Parallelepiped and Mahalanobis Distance based Classification for Forestry Identification in Pakistan

Umair Khan, Nasru Minallah, Ahmad Junaid, Kashaf Gul, Nasir Ahmad
Department of Computer Systems Engineering
University of Engineering and Technology (UET)
Peshawar, Pakistan.
[umairkhan91@gmail.com, n.minallah@uetpeshawar.edu.pk, e.ahmadjunaid@gmail.com, engrkashif1@gmail.com, n.ahmad@uetpeshawar.edu.pk]

***Abstract*—Rapid deforestation has been witnessed in Pakistan over the past few years. It is taking its toll on Pakistan economy, infrastructure, and environment in the form of frequent floods. In order to keep the numbers steady frequent surveys need to be conducted. Identifying lush green forests through remote sensing is quite effective when it comes to collecting ground truth reality through extensive ground surveys. In the following study two pixels based supervised classification algorithms i.e. Parallelepiped and Mahalanobis Distance Classification Algorithms are compared for classifying forests in Pakistan. For that purpose High Geometric Resolution Imagery of SPOT-5 (2.5m) is used as the base image. According to our results Parallelepiped Classification is proved to be the better one of the two with overall accuracy of 95.4% and kappa coefficient value of 0.937, with reference to the Mahalanobis Distance classifier with overall accuracy of 85.97% and kappa coefficient value equal to 0.8115. On the basis of these findings Parallelepiped Classifier is preferred to be used for the remote sensing of forestry in Pakistan.**

Keywords—Normalized Difference Vegetation Index, Kappa Coefficient, Confusion Matrix

I. Introduction

Forestry is the study of management of significant components of forests, conserving the natural resources for fulfilling the human needs and keeping intact the habitat of all the life forms it supports. For preservation of the forests and its inhabitants information about the area covered by the forests and their types is significant. Acquisition of such information requires space based and aerial based remote sensing techniques. Among the two, space based method is widely used due to its large coverage areas as in this case forests.[1]

Overall forests in Pakistan are limited. About less than 4% of the land mass is covered by forests. Rich in terms of biodiversity they are, the number is very much lower than the optimum value of 25% for a country. In Pakistan forests cover a total area of 5.01 hectares of land mass including both the natural and artificial plantation. Having the deforestation rate of 27,000 hectares per year it is ranked the highest among Asia. Rapid deforestation is taking its toll on Pakistan's economic status, its infrastructure, and environmental conditions in the shape of large scale floods for the past few years. Floods have catastrophic effects particularly in the developing countries lacking the appropriate infrastructure to counter above average water levels. It has been always believed that the native forest cover reduces the risks and severity of catastrophic floods. Due to lack of forests both people and environment are at risk. In order to keep the updated statistics of forests periodic surveys are necessary. However, due to land mafia ground surveys are hard to conduct. Remote sensing thus can be used to identify and delineate the forests and their types [2]. With increased spatial & spectral resolution it is not only accurate but also a lot easier than the ground surveys.

Forest identification based on subsequent multi date imaging over a time period is proved to be beneficial over a single date data. However for a selected region like forest, employing more training samples on single data images for recognizing spectral signature can be used to correctly identify the forestry.

In this study, two simple and fast pixel-based algorithms are compared in terms of accuracy on a high resolution image over a large study area. These algorithms include Parallelepiped classification and Mahalanobis Distance Classification. Pixels are considered to be the smallest unit of the image data and are classified on the basis of their individual spectral values [3] . The high resolution image has been taken by SPOT-5. It was a French satellite and offered greatly enhanced capabilities providing cost effective imaging solutions [4].

The paper is structured as follows. Section II highlights the methodology and briefly describes the pre-processing of the SPOT-5 image and division of data set into testing and training pixels, section III highlights the pixel based classification and covers the brief description of the two classification algorithms, section IV gives an insight of accuracy assessment of the two algorithms in question, section V discusses the results of the two classification algorithms and section VI hence concludes the paper.

