Turkish Journal of Electrical Engineering & Computer Sciences

http://journals.tubitak.gov.tr/elektrik/

Research Article

Turk J Elec Eng & Comp Sci
(2019) 27: 1696 – 1703

doi:10.3906/elk-1812-116

Can additional spectral bands be estimated from aerial color images?

Muhammet Ali DEDE*, **Erchan APTOULA**, **Yakup GENÇ**
Department of Computer Engineering, Faculty of Engineering, Gebze Technical University, Kocaeli, Turkey

Received: 17.12.2018 • **Accepted/Published Online:** 22.02.2019 • **Final Version:** 15.05.2019

Abstract: Inspired by the surprising performances of deep generative models, in this paper we present the preliminary results of an overly ambitious task: estimating computationally the additional spectral bands of a color aerial image. We have harnessed the expressive power of deep generative models to estimate the distribution of mostly infrared bands of aerial scenes, using only color RGB channels as input. Our approach has been tested from multiple aspects, including the reconstruction error of the additional bands and the effect of estimated bands on scene classification performance, as well as through the transfer potential of the trained network to a distinct dataset. To our surprise, the initial experiments have shown us that deep generative models can indeed learn to estimate additional bands up to a certain degree and can thus computationally reinforce datasets stemming from color-only sensors.

Key words: Aerial scene classification, auto-encoder, generative models, convolutional neural network, spectral super-resolution

1. Introduction

Thanks to advances in sensor technology, the spectral resolutions of remote sensing images have increased to unprecedented levels, paving the way for new applications and improving the performances of existing ones. High spectral resolution is known to improve the performance of a wide range of critical remote sensing applications, ranging from target detection [1] and scene classification [2] all the way to pixel classification [3–7], through the availability of complementary information.

However, even though such multi- and hyperspectral image acquisition devices are nowadays more accessible than even before, they are still evidently not as widespread as RGB color sensors. Consequently, we have chosen to investigate whether, given a RGB color scene, one can estimate computationally its appearance in subsequent wavelengths. Obviously, one cannot expect to estimate/guess accurately the full spectral response of an unknown material based only on its color. However, even a crude approximation can have enormous practical value when all one possesses is color. In fact, doing so with entire aerial scenes where spatial information is abundant and the surface material types are of a relatively limited variability might not necessarily be implausible. If it were to be even partially possible, it could lead the way for spectral super-resolution.

Our motivation for this investigation lies in improving aerial scene classification performance. Even though the number of datasets continues to increase [8], their majority remains color-only (UCM, AID, NWPU-RESISC45, etc.), and consequently contemporary aerial scene classification methods are mostly mere adaptations of color image analysis approaches. For example, Liu et al. [2] trained in parallel two convolutional architectures that are fused with a common loss function and showed the interest of using lower layer features

*Correspondence: madede@gtu.edu.tr

as opposed to exploiting the fully connected layers' output. Yu and Liu [4], on the other hand, focused on multilevel fusion and trained 3 convolutional neural networks (CNNs) concurrently, each with distinct resolutions of the same input image, that were then fused together to propose a single output. In addition, Weng et al. [3] presented the first use of extreme learning machines in the context of aerial scene classification by using it to replace the fully connected layer classifier. For a comprehensive recent survey on aerial scene classification the reader is referred to [8].

In this work we answer the question of whether it is possible to estimate additional EM spectra from RGB images with the help of other multispectral image data. Our inspiration/optimism in undertaking this task stems from the paradigm shift caused by deep neural networks in most visual data analysis fields, especially lately through successful generative models, such as generative adversarial networks [9] and variational autoencoders [10], known in particular for their capacity for approximating latent distributions. That is why, in this paper, we have explored harnessing the expressive power of deep encoder-decoder models through a custom network architecture and have used it with the aim of estimating the distribution of mostly infrared bands of aerial scenes, using as input only color RGB channels. We further propose an iterative estimation strategy as well. This approach may seem counterintuitive, since EM emissions and scattering of a nonblack body can not be estimated via numerical and algorithmic methods just by looking at the visual bandwidths. However, we trust that the proposed model does not only inspect the colors of the given image but also scene content and the relation between objects. Another drawback of the proposed method is the need for immense amounts of training data. Generative models require substantially more training data in order to capture the latent distribution of scene content. Even if a correlation exists between scene RGB content and the other spectra of the given images, this will be scene-dependent. To overcome this shortcoming the proposed method requires much training data with highly diversified content.

