I.J. Engineering and Manufacturing, **2018, 4, 13-20**
Published Online July 2018 in MECS (http://www.mecs-press.net)
DOI: 10.5815/ijem.2018.04.02

Available online at http://www.mecs-press.net/ijem

Remote Sensing Image Scene Classification

Md. Arafat Hussain, Emon Kumar Dey*

Institute of Infromation Technology, University of Dhaka, Dhaka-1000, Bangladesh

Received: 11 March 2018; Accepted: 06 June 2018; Published: 08 July 2018

Abstract

Remote sensing image scene classification has gained remarkable attention because of its versatile use in different applications like geospatial object detection, natural hazards detection, geographic image retrieval, environment monitoring and etc. We have used the strength of convolutional neural network in scene image classification and proposed a new CNN to classify the images. Pre-trained VGG16 and ResNet50 are used to reduce overfitting and the training time in this paper. We have experimented on a recently proposed NWPU-RESISC45 dataset which is the largest dataset of remote sensing scene images. This paper found a significant improvement of accuracy by applying the proposed CNN and also the approaches have applied.

Index Terms: Convolutional Neural Network, Remote Sensing Image, Scene Classification, CNN.

1. Introduction

Image or picture is a visual representation anything. Human can see a complex picture, understand it and can interpret properly without thinking twice but it is a very difficult task for machine. A machine can learn from the provided labelled data. Due to machines learning capability, image recognition and classification has become an active research area.

Image classification and recognition is important for various applications a like facial image classification, medical image classification, scene image classification, human pose classification, gender classification etc. Recently scene image classification has found remarkable attention for its growing importance. Researchers are trying to efficiently classify different types of scene images like natural scene, remote sensing scene, indoor scene etc.

Satellite and airborne observing the earth and taking a huge amount of remote sensing scene images. With the development of modern satellite and airborne, high spatial resolution (HSR) remote sensing images can

* Corresponding author.
E-mail address: emonkd@iit.du.ac.bd

provide detailed information. Remote sensing image classification has great importance in many applications such as natural vegetation mapping, urban planning, geospatial object detection, hazards detection, geographic image retrieval and environment monitoring [1]. The effectiveness of these applications depends on the classification accuracy.

The goal of the classification is to categorize a scene image in appropriate classes. In the past years, some datasets have proposed on remote sensing image. These datasets are very small and already found high accuracy using feature extraction based hand crafted methods. The benefits of deep learning approach cannot be found with these small datasets because of various reasons. Recently, few large datasets have been published and some models have already been applied also. But there is a huge scope to improve the performance of this dataset. For example, a recently published NWPU-RESISC45 dataset [2] with 311,000 images and 45 scene classes achieved 90.36% accuracy using fine tuning VGGNet-16 [3] which is the current state of the art solution.

On the other hand, current state-of-the-art hand crafted feature extraction methods showing poor performance on large dataset for scene classification. For example, the NWPU-RESISC45 achieved only 24.84% accuracy using the colored histogram features of this dataset. So the major objective of this research is how we can improve the performances remote sensing dataset using hand engineered feature extraction method as well as deep learning architecture.

Many experiments happened on focusing remote sensing scene image using different datasets. Local descriptors like local binary pattern, local ternary pattern, completed local binary pattern or histograms of oriented gradients have proven their worth in different scene classification. Hand engineered based feature extraction method was very popular for scene detection before the deep learning based approach was proposed. Still it is very attractive for its simplicities. Different researches have proposed different ways to solve this problem. Xiao et al. proposed the SUN for large scale scene detection in [4]. The dataset contains 899 categories of 130519 images. Which helped the scene detection research group of the world a lot. They also applied different feature extraction method to find the accuracy of the dataset and found a satisfactory state of the art result.

Lienou et. al. [5] conducted experiment on satellite images using sematic concept. They applied latent dirichle allocation model is used to training for each class. The model basically obtains their feature by extracting image words. They obtained 96.5% accuracy which is comparatively support vector machine which gained 85% accuracy on QuickBird image dataset.

Vatsavai et. al. [6] proposed an unsupervised way based on the Latent Dirichlet Allocation (LDA) method which is a semantic labelling framework. They collected 130 images of refineries, international nuclear plants, airports and coal power plants. Though there were reasonable overlapping between two classes they found 73% training accuracy and 67% testing accuracy.