II. Methodology

A. Study Area

Khyber Pakhtunkhwa, the province of Pakistan has almost 40 percent of the country's forest cover. It is embellished with prolific and dense forests covering the major portions of Hazara and Malakand division. Therefore, the location of our study includes forest intensive areas of Abbottabad districts, Khyber Pakhtunkhwa, Pakistan. Approximately 4599 km^2 area is studied which is a subset of the acquired SPOT5 imagery. It includes densely populated urban as well as rural areas. The

978-1-5090-0436-2/15/$31.00 2015 IEEE

pilot region spans over 459989 hectares, covered through SPOT5 (2.5m) satellite imagery, acquired on October 10, 2014.

B. Pre-Processing

SPOT-5 High resolution image is obtained from SUPARCO Pakistan. The image is then subjected to pre-processing for radiometric correction. Radiometric correction is to ameliorate the data due to sensor irregularities and atmospheric noise. Satellite images may get contaminated by different types of noises. Median filter is applied over the image to clean the image removing the additive noise preserving the finest details as much as possible. The objective is to restore the image as close as to the original scene.

C. Training and Testing Data

In any supervised classification, data set is divided into two subsets for training and testing purposes. We opted for a 70:30 division of our study data set for training and testing respectively. It is recommended that most data sets train well at the point of 70/30 resulting in high accuracy [5]. Training classes include shrubs and bushes, vegetation, forests, settlements, water bodies and barren lands. Table I lists the pixel distribution of training and testing data along with the mean value of Normalized Difference Vegetation Index (NDVI).

NDVI is a numerical indicator used to analyze the remote sensing measurements to observe the green vegetation. We can derive the NDVI information by focusing on the vegetation sensitive bands i.e. red and near infrared [6].

Once the ground surveys are conducted the assessment ability of SPOT-5 image is tested. Testing is done to see if the SPOT-5 image is able to distinguish between the green lush forests and the rest of the classes or not. The tests if failed, use of classification algorithms becomes useless. The ability of discriminating among the land covers is assessed on the basis of statistics and graphical data. Statistically, the separability factors like Transformed Divergence (TD) and Jeffries-Matusita (JM) are used to distinguish between the training classes. TD is based on variance-covariance matrix as a statistical difference between classes [7]. The values of both TD and JM lie in the range of [0-2], with the greater value representing higher separability between the two classes. These values depict the distinct spectral behavior of the training classes. The two classes with low value are grouped together into one having similar spectral properties.

III. Pixel Based Classification

Pixel is considered to be the smallest unit of an image. Classical pixel based classification automatically distinguishes all pixels into land cover classes pixel by pixel. Multispectral data is used for classification and the spectral value of each pixel is used as a numerical basis for categorization. The two statistical techniques which act as a foundation for pixel based approach are supervised and unsupervised classification. We have used supervised classification algorithms in our study

Supervised Classification:

In this technique class labels into which data is to be categorized are already known [8]. The analyst selects the samples for each land cover classes also known as the training sites. Image classification software thus uses these training sites to identify the classes in the entire image as shown in the Fig. 1. Common classifiers used are Parallelepiped, Maximum Likelihood, Mahalanobis and Minimum Distance to Mean.

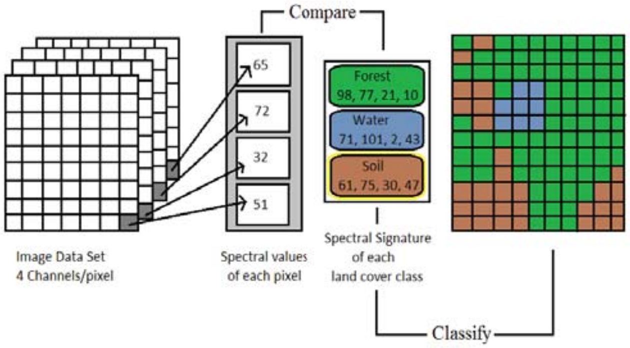

Fig. 1. Supervised Classification

A. Mahalanobis Distance Classification - MDC:

Mahalanobis Distance Classification algorithm is a direction sensitive classifier which uses statistics for each class based on covariance matrix. It is quite much similar to the Maximum Likelihood Classification (MLC) but unlike MLC it considers all class covariances to be the same. As far as complexity is concerned Mahalanobis is considered to be the simpler form of MLC and can be ranked in between Minimum