In order to prove our claim, the proposed approach has been tested from multiple aspects, including the reconstruction error of the additional bands, qualitative evaluation of estimated bands, and effect of estimated bands on scene classification performance, as well as through the transfer potential of the trained network to a distinct dataset. To our surprise, the preliminary experiments have shown us that deep generative models can indeed learn to estimate additional bands up to a certain degree and can thus computationally reinforce datasets stemming from color-only sensors.

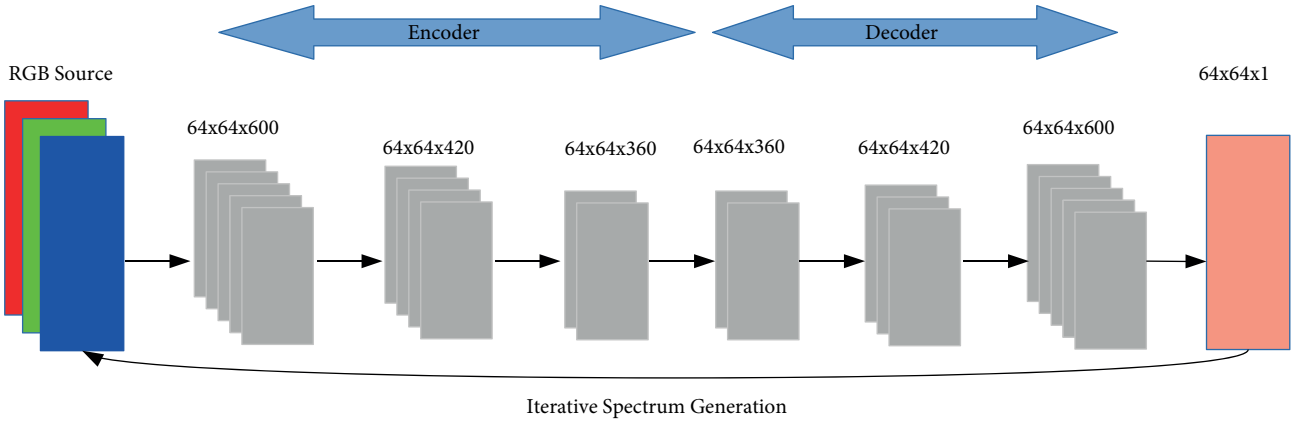

Figure 1. Illustration of the proposed encoder-decoder architecture for spectral band generation.

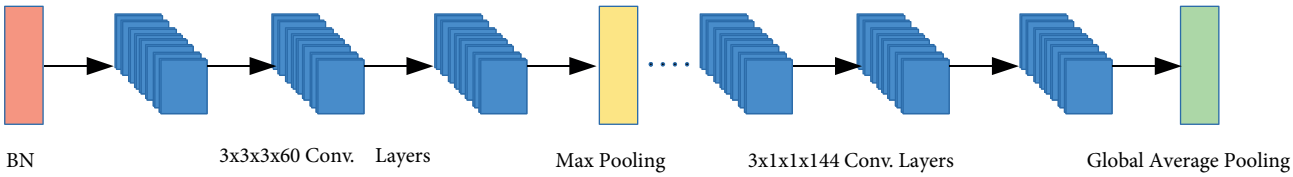

Figure 2. The convolutional blocks are composed of batch normalization, 3 convolutional layers, and a final max pooling layer. These blocks are repeated 4 times with 60, 72, 84, and 96 filters, respectively, and finally fed to three 1 × 1 convolution layers.

2. Explored approach

This section details the developed method for spectral band estimation.

2.1. Background on generative models

Although there are various definitions and mathematical models for generative models, in essence they assume that some observed variable x that is mapped to a label Y by a discriminative method is produced by a hidden process or variable z with some unknown distribution $p(x, z)$. Evidently, calculating this distribution is challenging. A solution to this problem comes from a simple but ingenious idea. Let q be a distribution that produces some x', which can be labeled Y by the same discriminative method. This makes one's job relatively easier by turning the problem from estimating an unknown distribution to generating x' similar to x so that our discriminative method cannot understand the difference. In other words, we just need to make x and x' similar using a distance or similarity function:

$$dist(x, x') = 0. \tag{1}$$

There have been further advances in generative models that employ more sophisticated assumptions about the a priori distribution, the choice of similarity function, or even the shape of the underlying manifold [10].

