Low Resolution Recognition of Aerial Images

Raghunath Sai Puttagunta*,Renlong Hang*, Zhu Li*, and Shuvra Bhattacharyya†

*University of Missouri-Kansas City
Kansas City, USA

†University of Maryland
College Park, USA

***Abstract*—Remote sensing for classification has been widely studied and is useful for a lot of applications like precision agriculture, surveillance, and military applications. Recently, due to tremendous results achieved by deep learning using Convolutional Neural Networks (CNN) for Imagenet dataset, there have been a large number of works which use deep learning for aerial image classification. Most of the works concentrate on original resolution and there are no works on low-resolution recognition of aerial images. This work is critical because aerial images are taken from a very high distance from the ground and the cost of installing high definition cameras is high, so it is hard to get a high resolution of the image. In this paper, we explore how we can do the better classification of aerial images for original spatial resolution and low spatial resolution in deep learning by using texture information. In our framework, we use YUV color space which is generally used for video coding and we also use Laplacian of Gaussian (LOG) information to exploit the texture information. We decouple RGB information into luminance information (Y channel), color information (UV) and texture information (LOG) and we train a separate CNN for each feature and combine them using autoencoder and with our results, we show that we do better than RGB images in original resolution and low resolution.**

***Index Terms*—Remote Sensing, Classification, Deep Learning, Low Resolution Recognition**

I. Introduction

Remote Sensing has attracted a lot of interest because of large applications in aerial image analysis. Aerial Images help us to get a clear picture of earth's surface and a great data source for earth observation. Classification of aerial images is widely researched and useful to understand the landscape. Aerial Images classification could be used to track the development of an area and helps in better planning of the area, especially in urban areas. Aerial imaging can also be used for surveillance, military applications. and tracking natural resources like forests to track the depletion rates. Recently aerial Images are also used to track the exact location and navigate in urban areas as the global positing system (GPS) is not reliable in urban areas. In Aerial Imaging due to the high cost of installing high definition cameras, it becomes hard to get a high-resolution aerial image which in turn causes images to be low resolution and which makes it harder for recognition tasks. In this work, we do better remote sensing classification on original and low-resolution aerial images.

The early works in aerial image classification used handcraft descriptors like SIFT and aggregated those key point features using some aggregating tools like Bag of Visual Words (BoVW) and pass it through Support Vector Machine (SVM) . With the popularity of data-driven methods like deep learning for classification in computer vision, Convolutional Neural Networks (CNN) methods were used in many works for remote sensing classification using deep learning and for classification and they achieved a good accuracy and convolutional neural networks like Visual Geometry Group (VGG) [1] network and Residual Network (ResNet) are generally used because they have achieved good accuracy's for ImageNet challenge [2].

In this paper, we make the first attempt to work on low-resolution recognition for aerial images using deep learning. Most of the works in low-resolution recognition are for face recognition and text recognition system. These previous works use deep learning based Super Resolution (SR) for preserving the original image details.In our work we used bicubic interpolation. For recognition, texture and color information is useful. In this work we propose a novel scheme where we decouple RGB image into YUV image and also add Laplacian of Gaussian (LOG) information for the classification task. We add LOG information because, in aerial images, texture information is more useful than color information . We use deep learning for the recognition tasks and we prove with our results a way how to do better recognition than just using RGB images in original resolution and also in low resolution.

The rest of the paper is as follows in section 2 we discuss some of the related works on the classification of aerial images. In section 3 we will discuss our framework. In section 4 we will discuss the experiments and results and in section 5 we will discuss the conclusion.

