortion done by the distant data points. The distant repulsion causes different materials to contain different abundance while local affinity ensures that the same material in different regions shall have similar abundance. Graph regularization is applied later to preserve the structural information where the vertices represent the reflectance at different data points. Details of this approach and the optimisation process can be found in [13].

### B. Band Group based Abundance Estimation

In this paper, we estimate the abundance of each endmember for groups of bands spanning over the entire spectral channels. Therefore, instead of calculating abundance over the entire spectra at once as in [13], we calculate abundances for a number of band groups with the goal to capture subtle spectral-spatial contribution from the image. In this way, each band group leads to distinct estimation of endmembers and their abundance. Such band grouping strategy has several advantages. First, smaller band groups provide better spectral-spatial estimation locally. Second, different band groups capture spectral information in different ranges of wavelengths which can contribute in classification performance by providing more material based information to the classifier. Finally, a large number of abundance maps representing the endmembers of the image are expected to benefit the CNN for classification in terms of both sufficient amount of training samples and useful spectral-spatial information as input to start the training with.

We group the original \( L \) bands into segments of \( B \) bands where \( B \ll L \). Therefore, endmember and abundance matrices can be estimated from each band group in a multi-task manner by the following linear mixture models as:

\[
I_1 = M_1 A_1 + E_1 \\
I_2 = M_2 A_2 + E_2 \\
\vdots \\
I_G = M_G A_G + E_G
\]

where \( G \) is the total number of band groups. \( M_g \) and \( A_g \) are the group-wise endmembers and abundances. \( E_g \) is the corresponding additive noise.

For each band group, the unmixing step follows the method in [13] with all parameters remaining the same across band groups. The sparsity constraint \( \lambda \), however, is band group dependent and is defined as:

\[
\lambda_g = \frac{1}{\sqrt{B}} \sum_{b=1}^{B} \frac{\sqrt{P - ||I_b||_1/||I_b||_2}}{\sqrt{P-1}}
\]

where \( I_b \) is the spectral responses in the \( b \)-th band and \( P \) is the total number of pixels in the image.

Furthermore, the local affinity and distance repulsion calculation is also undertaken within each band group. Finally, we obtain \( T \times G \) abundance maps where \( T \) is the number of endmembers in each band group.

### C. Feature Learning and Classification Using CNN

In this section, we explain the process of using the abundance matrices as the input to the CNN for generating features and later classifying hyperspectral images. Deep learning based methods build a deeper hierarchical model, usually deeper than three layers. More abstract and invariant features can therefore be learned by using these models, which is very beneficial for classification tasks. Recently, CNNs have achieved tremendous success in classification of hyperspectral images [22]. The architecture of CNN is different from other deep learning models in two ways. Firstly, CNN uses the local connection between the neurons of adjacent layers. Secondly, some of these connections are replicated across the layer by sharing the same weights. With these two properties,
CNN produces better performance in classification tasks than traditional classifiers that take hand-crafted features as input.

In our method, instead of using raw image data, we treat the abundance maps of the endmembers as input to the CNN. These abundance matrices contain the estimation of contribution that each endmember has in each band group. Therefore, our CNN implicitly receives spectral information across the entire spectral bands through these abundance matrices in spite of using convolution operations across the spatial domain only. For $T$ classes and $G$ band groups, we will have $T \times G$ abundance matrices to be provided as the input to the CNN.

Our CNN includes several convolution layers, pooling layers, fully connected layers and activation functions. During the CNN training, repeated convolution operations between kernels and the abundance matrices produce multiple feature maps. Let $(x, y)$ defines a location in the spatial dimension whose value on a feature map is given by [23]:

$$val_{ij}^{xy} = f(\sum_{i=1}^{m} \sum_{c=0}^{G_{\text{c}}} \sum_{g=0}^{G_{\text{g}}} k_{ij}^{cg} \cdot val_{(i-1)j}^{(x+c)(y+g)} + \delta_{ij})$$

where $l$ is the current layer that is being considered, $m$ is the number of feature maps in the $(l - 1)$-th layer (previous layer), $j$ is the current kernel number, $i$ is the current feature map in the $(l - 1)$-th layer connected to the feature map of the $l$-th layer, $k_{ij}^{cg}$ is the $(c, g)$-th value of the kernel connected to the $i$-th feature map in the previous layer and $\delta$ is the bias of the network. $C_l$ and $G_l$ are the height and the width of the kernel respectively.

Each feature map is treated independently. Therefore, $val_{ij}^{xy}$ is calculated by convolving a feature map of the previous layer with a kernel of the current layer. In this process, the number of feature maps in the previous layer will be multiplied by the number of kernels in the current layer which will produce as many feature maps as the output of the $l$-th convolution layer.