TABLE-I: Training and testing data detail of SPOT-5 Image

Name	Training Data		Testing Data	
	(Number of Pixels)	*NDVI Mean*	*(Number of Pixels)*	*NDVI Mean*
Shrubs & Bushes	17750	0.2844	7607	0.2841
Sparse Vegetation	14217	0.2258	6031	0.2259
Settlements	8443	0.1855	3616	0.1856
Forest	62515	0.2980	26795	0.2981
Water Bodies	8569	0.2325	3520	0.2313
Barren Land	56464	0.1357	24209	0.1357

distance and Maximum Likelihood Classification [9]. Mathematically Mahalanobis Distance Classifier is represented as:

$$M_h^2 = \frac{1}{2}(x - m_i)^t C_i^{-1}(x - m_i) \qquad (1)$$

where M_h^2 = Mahalanobis Distance
$|C_i|$ = Determinant of covariance matrix of class i
x = A pixel's n dimension matrix
m_i = Mean vector
t = transpose of base matrix

B. Parallelepiped Classification:

Each land cover class or training set is characterized by a unique spectral signature. Parallelepiped classifier also known as the box decision rule uses the dimensions and boundaries of each class and its respective signature. Based on the acquired information it can identify whether a pixel belongs to a particular class or not [10]. Spectral values from each pixel of the multispectral images are used to construct an n dimensional mean vector given as:

$$M_c = (\mu_{ck1}, \mu_{ck2}, \mu_{ck3}, \mu_{ck4}, \mu_{ck5} \ \dots\dots\dots \ \mu_{ckn}) \qquad (2)$$

where M_c = multispectral mean vector
μ_{ck} = mean of training data for class c in band k

The decision boundaries form an n dimensional parallelepiped in space. If a pixel value lies within the two thresholds for all the n bands the pixel is assigned to that particular class. Parallelepiped is an efficient computation method for distinguishing remote sensing data. However there are a few shortcomings of this particular algorithm. A pixel value may lies within more than one class i-e overlapping of the two classes. In such situation the pixel is classified as of the class of which it satisfies all the criteria.

The performance of the classifier is illustrated by the Receiver Operation Characteristic curve (ROC) and the Probability of Detection versus Threshold graph (PDT) [11]. The combination of the two is used to find out the optimum threshold for the classifiers where their performance will be at maximum. The optimum threshold is the point at which the probability of detection is at 80%, having the probability of false alarm as lower as possible.

Post Processing - Post processing is done to enhance the results and to measure its effect on accuracy. It is then applied to the classification results. Post processing includes the application of filters like majority, median, sieve, clump and combination of these filters etc.

IV. Accuracy Assessment

The accuracy of a classification is measured by comparing the classified pixels with some reference data reflecting the ground truth reality. The accuracy of each classifier is assessed on the basis of the testing pixels with the help of confusion matrix and two other well-known parameters. *Overall Accuracy* – It is measured as the ratio of number of pixels classified the same in satellite image and on ground to the total number of pixels.

$$\text{Overall Accuracy} = \frac{\text{Number of Classified Pixels}}{\text{Total Number of Pixels}} \qquad (3)$$

As the name indicates, it shows the overall accuracy of the classification rather than the accuracy of each class being identified individually.

Kappa Coefficient – Kappa Coefficient of statistic was developed by Cohen. It is used to measure the observed agreement between two classifiers which can classify N items into C classes [12]. KC evaluates the performance of classifiers statistically and denotes their accuracy with respect to a random classifier.

$$K = \frac{\left(N \sum_{i=1}^{j} x_{ii} - \sum_{i=1}^{j} x_{i+} x_{+i}\right)}{N^2 - \sum_{i=1}^{j} x_{i+} x_{+i}} \qquad (4)$$

where N = total number of pixels in all classes
j = total number of classes
x_{ii} = pixels on diagonal of confusion matrix
x_{+I} = summation of all rows on column i
x_{i+} = summation of all columns on row i

V. Results And Discussion

A. Visual Assessment of Classified Image

Fig. 2(a) represents the original image of the study area. Fig. 2(b) shows the classification results of parallelepiped algorithm. Fig. 2(c) shows the results of maximum likelihood algorithm. Fig. 2(d) lists the mapping of color scheme to the land cover classes. Each distinct color represents a separate class. However black color represents the unclassified pixels, pixels which did not belong to any class with respect to its spectral signature.