II. Related Work

There has been significant research in the aerial image classification using both handcrafted features and deep learning features for remote sensing classification. Earlier efforts for remote sensing classification have been using handcrafted features. These features try to exploit the color, texture, shape and spatial information of the images which are useful for classification. The handcraft features include color histogram, texture descriptors like SIFT, GIST, HOG. The color histogram is the easiest handcraft feature to compute and exploits the color information for classification. The color histogram descriptor is invariant to rotation and translation for the viewing axis, but the major disadvantage is it doesn't convey any spatial information. For texture information, there have been works such as Gabor Transform [3] and Local Binary Patterns [4]

978-1-7281-3723-0/19/$31.00 ©2019 IEEE

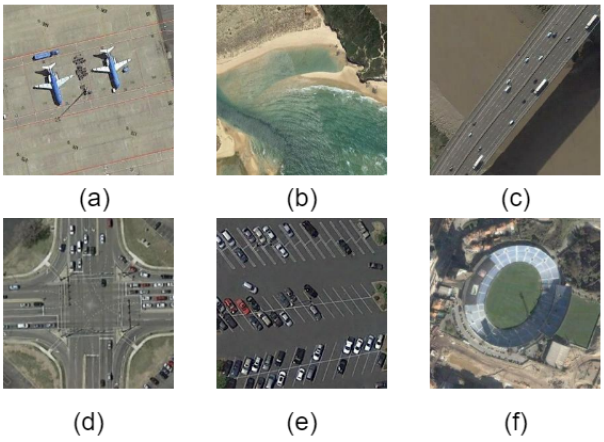

Fig. 1. Aerial Images from NWPU-RESISC 45 dataset with 6 different classes a) Airplane b) Beach c) Bridge d) Intersection e) Parking lot f) Stadium

which exploit the texture features. There also have been works using GIST features which are global features and can be used for classification directly. SIFT features are used for key point detection in an image and unlike color histogram and GIST [5] which are global features SIFT is a local feature and cannot be used directly for classification, so we use aggregating tools like BoVW with SIFT for classification. HOG feature [6] like SIFT [7] is also a local feature and is represented by computing the gradient intensities and orientation in spatial regions. Like SIFT it also needs an aggregating tool to classify it. The handcraft features fail when the images become challenging. To overcome this there were works on unsupervised feature learning which was an alternative to handcraft features. In unsupervised learning, we don't have the label data and we set to learn discriminative features by learning a set of basis functions whose inputs are pixel intensity values or handcraft features. Some of the unsupervised methods works for aerial images include PCA, k-means clustering and auto encoders

The most recent works in aerial image classification are based on deep learning because of the availability of large datasets and their high accuracies in classifying Imagenet dataset. Initially works done, use state of the art performing CNN's like VGG-16 network and use them as extractors and classify the features using Support Vector Machine (SVM) or just fine-tune the weight of the networks. Also, there are other works which have used features from multiple convolutional layers and use a coding technique like VLAD or Fisher vector for classification. Even though deep learning based methods have achieved high accuracy's using CNN as either feature extractors or classifiers none of them have adopted the texture based information for classification using deep learning or worked on low-resolution recognition of aerial images. In [8] they have used multiple layers features of CNN and they combined them using fisher kernel coding method to get good classification accuracies. In this paper, we discuss how we use deep learning with texture information to have better recognition in low resolution. In [9] instead of using cross-entropy loss for classification, they have metric learning based method to achieve better classification results. There have been fewer works on low-resolution recognition. Most of the works deal with facial images and text recognition. In [10] they used CNN's for low-resolution recognition. They have worked using CIFAR-10 and CIFAR-100 datasets whose pixel resolutions are 32 x 32 and they conducted their experiments using 2 and 4 times downsampled images and they extended it for facial recognition and text recognition. The original resolution of the datasets is not really large. In our work, our pixel resolution of the original image is 256 x 256. In [11] they used deep learning for text recognition where the resolution is low. In [12] they used low-quality videos to recognize human emotions. In [13] they preserved gradient information for doing Super Resolution of low resolution images.