More specifically, most of the contemporary generative models stem from two major neural network disciplines. The first is generative adversarial models that work through the interplay between two semiseparate networks: a generator and a discriminator. The goal of the discriminator is to tell the difference between the data generated by the generator and the real-world data we are trying to model [9], while the others are auto-encoders, a fully unsupervised tool. They rely on standard backpropagation and set the target output values of the network to be equal to their inputs; in other words, they learn an approximation of the identity function, so as to output $\hat{x}$ that is similar to x. In doing so, and through constraints on the network, such as limited hidden units, auto-encoders can perform dimension reduction [11] as well as feature extraction [12].

Advances in recent years have evolved auto-encoders into more sophisticated variational forms [10]. These networks are rooted in the pure mathematical realm of variational Bayes theory and attempt to approximate a latent space embedding distribution with the help of a prior distribution while trying to reconstruct the given data from estimated latent parameters.

2.2. Explored estimation strategy

Our approach for estimating spectral bands from RGB images consists of two components.

For the task of band generation we have chosen to use a CNN-based encoder-decoder network similar to a convolutional auto-encoder that can receive as input arbitrary image bands (e.g., RGB color bands) and

aims to construct the band that was presented to it as ground truth (e.g., a near infrared band) (Figure 1). It possesses 3 encoding and 3 decoding convolutional layers, with 600, 420, and 360 filters, respectively, and is equipped with a mean squared error loss function.

Since, however, reconstruction loss is not a convincing metric for the success of our strategy, we also used a CNN (i.e. the discriminative model) to confirm whether our generative approach generates output similar to the original bands of the image.

In particular, our discriminative model is a traditional CNN, bearing similarities to the well-known VGG architecture. Our model is built with 4 special convolutional blocks (Figure 2), where every block consists of one batch-normalization, 3 convolutional, and 1 max pooling layers. At the end of the final block, we employed three 1×1 convolutional layers and global average pooling instead of a more classical dense layer and a final flattening layer. This network's loss function is cross-entropy.

We explored two distinct approaches for estimating the spectral bands of an aerial scene. In the first direct strategy, we have used only RGB color bands as a priori distributions and have tried to estimate each spectral band under the guidance of the original spectral bands of the input image. In the second iterative strategy, we have taken the best generated band with respect to the reconstruction loss and have combined it with the original RGB image, and used them together in estimating the next band. After each training, the generated band has been once again combined with the input and used to reestimate another band.

The next section will elaborate on the effect of generated spectra on the performance of aerial scene classification.

2.3. Datasets

We have used the recently presented Eurosat [13] dataset for training our network. It contains 27,000 64×64 pixel, 13-band images labeled into 10 categories. Band descriptions and resolutions are provided in Table 1. We have selected 2000 images for training and 500 for validation, and the rest have been kept for testing purposes. The reason for this uneven split is the number of reported saturated performances when using even (50-50 or 80-20) train/test splits.

3. Experiments and discussion

In the later stage of our experiments, where we have explored the possibility of transferring the trained network's band estimation skill to another dataset, we have used the UC Merced [14] (UCM) and AID [15] datasets. UCM data comprise 2100 color scenes of 21 categories, each with 100 samples, at a spatial resolution of 0.3 m. Original images of the UCM dataset are directly downloaded from United States Geological Survey (USGS) databases and resized into 256×256 pixel images. This dataset contains highly overlapping land use cases like dense residential areas, medium residential areas, and sparse residential areas. This unique property makes UCM a challenging dataset for discriminative tasks, but from the perspective of generative tasks, our proposed method can learn the differences in NIR bands between many visually similar scenes. The AID dataset contains 10,000 RGB color images of size 600×600 pixels, with 30 classes. The AID dataset is directly sampled from Google Earth images. Unlike UCM images, the AID dataset is multiresolution. The spatial resolution ranges from about 8 m to about 0.5 m.