Deep learning has come with a revolutionary change in the field of machine learning. Accuracy of different datasets jumped after applying deep learning approach [21-23]. The typical Convolutional Neural Networks (ConvNet) including Alexnet [7], VGGNET [8], GoogleNet [9] has been applied for scene classification purpose. As, in most of the existing scene dataset number of images are small, the activations of different layers of ConvNet or their fusion considered as the visual representation sent to the classification layer of ConVNet. For example, in [10-12] the authors used deep learning model for scene classification based on 7 layer AlexNet or CaffeNet [13]. Castelluccio, Marco [14] achieved state of the art accuracy on UC Merced dataset and the Brazilian coffee scene dataset. They applied CaffeNet and GoogLeNet with various learning procedures on these datasets. They used these two network from scratch and also used fine-tuning on their experimental dataset and they used the procedure to reduce overfitting. They found 97.10% accuracy using fine-tuned GoogleNet on UC Merced dataset and 91.83% accuracy using GoogleNet on Brazilian coffee scene dataset.

VGG architecture consisting 16 layers have also been applied in [15-16]. Wang, Limin, et al. used [17] VGG networks to find state of the art accuracy on Place205, MIT67 and SUN397. They used Caffe toolbox to design ConvNets. They designed three types of VGG nets named VGG11, VGG13, VGG16. Trained the VGG nets on

Place205 dataset. To reduce computation cost they loaded weight to VGG13 from VGG11 and from VGG13 to VGG16.

Many deep learning based models have proposed to classify images of different dataset. But there is no specific model or approach which is best suited for all dataset. A deep knowledge in this field and many year research experience can give an appropriate path to design a good model for a dataset. In this paper we have focused on the improvement of the performances of large-scale image dataset. We have experimented on NWPU-RESISC45 using our proposed architecture of deep learning model. We have found a reasonable accuracy.

2. Methodology

In this paper we have applied two types of approaches to evaluate the dataset. We applied both transfer learning approach and training from the scratch. Transfer learning fine tunes the existing deep convolutional neural network models. We considered VGG16 and ResNet50 for the transfer learning. In 2.3 we described our proposed from the scratch. Following sub-sections describe the methodologies.

2.1. Transfer Learning using Pre-trained VGG16 Model

We fine-tuned the VGG16 model which was trained on ImageNet dataset [7]. As this dataset covers a huge variety of images, a model that is trained on this dataset with a good accuracy rate has a huge knowledge indeed. We just want to fit this knowledge for our dataset. We consider two factors, one is the size of the dataset another is the similarity with the ImageNet dataset. Our dataset is comparatively small and partially similar with ImageNet dataset. In this situation, partially training the VGG16 model will be best. We kept the convolutional part same as before but changed the fully connected layers. We replaced the first fully-connected layer with 256 neurons, second fully-connected layer with 256 neurons and third with 45 neurons. We replaced 4096 by 256 to reduce overfitting by reducing the parameters.

2.2. Transfer Learning using Pre-trained ResNet50 Model

We applied fine-tuning on pre-trained model not only changing the output layer and training the classifier but also training some layers near the last layer. Here last layer is replaced with six layers. After the last average pooling we added a global average pooling layer to reduce dimensionality of input. It also works an alternative of flatten layer by producing one dimensional vector from 3 dimension. Then a fully-connected layer is added with 256 neurons. After that a dropout layer is added and another fully-connected layer of 256 neurons follows the dropout layer. Another dropout layer follows the second fully connected layer. The last layer of the model is a fully-connected layer with 45 neurons with a softmax classifier.

2.3. Proposed CNN model from Scratch

The model is a VGGNet based seven-layer convolutional neural network. First four layers are convolutional layer and other three are fully connected layers. There are four max-pooling layer followed by each convolutional layer. Two dropout layers are added which follows first two convolutional layers. Last fully-connected layer is basically classification layer which classify images into different classes. Figure 1 shows the basic architecture of our proposed model. Table 1 shows every layers, their output and parameters. Convolutional layers (marked in blue) and fully-connected layers (marked in yellow) are the skeleton of our model.