III. Proposed Method

Our proposed framework decouples RGB image into YUV and LOG features, we use the texture information because it is useful for aerial images classification. The texture information becomes critical in the case of low-resolution recognition. In our work, we divide the YUV color space into the Y-channel (luminance) which gives us brightness and UV-channel which contains the color information. YUV color space are generally used for video coding and Y component is very similar to gray scale image and UV channel is also called as chrominance components.For texture information we use LOG, we use laplacian filters for LOG which are derivative filters and are used to find edge information (gradient information). Generally when you use laplacian filters they are highly susceptible to noise and to overcome that we use smooth the filter before giving it to laplacian filter and these two steps are called LOG. The LOG for an image is given by

$$LoG(x,y) = -\frac{1}{\pi\sigma^4}\Big[1 - \frac{x^2+y^2}{2\sigma^2}\Big]e^{-\frac{x^2+y^2}{2\sigma^2}} \qquad (1)$$

The sigma value in the above equation is the standard deviation of the Gaussian filter and x,y are co-ordinates of pixels. In our work we generate LOG by passing the luminance information (Y-channel) to the filter. The sigma we used for our work was 0.5.

For classification in remote sensing we used CNN's and VGG-11 network was used in our work. We changed the fully connected layers of the network and finetuned it for our own dataset and the architecture is shown in figure 2. For low-resolution recognition, we use the images which were downsampled 2 and 4 times. To have a better recognition at the original resolution we train 3 VGG-11 networks separately for luminance information, color information and texture (LOG information). Each network was trained using cross-entropy loss and we extract the features and we combine them using an auto encoder which has one hidden layer and one output layer as shown in figure 3. The inputs to the output layer are addition of networks information after input layer. In the low resolution, we deal with the lower spatial resolution of the image by just downsampling the image and upsampling it back to the original size by using bicubic upsampling. The images

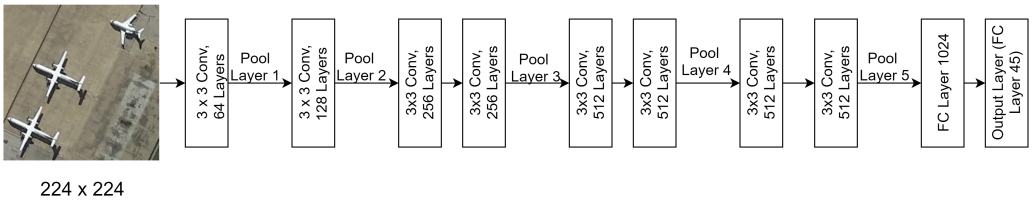

Fig. 2. VGG 11 modified Architecutre

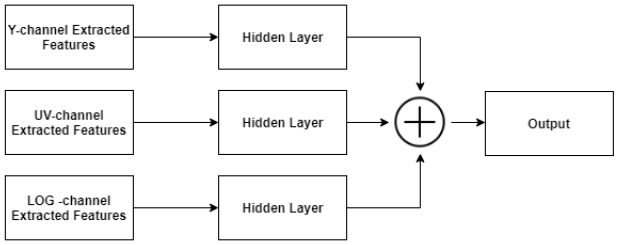

Fig. 3. Our framework to combine luminance, color information and LOG information

lose a lot of information by using that especially in case of low resolution like 4 times downsample of an image. In the case of low spatial resolution, we use the same framework as we have done for the original spatial resolution but instead of retraining the VGG-11 network with low resolution, we use weights from the original trained VGG-11. Using the original resolution network, we freeze the training of convolutional layers of the network and train only the fully connected layer features and then combine them by training the auto encoder as shown in figure 3.

IV. Experiments and Results

In our experiments, we use NWPU-RESISC 45 dataset [14], a publicly available Remote Sensing dataset. This dataset has 31,500 images with 45 classes and each class has 700 images. Some of the classes from the dataset are shown in figure 1. This dataset is suitable for deep learning based methods because we have large amounts of data to train the CNN's for classification. Each image in the dataset is of size 256 x 256 which is really useful for CNN's because most of the networks accept 224 x 224 as the input resolution size. We divide our dataset into 50% training and 50% testing (validation). In our experiments for the low resolution, we have used 2 and 4 times downsampled images. The downsampling was done by bicubic interpolation and the downsampled images were upsampled back to original resolution so that we can feed them as an input to the VGG-11 network. For LOG we use Gaussian filter of sigma 0.5 and we use luminance information to get LOG. We tested multiple values of sigmas for LOG and we achieved the highest accuracy for sigma 0.5 so we chose that value. We also tried different values for auto encoder and it was found out that the value of 256 is the best value for the hidden layer. We also tried multiple hidden layers sizes and found that only one hidden layer size was working the best.