3.1. Setup

Our experiments consist of two parts. In the first part (Part A), the networks presented in the previous section are trained and tested with the Eurosat dataset. More specifically, we started by conducting scene classification

Table 1. Properties of the Eurosat dataset [13].

Band	Spatial resolution (m)	Central wavelength (nm)
Aerosols	60	443
Blue	10	490
Red	10	560
Green	10	665
Red Edge 1	20	705
Red Edge 2	20	740
Red Edge 3	20	783
NIR	10	842
Red Edge 4	20	865
Water Vapour	60	945
Cirrus	60	1375
SWIR 1	20	1610
SWIR 2	20	2190

experiments with the actual data (*actual*) in order to form a baseline where no generated data are employed. Then we repeated the experiments with directly generated bands from RGB input (*direct*), as well as using the iterative generation method (*iterative*). The results of this batch of experiments are shown in Table 2.

In the second part of our experiments (Part B), we have used the network trained with Eurosat data in order to estimate spectral bands of another distinct dataset (UCM) acquired from a different sensor, over different geographical regions. UCM is a purely color dataset; hence, even though we cannot validate the accuracy of the computed estimations, we can all the same measure their effect on scene classification (Table 3).

Furthermore, a core question is which bands to estimate. Two approaches have been explored in this regard. In the first, the decision was intuitively made depending on which band minimized the reconstruction loss in the noniterative experiments. This, however, did not lead to satisfactory validation losses (Figure 3). Consequently, we adopted a different second approach, where the bands between RGB and the band with the least reconstruction loss were selected. This led to a significant drop in validation loss (Figure 4). This phenomenon is presumed to be related to the strong affinity of neural networks to interpolate between given boundaries.

3.2. Training

We employed aggressive data augmentation in training both our generative and discriminative models, through flipping, mirroring, random translations around 10 pixels, and random rotations between –20 and 20 degrees. Both of our models operate with batches of 16 images and no other regularization other than batch normalization has been employed.

Training started with a very high learning rate of 0.05 and lasted for 5 epochs. If the training loss descends below a certain threshold then it is further dropped to 0.01 for 15 more epochs and training continues for 60 more epochs with the learning rate being halved every 10 epochs. If the loss value does not fall rapidly enough, training is stopped completely for that particular band. Please note that in Part B of our experiments, no hyperparameter search has been conducted for the UCM dataset; instead, in order to simulate a practical scenario, all model-

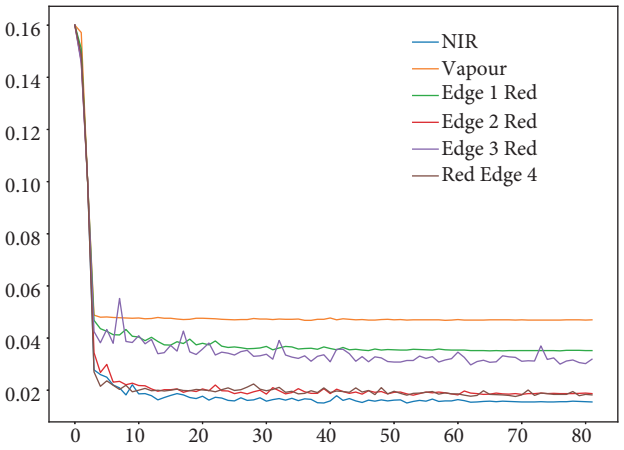

Figure 3. Validation loss of the direct generative process w.r.t. epochs.

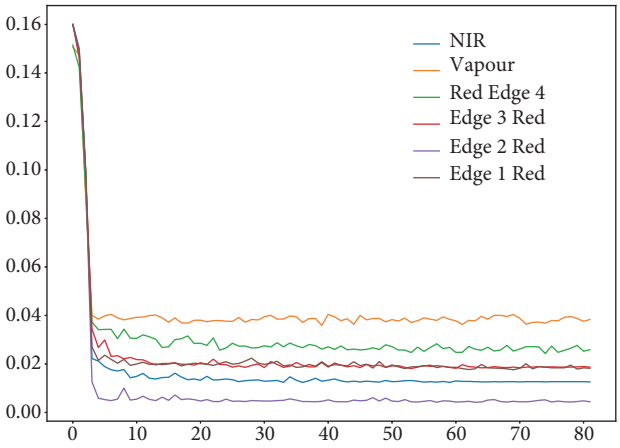

Figure 4. Validation loss of iterative generative process w.r.t. epochs. Readers should notice the significant drop in the reconstruction loss in the NIR, Vapour, and Red Edge 2 bands.

related hyperparameters have been transferred from Eurosat. We performed all our experiments on a humble computer with i5 processor, 16 GB of RAM, and 2 SLI'ed NVIDIA GTX 1070 graphics card.