B. Classification Results

Confusion matrix is used to determine the accuracy of classifiers applied. Results from the confusion matrix are given in the table II. These results list the overall accuracy, Kappa statistics, user accuracy and producer accuracy of land cover classes i.e. forestry, water bodies, shrubs and bushes, settlements, and barren lands.

Among the two classification algorithms Parallelepiped classified the image to an overall accuracy of about 95.4% with kappa coefficient having a value of 0.937, whereas Mahalanobis Distance Classification algorithm showed an overall accuracy of 85.9% with the value of kappa coefficient as 0.8115. The post-processing methods mentioned before show a positive improvement in the overall accuracy of the two classification algorithms when applied both individually as well as in combination. They show quite an improvement in the statistical results. Post-processing statistics are also listed in the Table II.

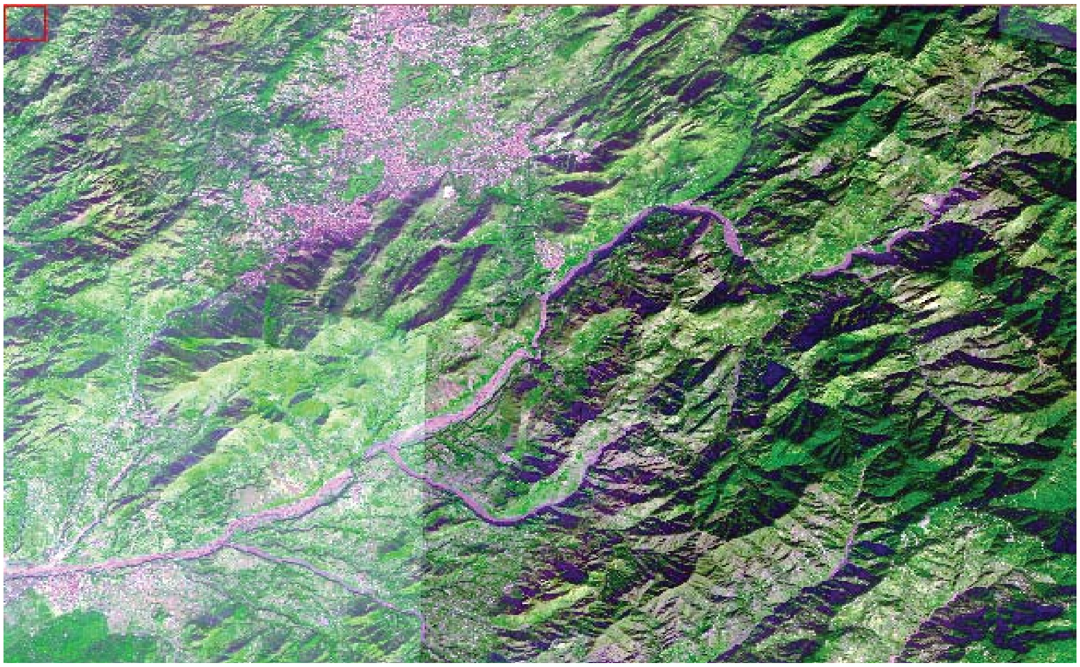

Fig. 2(a). Original Image from SPOT-5

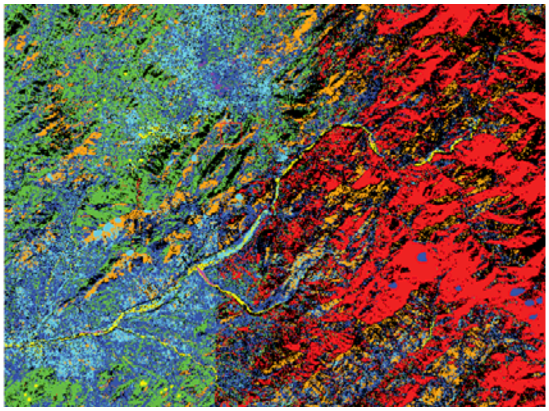

Fig. 2(b). Classified Image – Parallelepiped

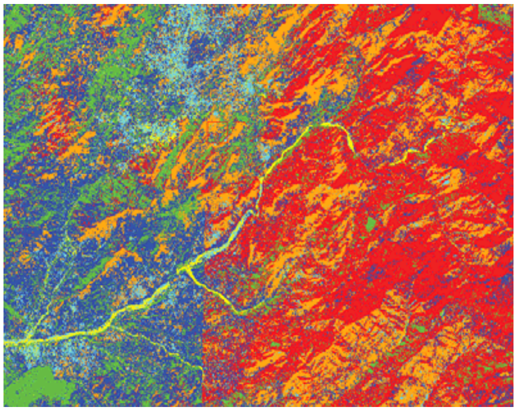

Fig. 2(c). Classified Image – Mahalanobis Distance

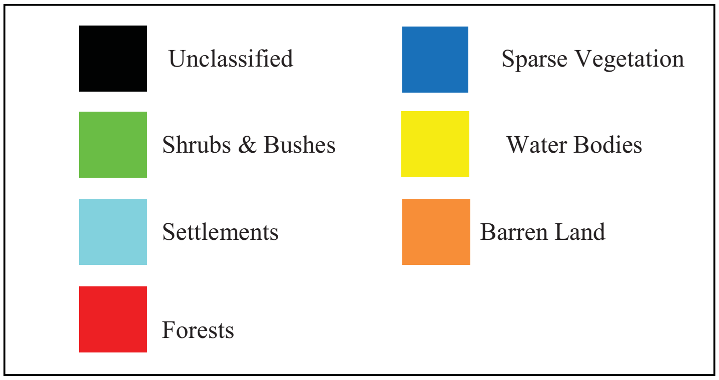

Fig. 2(d). Color Class Mapping Chart

TABLE II. Classifiers results from confusion matrix before and after post-processing

		No Post Processing	Majority Filter	Median Filter	Majority + Sieve Clump Filter
Parallelepiped Classification	Overall Accuracy	95.4%	95.64%	96.02%	96%