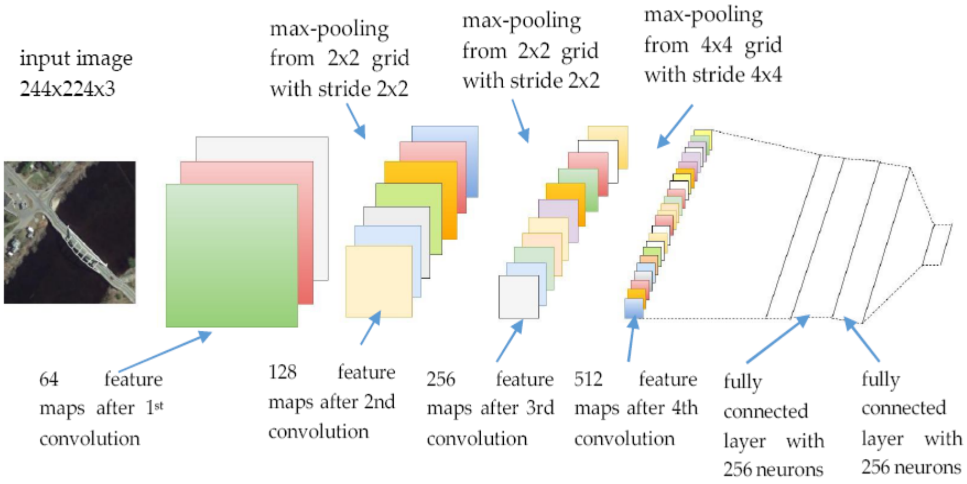
input image
244x224x3
max-pooling
from 2x2 grid
with stride 2x2
max-pooling
from 2x2 grid
with stride 2x2
max-pooling
from 4x4 grid
with stride 4x4
64 feature maps after 1st convolution
128 feature maps after 2nd convolution
256 feature maps after 3rd convolution
512 feature maps after 4th convolution
fully connected layer with 256 neurons
fully connected layer with 256 neurons

Fig.1. Proposed Model

Table 1. Layers, Output Shape and Parameters of Proposed

Layer (type)	Output Shape	Parameter
zero_padding2d_1 ((Zero Padding))	3, 226, 226	0
conv2d_1 (Conv2D)	64, 224, 224	1792
max_pooling2d_1 (MaxPooling2	64, 112, 112	0
zero_padding2d_2 (Zero Padding)	64, 114, 114	0
conv2d_3 (Conv2D)	256, 56, 56	295168
max_pooling2d_2 (MaxPooling2D)	128, 56, 56	0
zero_padding2d_3 (Zero Padding)	128, 58, 58	0
conv2d_3 (Conv2D)	256, 56, 56	295168
max_pooling2d_3 (MaxPooling2	256, 28, 28	0
zero_padding2d_4 (Zero Padding)	256, 30, 30	0
conv2d_4 (Conv2D)	512, 28, 28	1180160
max_pooling2d_4 (MaxPooling2)	512, 7, 7	0
flatten_1 (Flatten)	25088	0
dense_1 (Dense)	256	6422784
dropout_1 (Dropout)	256	0
dense_2 (Dense)	256	65792
dropout_1 (Dropout)	256	0
dense_3 (Dense)	12	3084
softmax	12	0

3. Experimental Results

In this paper we have used NWPU-RESISC45 dataset. We applied some hand engineered learning procedure and deep learning procedure in this experiment. In deep learning procedure we used two well-known deep learning model VGG16 and ResNet50. We, trained them from scratch as well as applied transfer learning for the dataset. We used Theano, Tensorflow as machine learning library, Keras as deep learning API. The experiment was implemented on Windows 10 operating system.

3.1. Dataset

NWPU-RESISC45 dataset is used in this experiment. It is one of the largest remote sensing image scene dataset. The dataset contains 31,500 images in total with 45 scene classes and each class contain 700 images. The dataset is developed from the inspiration of applying many data driven algorithms like deep learning. Scarcity of data severely limits the learning capability of deep learning models which was the main motivation behind this dataset. So, the dataset is produced to face the problem by combining a number of datasets. Firstly we have used twelve scene classes for our experiment due to our resource constraints. We split the dataset into two parts where 80% data is used for training and another 20% for validation. Images were randomly chosen for training and validation. Figure 2 shows some sample images from NWPU-RESISC45 dataset.