After obtaining the convolved features, we use pooling to reduce the size of the feature maps. This operates independently on a small patch of the feature map and resizes it spatially, mostly by using MAX operation. The MAX operation takes the maximum value over the patch and discards the rest of the activation. The use of pooling brings invariance property by reducing the amount of parameters. We also used rectified linear unit (ReLU) as the activation function which is faster than other saturating activation functions. After several layers of convolution, pooling and ReLU layers, the intermediate feature maps are trained. The feature maps are then flattened into a feature vector and fed into a fully-connected layer which extracts the final learned deep features. Finally, we use logistic regression (LR) as a classifier to generate the required classification results.

During the CNN training, all the connections/weights are being updated by using the gradient descent back propagation algorithm. We randomly initialize the model parameters. A cost function is required to update the weights during the training. In our training process, we used mini-batch update procedure which is computed on a mini-batch of inputs [24]:

$$d = -\frac{1}{s} [y_i \log(z_i) + (1 - y_i) \log(1 - z_i)]$$

where $s$ is the mini-batch size, $y_i$ and $z_i$ are the i-th predicted label and label in the mini-batch respectively. We aim to optimize cost $d$ using stochastic gradient descent.

LR uses soft-max function to classify the learned features from the CNN. By using soft-max, LR triggers the output units to 1 in order to represent the results as a set of conditional probabilities. For the given input $X$, the probability that the input belongs to class $i$ over a total of $K$ classes is estimated as:

$$P(y_i|X, W) = \frac{e^{W_{it}^l}}{\sum_{k} e^{W_{kt}^l}}$$

where $W$ are the weights of the LR layer.

In this paper, we proposed to train our CNN with local spatial-spectral information through the use of abundance matrices so as to let different wavelength ranges make distinct contribution to the final classification outcome. Because the convolution kernels are applied to each of these different abundance matrices, there is a high possibility that the intermediate feature maps will become more meaningful and interesting throughout the training. We present the layer-wise intermediate feature map results in the next section.

III. EXPERIMENTAL RESULTS

In this section, we present the experimental results on real-world hyperspectral remote sensing images. Then we analyse the performance of the proposed method in comparison with several alternatives.

A. Dataset

For better evaluation of our proposed method, we created a dataset “Griffith-USGS” by collecting AVIRIS images from the USGS database\footnote{https://earthexplorer.usgs.gov/}. After downloading data located in the region of north America spanning over the United States of America, Canada and Mexico, we performed a subsequent pre-processing step to make the images compatible for use in hyperspectral image analysis. In our research, we downloaded 19 scenes to build the training and testing sets for deep learning. We cropped those scenes into a number of individual portions to build 150 training images and 100 testing images. As we captured scenes from multiple locations, the spatial resolutions of the scenes used in this dataset range from 2.4 to 18 meters per pixels. Each image consists of approximately $224 \times 224$ pixels.

The data that we obtained from AVIRIS were not labeled. To fit into our proposed supervised training framework, we performed a pixelwise manual labeling on the images. We created a training set containing six classes, including road, water, building, grass, tree and soil. The labeling was done with the help of high-resolution color images in Google Earth.
B. Design of the CNNs

For the CNN used in our method, we used $5 \times 5$ convolution kernels. We adopted four convolution layers and three pooling layers with $2 \times 2$ pooling kernel in each layer. ReLU layers were used as well and cut off the features that were less than 0. The pooling layers reduced the size of the feature maps. The size of the mini-batch was set to 10. For the logistic regression, the learning rate was set to 0.005 and the number of epochs was 550. The weights were randomly initialized and gradually trained using the back propagation algorithm. Each convolution kernel extracts distinct features from the abundance matrices as input and with the weights being learned. The features convey meaningful structural information about the data. Different kernels used in the convolution layers are able to extract different features on the way to form a powerful representation.

C. Results and Comparison

We attempted to use unmixing as an important feature extraction approach to be provided abundance maps as the input for classification of hyperspectral images. That is why we evaluated the effectiveness of including unmixing results not only for our proposed CNN model but also for other classifier as well. In this regard, we used a widely used classifier support vector machine (SVM) to observe the rate of improvement these classifiers achieve after incorporating unmixing results in the form of abundance matrices for the defined endmember of our hyperspectral images.

As mentioned in the previous section, we generated multiple abundance matrices for each spectral signature from 20 groups of bands across the spectral channel. Fig. 2 shows an example of measuring an endmember “water” with its estimated abundances as realized by the brightness of the pixels covering the region, though most of them are quite consistent with each other.

Next, we evaluated the performance of SVM for using unmixing output as an input to the model. At first, we simply executed a multi-class SVM by giving raw spectral information for the pixels. After this, we replaced the raw spectral information with the band group based abundance matrices computed earlier to train the model. Later, we made a comparison between these two different input information to test the usefulness of unmixing output. For performance evaluation, we calculated the overall accuracy (OA) and average accuracy (AA) with the corresponding standard deviations. We repeated our experiments for ten times over the randomly split training and testing data. From Table I, we see that the classification accuracy of SVM improved quite significantly after including abundance information for each endmember as input. Also, to justify the usefulness of generating abundance matrices for smaller band groups instead of whole spectral channel at once, we executed our method in two settings: (1) abundance matrices generated from the whole spectral channels and (2) band group based abundance matrices. Table I shows that band groups lead to superior performance. With different wavelength ranges, the abundance information captured better local information distinct to each group and hence, the variety of spectral-spatial information had been estimated for each underlying endmember. Thanks to these advantages, SVM was able to classify the pixels better than its counterpart where abundance information were collected from the whole spectral channel at once.