	Kappa Coefficient	0.937	0.94	0.945	0.945
	Shrubs &Bushes PA, UA (%)	98.92, 84.85	99, 84.97	99.25, 85.27	99.07, 85.55
	Sparse Vegetation PA, UA (%)	76.31, 98.14	76.55, 98.46	77.27, 98.77	77.67, 98.61
	Settlements PA, UA (%)	98.37, 75.65	98.73, 76.87	99.23, 78.49	99.12, 78.72
	Forests PA, UA (%)	99.35, 100	99.41, 100	99.55, 100	99.50, 100
	Water Bodies PA, UA (%)	95.31, 99.94	95.43, 99.94	95.91, 99.94	95.74, 99.94
	Barren Lands PA, UA (%)	94.34, 99.99	94.74, 99.99	95.32, 100	95.30, 99.99
Mahalanobis Distance	Overall Accuracy	85.97%	87.57%	88.84%	89.03%
	Kappa Coefficient	0.8115	0.8324	0.89490	0.8513
	Shrubs &Bushes PA, UA (%)	70.28,61.55	70.50, 62.53	70.95, 64.34	70.03, 64.52
	Sparse Vegetation PA, UA (%)	78.43,46.84	79.90, 51.18	81.15, 54.59	80.72, 55.84
	Settlements PA, UA (%)	94.03, 89.24	96.54, 91.27	97.70, 92.39	96.35, 92.93
	Forests PA, UA (%)	83.72, 98.35	85.84, 98.57	87.83, 98.76	88.44, 98.52
	Water Bodies PA, UA (%)	92.87, 94.29	94.97, 96.65	96.45, 97.70	96.70, 96.43
	Barren Lands PA,UA (%)	93.07, 98.36	94.37, 98.61	95.08, 98.81	95.53, 98.54

VI. Conclusion And Future Work

Two simple pixel based classification algorithms are used and compared for classifying forests in Pakistan. Classification is performed on a very high geometric resolution image of SPOT-5(2.5m). The study area imagery is provided by SUPARCO Pakistan. Classification results show an overall accuracy of 95.4% of the Parallelepiped algorithm whereas the Mahalanobis Distance classification showed an overall accuracy of 85%.