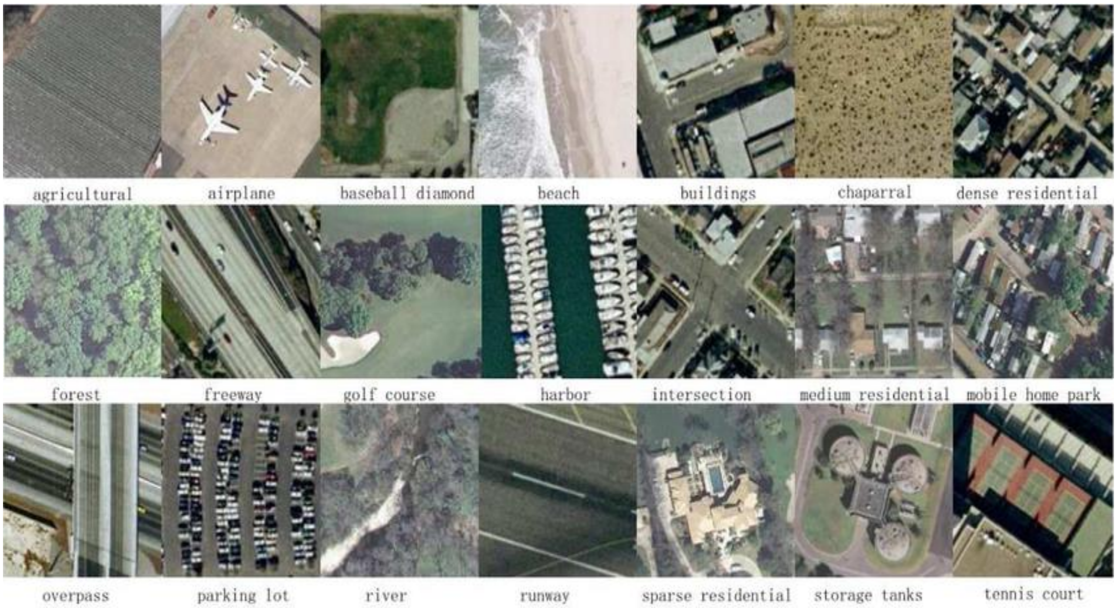

Fig.2. Sample Images from NWPU-RESISC45 Dataset

3.2. Results

We have applied the proposed models and compared the result with existing methods. Table 2 shows the experimental results.

Table 2 shows that the handheld feature extraction method NABP shows 71.76% accuracy, again 72.28% accuracy have been found using CENTRIST and 73.80% using tCENTRIST. All of these three methods was performed using 5 fold cross validation with SVM classifier.

Table 2. Experimental Results of Different Methods.

Categories	Method	Accuracy (%)
Hand engineered feature extraction method	NABP [18]	71.76%
	CENTRIST [19]	72.28%
	tCENTRIST [21]	73.80%
Training from scratch(Deep learning)	VGG11	76.9%
	VGG16	70.52%
	7 layer CNN	87.57%
Transfer learning (Deep learning)	VGG16 (classifier)	91.9%
	VGG16(fine-tuned)	90.98%
	ResNet50 (classifier)	95.9%
	ResNet50 (fine-tuned)	96.91%

In the next phase we applied VGG11, VGG16 and our proposed 7 layer CNN from the scratch. VGG11 and VGG16 shows comparatively less accuracy than our proposed CNN trained from scratch. VGG16 and VGG11 shows 70.52% and 76.9% of accuracy respectively and our proposed 7 layers CNN shows 87.57%. The result improved more than 10% than VGG11. Results of initial two models produce less accuracy. The main reason is the models are large enough and huge number of parameters but we could not feed enough data to these models due to resource constraints. So, the models over fits the data and yields a poor accuracy rate. We have reduced a number of parameters to reduce overfitting in our model.

We found our best results using transfer learning. 91.9% accuracy is found when we used VGG16 as fixed feature extractor and trained only the softmax classifier, 90.48% accuracy found by fine-tuning the last three layers of VGG16.