To further evaluate the usefulness of including band group-wise abundance estimations, we made comparisons with methods mentioned in our literature where abundance maps were generated from whole spectral channel and used as input to SVM for classification. First, we compared with [16] and to make a fair comparison, we experimented on a widely used AVIRIS dataset “Indian Pines” containing sixteen classes. As experimented in their paper with this dataset, we also maintained the same experimental settings and randomly selected 15% samples for training. We also implemented probabilistic SVM included in the LIBSVM [25]. For convenience, we call this method "FCLS-SVM". Later, we also compared with [17]
TABLE II
Comparison with FCLS-SVM (Indian Pines)

<table>
<thead>
<tr>
<th></th>
<th>FCLS-SVM [16]</th>
<th>Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>OA(%)</td>
<td>91.10±1.42</td>
<td>97.71±1.96</td>
</tr>
<tr>
<td>AA(%)</td>
<td>90.73±1.73</td>
<td>95.90±1.48</td>
</tr>
</tbody>
</table>

TABLE III
Comparison of Overall Accuracy (in %) with MTMF-SVM (PV)

<table>
<thead>
<tr>
<th>Training Samples</th>
<th>MTMF-SVM [17]</th>
<th>Proposed Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>5%</td>
<td>93.38</td>
<td>95.63</td>
</tr>
<tr>
<td>10%</td>
<td>94.40</td>
<td>97.55</td>
</tr>
<tr>
<td>15%</td>
<td>94.87</td>
<td>97.71</td>
</tr>
</tbody>
</table>

on another AVIRIS dataset “Pavia University (PV)” containing nine classes. To make fair comparisons, we randomly selected 5%, 10%, 15% samples for training and used an SVM classifier with a Gaussian kernel. We call this method “MTMF-SVM”.

Table II and Table III show the classification comparisons and we see that our proposed method of band groupings for abundance estimation produces better classification accuracy and hence shows the significant potential of extracting distinct groupwise abundance information for the endmembers.

In our method, we proposed to use CNN as an integral model to build powerful features representing our data. Unlike most of the CNN based methods which learn features automatically from raw data, we provided abundance matrices to the model for better feature construction. To test the usefulness of doing this, we compared our CNN based model with other methods as well. Under the same experimental settings, we tested an SVM-based classification algorithm [26] which itself is divided into two parts: (1) SVM with composite kernel (SVM-CK) and (2) SVM with generalized composite kernel (SVM-GCK). We compared our results with SVM-GCK as it outperformed its counterpart SVM-CK.

We also compared our method with a spatial-spectral-based method (MPM-LBP-AL) [27]. In this method, active learning (AL) and loopy belief propagation (LBP) algorithms were used to learn spectral and spatial information simultaneously. Then the marginal probability distribution were exploited, which used the whole information in the hyperspectral data. We made comparison with another supervised method (MLR_subMLL) [28] that integrated spectral and spatial information into a Bayesian framework. We presented our results in Table IV. It shows that our proposed method achieves much better classification accuracy compared to other methods.

We can draw two significant conclusions from these results. Firstly, the use of abundance matrices provide useful information as input for the classifier. Secondly, CNN is able to learn better features from those abundance matrices as it achieves significantly better classification accuracy than SVM which also considers abundance matrices under the same experimental settings.

Fig. 3 shows the intermediate features generated during
the CNN training. We can observe that the four convolutions layers used in our model gradually constructed more structured representation of the data.

Fig. 4 illustrates the CNN based classification results on our dataset. The first column is the ground truth. The second column is the classification map generated by the CNN. The third column is a binary map that shows the effect of corresponding misclassification obtained by comparing with the ground truth. The white pixels indicate the parts of the image that were not correctly classified. We further analysed that maximum amount of misclassification happened in cases of classifying “road” and “soil”. Comparatively greater spectral similarities shared by these two different materials might possibly have led to such misclassification. In future, we plan to combine more distinctive features in the unmixing and classification approaches to solve such complicated cases.

IV. CONCLUSION

In this paper, we presented a CNN based classification model by incorporating unmixing results during the training procedure of the model. We extended an existing region based structure preserving nonnegative matrix factorization method to estimate the individual spectral responses from different materials in different groups of wavelengths. The estimated abundance maps of the materials are used as important features to train a deep CNN model. Instead of learning from raw data, our proposed CNN model receives significant spectral-spatial information in the data to produce better and powerful features so as to achieve improved classification performance. Comparison with several state-of-art methods shows the potential of using unmixing in deep learning based classification framework.

REFERENCES