Our generative model has over 8.4 million parameters and training per image required 19 ms and testing consumed 9 ms. The discriminative model was of course faster. Training a batch lasted only 30 ms and average prediction duration was no more than 2 ms. All implementations were based on TensorFlow.

3.3. Results

The results of our quantitative scene classification experiments are reported in Tables 2 and 3 as classification accuracies averaged across 5 training runs for Eurosat and 50 runs for UCM. The standard deviations were negligible, so they were omitted.

In Part A of our experiments (Table 2), where we worked exclusively with the Eurosat dataset, our baseline (*Actual*) performance was measured using the available original bands progressively together with RGB to obtain a baseline of how well the network performs without any estimation involved. As expected, classification performance increased from 0.868 using only RGB data, all the way to 0.937 with 6 more additional spectra.

Next, using the *Direct* strategy, were each additional spectral band was estimated only from RGB, the estimated spectral bands were progressively stacked together and a minor improvement of 1.5 percentile points was observed; albeit minor this result convinced us that the network can in fact produce useful data.

Then, with the *Iterative* strategy, the best generated band with respect to the reconstruction loss was combined with the original RGB image and used together in estimating the next band. In this case, the improvement w.r.t. using only RGB was recorded as 3.6 percentile points (0.904), indicating that the network has estimated a significant amount of classification-wise useful data.

Nevertheless, despite the promising improvements, one can easily argue the practical interest of this validation strategy, since Eurosat already possesses non-RGB bands. That is why we put the method to test in Part B using UCM, which is an RGB-only dataset, at a much finer spatial resolution and explored whether a network trained to estimate non-RGB bands with one dataset (Eurosat) can transfer its know-how to another that it has not witnessed before. Surprisingly, it turned out (Table 3) that this is not impossible, as

an improvement of 1.4 percentile points was observed (0.931 from 0.917). In order to safely rule out network biases, these values were computed as means across 50 runs.

Table 2. Scene classification accuracy with the Eurosat dataset (Part A). Each column represents the spectral band included along the data to its left during testing; e.g., "+ Red 2" denotes "RGB, Red 1 and Red 2". Best results are given in bold font.

	RGB	+ Red 1	+ Red 2	+ Red 3	+ NIR	+ Red 4	+ Vapour
Actual	0.868	0.883	0.925	0.931	0.933	0.935	**0.937**
Direct	0.868	0.882	0.879	**0.883**	0.879	0.880	0.876
Iterative	0.868	0.882	0.895	0.898	0.896	0.894	**0.904**

Table 3. Scene classification accuracies with the AID and UCM datasets, using estimated channels along with RGB data (Part B). Best results are given in bold font.

	RGB	..+Red 1	..+Red 2	..+Red 3	..+NIR	..+Red 4	..+Vapour
UCM	0.917	0.922	0.926	0.924	**0.931**	0.922	0.927
AID	0.864	0.872	0.874	0.87	0.874	**0.88**	0.879

4. Discussion

Our work proposes that scene information can be used to estimate additional spectra of that particular scene. Numerical results presented in the previous section are encouraging in this direction. Intuitively, this claim is not bulletproof. From a scientific point of view, it not possible to estimate EM emission or scattering of a nonblack body from its visual spectra. We are well aware that our empirical demonstration does not present proof that the developed generative model is in fact estimating non-RGB bands from RGB, but our assumptions and the proposed method do not rely on deterministic mathematical models but rather draw power from statistical learning theory. We assume that there are statistically significant amounts of data to capture the relations between various channels and between scene content. If there is a statistical relation between these objects and channels, the generative model tries to approximate the latent distribution between content and the channels. In Table 3 we presented that there is a small but significant correlation between channels and content.