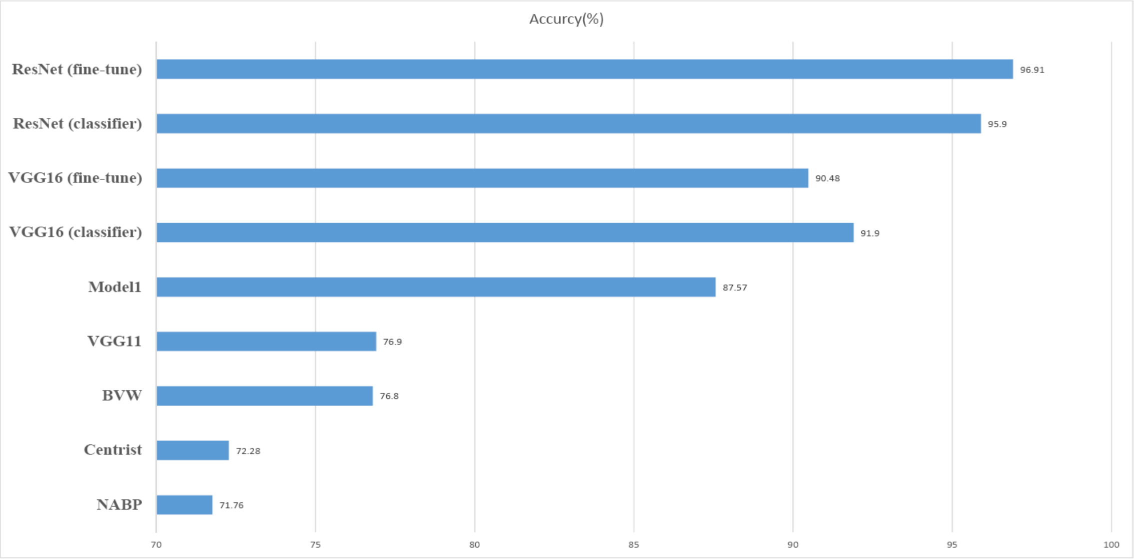

Fig.3. Experimental Results Comparison for Different Methods on NWPU-RESISC45 Dataset

Lastly we applied ResNet50 to classify the remote sensing image dataset and found best results. Using original ResNet50 we found 95.9% and after applying some fine tuning described previously we got 96.91% accuracy for our dataset. Figure 3 shows a comparison of results based on different findings using different procedures.

4. Conclusion

The goal of this experiment was to increase accuracy of remote sensing scene images. We have chosen NWPU-RESISC45 dataset. The main dataset contains 31,500 images with 45 classes. Different machine learning procedures are applied on this dataset including deep learning. We applied various types of procedures on the dataset. We proposed a new model which yields the best result (87.57%) for the models learned from scratch. Best result has been found using transfer learning (fine-tuned ResNet50 96.91%). ResNet50 is a large model compared to the proposed 7 layer CNN model. It takes long time to find output from input using ResNet. Proposed CNN takes less than half time compared to ResNet50 to take decision. But, as accuracy is our main concern, transfer learning is the best suited for the dataset we have used. Recently a new large dataset on scene image has been published named by PatternNet [20]. It is a high scaled remote sensing image dataset containing 38 classes. In future we will try to do our research on a better environment on these largest datasets. These will help us to accurately classify remote sensing scene images.

References

[1] Guo, Zhenhua, Lei Zhang, and David Zhang. "A completed modeling of local binary pattern operator for texture classification." *IEEE Transactions on Image Processing* 19.6 (2010): 1657-1663.

[2] Cheng, Gong, Junwei Han, and Xiaoqiang Lu. "Remote sensing image scene classification: benchmark and state of the art." *Proceedings of the IEEE* 105.10 (2017): 1865-1883.

[3] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[4] Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010, June). Sun database: Large-scale scene recognition from abbey to zoo. *In Computer vision and pattern recognition (CVPR)*, 2010 IEEE conference on (pp. 3485-3492). IEEE.

[5] Lienou, Marie, Henri Maitre, and Mihai Datcu. "Semantic annotation of satellite images using latent Dirichlet allocation." *IEEE Geoscience and Remote Sensing Letters* 7.1 (2010): 28-32.

[6] Vatsavai, Ranga Raju, Anil Cheriyadat, and Shaun Gleason. "Unsupervised semantic labeling framework for identification of complex facilities in high-resolution remote sensing images." Data Mining Workshops (ICDMW), 2010 *IEEE International Conference on*. IEEE, 2010.