For our recognition task, we use VGG-11 network. Initially, we separately train Y-channel, UV-channel, and LOG with VGG-11 network with our training dataset. We compare our results with RGB images. Each network is trained till 100 epochs and we use the cross-entropy loss to train the network. We use Adam optimizer with an initial learning rate of 1e-4 and we decay the learning rate by 0.1 at multiple steps at 25, 40 and 70th epoch. We can observe from the results in table 1 that individually Y-channel, UV-channel, and LOG-channel don't perform well but when we combine and train them as shown in figure 3 the accuracy is almost 2% more than what we have just using RGB images. The autoencoder used to combine the features was trained using cross-entropy loss and Adam optimizer was used for that as well.

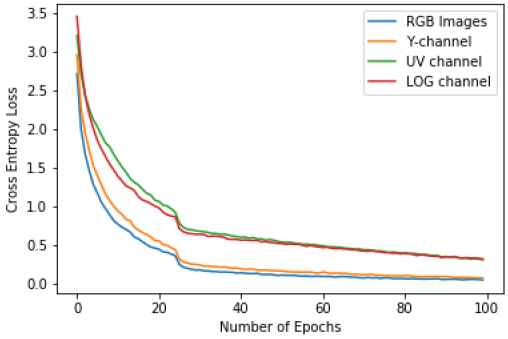

Fig. 4. Loss Values per epoch for original Resolution

TABLE I
Accuracy values of Experiments

Feature	Resolution	Accuracy
RGB Image	Original Resolution	88.92
Y channel	Original Resolution	86.50
UV channel	Original Resolution	73.72
LOG channel	Original Resolution	79.49
Y+UV+LOG	Original Resolution	91.25
RGB	2x downsampled Image	86.30
Y channel	2x downsampled Image	82.30
UV channel	2x downsampled Image	65.65
LOG channel	2x downsampled Image	67.74
Y+UV+LOG	2x downsampled Image	87.65
RGB	4x downsampled Image	78.31
Y channel	4x downsampled Image	71.62
UV channel	4x downsampled Image	51.02
LOG channel	4x downsampled Image	50.95
Y+UV+LOG	4x downsampled Image	78.89

In case of low resolution, we don't train the entire network we only train the fully connected layers of the network for RGB, Y, UV, and LOG. We use Adam optimizer for low resolution as well. We start with a learning rate of 1e-4 We again combine the features extracted from Y, UV, and LOG by training an autoencoder. We train the autoencoder for 150 epochs with cross-entropy loss and we can see from table 1 that the accuracy of the combination of features is around 1% more than what we have for RGB in case of 2 times downsampled images and combination of features around 0.8% more than what we had for 4x downsampled images. These results show that the texture information is really useful for aerial image classification for original and low-resolution images. We can also observe from figure 5 that UV and LOG for low resolution didn't train as much but combining it with Y-channel and training the autoencoder worked better than what we had for just using RGB images.