The main weakness of the proposed method is that the generative model needs immense amounts of data to capture the latent distribution. This is not unique to our case. This behavior can be seen in almost all statistical learning machinery. Lacking the necessary amount of training data hurts our model severely, since our solution depends not only on relations between bands but also relations between content and source of these multiband image samples. Intuitively, another drawback of our model is that training samples should cover a vast number of multiband images with highly diversified scene content; in other words, it needs very different scene images with high intraclass variance so that it can encode the relation between content, scene, and the channels.

Nevertheless, the proposed method shows promising results although the performance of estimated bands is not as high as original bands on classification tasks. We show that our method encoded some information between scene bands and scene contents. We chose to refer to this information as side-channel information, and the practical potential of this information has been measured in the context of our scene classification tests as positively nonnegligible. The knowledge transfer test to a distinct dataset (from Eurosat to UCM) was particularly important from this regard, as it showed that the learned relations between RGB and near infrared bands are not specific to the training Eurosat dataset and are not impossible to generalize. This opens a wide

direction of research especially in terms of pixel-based spectrum estimation and its effect on pixel classification performance and land-cover map calculation.

5. Conclusion

This paper has explored the estimation of infrared channels through neural generative models and their practical use in a scene classification context. Two alternative strategies have been explored; the first relies on the use of only RGB bands and the second approach, which we named iterative generation, is based on merging the estimated bands with RGB to subsequently predict unseen bands. The selection of these bands to target for estimation has been also investigated. Overall, at this preliminary stage our qualitative and quantitative results have shown promising empirical indications that generative models can contribute up to a certain degree to the estimation of non-RGB bands and definitely merit further research.

References

[1] Nasrabadi NM. Hyperspectral target detection: an overview of current and future challenges. IEEE Signal Processing Magazine 2014; 31: 34–44.

[2] Liu Y, Liu Y, Ding L. Scene classification based on two-stage deep feature fusion. IEEE Geoscience Remote Sensing 2018; 15: 183–186.

[3] Weng Q, Mao Z, Lin J, Guo W. Land-use classification via extreme learning classifier based on deep convolutional features. IEEE Geoscience and Remote Sensing Letters 2017; 10: 704–708.

[4] Yu Y, Lu F. Aerial scene classification via multilevel fusion based on deep convolutional neural networks. IEEE Geoscience and Remote Sensing Letters 2018; 15: 287–291.

[5] Aptoula E, Dalla Mura M, Lefevre S. Vector attribute profiles for hyperspectral image classification. IEEE Transactions on Geoscience and Remote Sensing 2016; 54: 3208–3220.

[6] Pham MT, Lefevre S, Aptoula E. Local feature-based attribute profiles for optical remote sensing image classification. IEEE Transactions on Geoscience and Remote Sensing 2018; 65: 1199–1212.

[7] Pham MT, Aptoula E, Lefevre S. Feature profiles from attribute filtering for classification of remote sensing images. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing 2018; 11: 249–256.

[8] Cheng G, Han J, Lu X. Remote sensing image scene classification: benchmark and state of the art. Proceedings of the IEEE 2017; 105: 1865–1883.

[9] Chen X, Duan Y, Houthooft R, Schulman J, Sutskever I et al. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In: Proceedings of NIPS; Barcelona, Spain; 2016. pp. 2172–2180.

[10] Kingma DP, Welling M. Auto-encoding variational Bayes. CoRR, vol. abs/1312.6114, 2013.

[11] Hinton G, Salakhutdinov R. Reducing the dimensionality of data with neural networks. Science 2006; 31: 504–507.

[12] Masci J, Meier U, Ciresan D, Schmidhuber J. Stacked convolutional auto-encoders for hierarchical feature extraction. In: Proceedings of the Artificial Neural Networks and Machine Learning Conference; Espoo, Finland; 2015. pp. 52–59.

[13] Helber P, Bischke B, Dengel A, Borth D. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. CoRR, vol. abs/1709.00029, 2017.

[14] Yang Y, Newsam S. Geographic image retrieval using local invariant features. IEEE Transactions on Geoscience and Remote Sensing 2013; 10: 818–832.

[15] Xia GS, Hu J, Hu F, Shi B, Bai X et al. Aid: A benchmark data set for performance evaluaton of aerial scene classification. IEEE Transactions on Geoscience and Remote Sensing 2017; 55: 3965–3981.