[7] A. Krizhevsky, I. Sutskever, G. E. Hinton, Imagenet classification with deep convolutional neural networks, in: *Advances in Neural Information Processing Systems (NIPS)*, 2012, pp. 1097-1105.

[8] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37 (9) (2015) 1904-1916.

[9] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[10] F. P. S. Luus, B. P. Salmon, F. van den Bergh, B. T. J. Maharaj, Multiview deep learning for land-use classification, *IEEE Geoscience and Remote Sensing Letters* 12 (12) (2015) 2448-2452.

[11] B. Zhao, B. Huang, Y. Zhong, Transfer learning with fully pretrained deep convolution networks for land-use classification, *IEEE Geoscience and Remote Sensing Letters* 14 (9) (2017) 1436-1440.

[12] F. Hu, G.-S. Xia, J. Hu, L. Zhang, Transferring deep convolutional neural networks for the scene classification of high-resolution remote sensing imagery, *Remote Sensing* 7 (11) (2015) 14680-14707.

[13] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Ca_e:

Convolutional architecture for fast feature embedding, in: *ACM International Conference on Multimedia*, 2014, pp. 675-678.

[14] Castelluccio, Marco, et al. "Land use classification in remote sensing images by convolutional neural networks." arXiv preprint arXiv:1508.00092 (2015)

[15] K. Nogueira, O. A. Penatti, J. A. dos Santos, Towards better exploiting convolutional neural networks for remote sensing scene classi_cation, *Pattern Recognition* 61 (2017) 539 -556.

[16] G. Cheng, J. Han, X. Lu, Remote sensing image scene classi_cation: Benchmark and state of the art, *Proceedings of the IEEE* 105 (10) (2017) 1865-1883.

[17] Wang, Limin, et al. "Places205-vggnet models for scene recognition." arXiv preprint arXiv:1508.01667 (2015).

[18] Rahman, Md Mostafijur, et al. "Noise adaptive binary pattern for face image analysis." *Computer and Information Technology (ICCIT)*, 2015 18th International Conference On. IEEE, 2015.

[19] Wu, J.; Rehg, J.M. CENTRIST: A visual descriptor for scene categorization. *IEEE Trans. Pattern Anal. Mach. Intell*. 2011, 33, 1489–1501

[20] Zhou, W., Newsam, S., Li, C., & Shao, Z. (2017). Patternnet: a benchmark dataset for performance evaluation of remote sensing image retrieval. arXiv preprint arXiv:1706.03424.

[21] Dey, E. K., Tawhid, M. N. A., & Shoyaib, M. (2015). An automated system for garment texture design class identification. *Computers*, 4(3), 265-282.

[22] Islam, S. S., Rahman, S., Rahman, M. M., Dey, E. K., & Shoyaib, M. (2016, May). Application of deep learning to computer vision: A comprehensive study. *In Informatics, Electronics and Vision (ICIEV)*, 2016 5th International Conference on (pp. 592-597). IEEE.

[23] Islam, S. S., Dey, E. K., Tawhid, M. N. A., & Hossain, B. M. (2017). A CNN Based Approach for Garments Texture Design Classification. *Advances in Technology Innovation*, 2(4), 119-125.

[24] Rahman, M. M., Rahman, S., Dey, E. K., & Shoyaib, M. (2015). A gender recognition approach with an embedded preprocessing. *International Journal of Information Technology and Computer Science (IJITCS)*, 7(7), 19.

Authors' Profiles

Md. Arafat Hussain Completed Bachelor of Science in Software Engineering from Institute of Information Technology (IIT), University of Dhaka, Bangladesh in 2017. His area of interest in research is machine learning, image processing and pattern recognition.

Emon Kumar Dey is currently an assistant professor in Institute of Information Technology (IIT), University of Dhaka. He received his M.S degree from the Department of Computer Science and Engineering, University of Dhaka, Bangladesh in 2011. His research area include pattern recognition, machine learning, image processing, LiDAR data processing, 3D building modelling etc.

How to cite this paper: Md. Arafat Hussain, Emon Kumar Dey,"Remote Sensing Image Scene Classification", International Journal of Engineering and Manufacturing(IJEM), Vol.8, No.4, pp.13-20, 2018.DOI: 10.5815/ijem.2018.04.02