Post-processing filters when applied individually as well as in combination showed positive improvement in the accuracy of the classification results. In the imagery of 10 Oct 2014, Parallelepiped classification algorithm is compared with the Mahalanobis Distance Classifier in terms of overall accuracy, kappa statistics, user assessment and product assessment. Results are also compared after and before post-processing techniques. It is concluded from the results and statistics that the Parallelepiped algorithm showed an overall better accuracy as compared to Mahalanobis Distance in the classification of land cover classes. In future our emphasis will to compare the Parallelepiped algorithm and Mahalanobis Distance with other supervised classification algorithms in terms of accuracy and computational complexities. We will also analyze the classification results of pixel based versus object based classification along with their comparison. We will further explore other classifiers and post-processing techniques to improve the classification of forestry.

References

[1] B. Shahbaz, T. Ali, and A. Suleri, “A Critical Analysis of Forest Policies of Pakistan: Implications for Sustainable Livelihoods,” *Mitig. Adapt. Strateg. Glob. Chang.*, vol. 12, no. 4, pp. 441–453, 2007.

[2] J. Miettinen, H.-J. Stibig, and F. Achard, “Remote sensing of forest degradation in Southeast Asia—Aiming for a regional view through 5–30 m satellite data,” *Glob. Ecol. Conserv.*, vol. 2, pp. 24–36, Aug. 2014.

[3] D. C. Duro, S. E. Franklin, and M. G. Dubé, “A comparison of pixel-based and object-based image analysis with selected machine learning algorithms for the classification of agricultural landscapes using SPOT-5 HRG imagery,” *Remote Sens. Environ.*, vol. 118, pp. 259–272, 2012.

[4] “SPOT-5 Satellite Sensor.” [Online]. Available: http://www.satimagingcorp.com/satellite-sensors/other-satellite-sensors/spot-5/. [Accessed: 29-Aug-2015].

[5] P. S. Crowther and R. J. Cox, “A method for optimal division of data sets for use in neural networks,” *Knowledge-Based Intell. Inf. Eng. Syst.*, vol. 20, pp. 1–7, 2005.

[6] A. Ghorbani, A. M. Mossivand, and A. E. Ouri, “Utility of the Normalised Difference Vegetation Index (NDVI) for land / canopy cover mapping in Khalkhal County (Iran),” *Ann. Biol. Res.*, vol. 3, no. 12, pp. 5494–5503, 2012.

[7] V. N. Mishra, P. Kumar, D. K. Gupta, and R. Prasad, “Classification of various land features using RISAT-1 dual polarimetric data,” *ISPRS - Int. Arch. Photogramm. Remote Sens. Spat. Inf. Sci.*, vol. XL–8, no. December, pp. 833–837, 2014.

[8] T. Rao and T. V. Rajinikanth, “Supervised Classification of Remote Sensed data Using Support Vector Machine,” *Glob. J. Comput. Sci. Technol. C Softw. Data Eng.*, vol. 14, no. 1, pp. 71–76, 2014.

[9] A. Ahmed, M. Muaz, M. Ali, M. Yasir, S. Ullah, N. Minallah, and S. Khan, “Mahalanobis Distance and Maximum Likelihood Based Classification for Identifying Tobacco in Pakistan,” in *7th International Conference on Recent Advances in Space Technologies, IEEE RAST*, 2015.

[10] Padmini.D, “Different Levels of Image Fusion Techniques in Remote Sensing Applications and Image Classification,” *Int. J. Adv. Res. Comput. Sci. Softw. Eng.*, vol. 4, pp. 199–206, 2014.

[11] K. Søreide, “Receiver-operating characteristic curve analysis in diagnostic, prognostic and predictive biomarker research.,” *J. Clin. Pathol.*, vol. 62, no. 1, pp. 1–5, 2009.

[12] A. J. Viera and J. M. Garrett, “Understanding interobserver agreement: The kappa statistic,” *Fam. Med.*, vol. 37, no. 5, pp. 360–363, 2005.