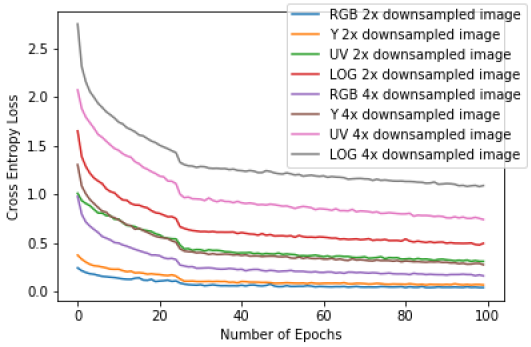

Fig. 5. Loss Values per epoch for Low Resolution

V. Conclusion

In this paper, we introduce a new approach to do better recognition of aerial images in low resolution. We take an RGB image and we decouple them into the Y channel, UV channel, and LOG channel. By doing this we have luminance, color and texture information. We train these features (information) with a VGG-11 network separately and combine them by training an autoencoder. Using this process we have shown that we have better recognition than just using an RGB image for classification of aerial images. Our results show that for aerial images adding texture information is useful for recognition in aerial images since luminance and color information are already in RGB color space. Our current method performs better than using RGB images by 2% for original resolution and around 1% for 2x and 0.5% for 4x low resolution.

In future work, we can extend the work we did and work on the low resolution by preserving gradient information in low resolution like in [13] where they preserved the Difference of Gaussian (DOG) information in low resolution. We can also develop a deep learning based CNN which is scalar invariant using triplet loss.

VI. Acknowledgement

This research was supported in part by the U.S. Air Force Office of Scientific Research under the Dynamic Data Driven Applications Systems (DDDAS) Program.

References

[1] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. [Online]. Available: http://arxiv.org/abs/1409.1556

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 2016, pp. 770–778. [Online]. Available: https://doi.org/10.1109/CVPR.2016.90

[3] A. K. Jain, N. K. Ratha, and S. Lakshmanan, "Object detection using gabor filters," *Pattern Recognition*, vol. 30, no. 2, pp. 295–309, 1997. [Online]. Available: https://doi.org/10.1016/S0031-3203(96)00068-4

[4] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, July 2002.

[5] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision*, vol. 42, no. 3, pp. 145–175, May 2001. [Online]. Available: https://doi.org/10.1023/A:1011139631724

[6] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, 2005, pp. 886–893. [Online]. Available: https://doi.org/10.1109/CVPR.2005.177

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, Nov 2004. [Online]. Available: https://doi.org/10.1023/B:VISI.0000029664.99615.94

[8] E. Li, J. Xia, P. Du, C. Lin, and A. Samat, "Integrating multilayer features of convolutional neural networks for remote sensing scene classification," *IEEE Trans. Geoscience and Remote Sensing*, vol. 55, no. 10, pp. 5653–5665, 2017. [Online]. Available: https://doi.org/10.1109/TGRS.2017.2711275

[9] G. Cheng, C. Yang, X. Yao, L. Guo, and J. Han, "When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns," *IEEE Trans. Geoscience and Remote Sensing*, vol. 56, no. 5, pp. 2811–2821, 2018. [Online]. Available: https://doi.org/10.1109/TGRS.2017.2783902

[10] Z. Wang, S. Chang, Y. Yang, D. Liu, and T. S. Huang, "Studying very low resolution recognition using deep networks," *CoRR*, vol. abs/1601.04153, 2016. [Online]. Available: http://arxiv.org/abs/1601.04153

[11] M. Jaderberg, A. Vedaldi, and A. Zisserman, "Deep features for text spotting," in *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV*, 2014, pp. 512–528. [Online]. Available: https://doi.org/10.1007/978-3-319-10593-2_34

[12] B. Cheng, Z. Wang, Z. Zhang, Z. Li, D. Liu, J. Yang, S. Huang, and T. S. Huang, "Robust emotion recognition from low quality and low bit rate video: A deep learning approach," in *Seventh International Conference on Affective Computing and Intelligent Interaction, ACII 2017, San Antonio, TX, USA, October 23-26, 2017*, 2017, pp. 65–70. [Online]. Available: https://doi.org/10.1109/ACII.2017.8273580

[13] D. F. Noor, Y. Li, Z. Li, S. Bhattacharyya, and G. York, "Gradient image super-resolution for low-resolution image recognition," in *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2019, pp. 2332–2336.

[14] G. Cheng, J. Han, and X. Lu, "Remote sensing image scene classification: Benchmark and state of the art," *CoRR*, vol. abs/1703.00121, 2017. [Online]. Available: http://arxiv.org/abs/1703.00121