K-Means Clustering Algorithm with Color-based Thresholding for Satellite Images

Gurudatta V Nayak
Information Technology
K J Somaiya College of Engineering, Vidyavihar
Mumbai, India

Anuja A Rao
Information Technology
K J Somaiya College of Engineering,Vidyavihar
Mumbai,India

NandanaPrabhu
Faculty, Information Technology
K J Somaiya College of Engineering,Vidyavihar
Mumbai,India

ABSTRACT
Land cover classification is an essential input to environmental and land use planning .Clustering is a technique used for land cover classification. Clustering is the assignment of objects into groups called clusters so that objects from the same cluster are more similar to each other than objects from different clusters. The proposed work presents an algorithm Using K means clustering algorithm with color based thresholding for classification of a satellite image. It is observed that this method gives better accuracy as compared using only K means clustering algorithm. The image quality metrics used are overall accuracy, user's accuracy, producer accuracy, average accuracy for user and producer.

Keywords
Image segmentation; Remote sensing; K-means; K-means with color based thresholding, Confusion Matrix.

1. INTRODUCTION
Land cover refers to the features of the land surface. These can be natural or man-made. The main reason for producing land cover maps is to give us idea of what all natural and built resources exist [1]. Satellite images are very powerful sources of maps as digital images.

Partitioning of an image into several constituent components is called segmentation. Image segmentation [6] has been subject of considerable research activity over the last three decades. In remote sensing, the process of image segmentation is defined as: "the search for homogenous regions in an image [3] and later the classification of these regions". It also means the division of an image [8] into meaningful regions based on homogeneity or heterogeneity criteria (Haralick et al; 1992). Many unsupervised algorithm have been developed for segmenting the gray scale images. In computer vision literature, various methods dealing with the segmentation and feature extraction are discussed, such as split and merge[8], region based techniques and so on. However, because of the wide range and composite nature of images, robust and efficient segmentation algorithm on coloring images is still a very challenging task and fully automatic segmentation are far from satisfying in practical situations.

Anil [2] proposed the segmentation method called Color – based K-means clustering,by first enhancing color separation of satellite image using decorrelation stretching then the grouping the regions a set of five classes using K-means clustering algorithm. Quanhua Zhao [4]the algorithm based on Levenberg-Marquardt (L-M) is used to improve the BP neutral network and then be applied in recognition of land cover with RS image. Ashok [7] in his paper deals with the performance study of Median and Wienerfor de noising. R. Weih.[9]developed the methodology of object-based classification using Feature Analyst, Imagine, and ArcGISsoftware.

This paper uses high pass filtering for preprocessing, explains the clustering algorithm, and how its efficiency can be improved by using it with the color-based thresholding. The rest of the paper is organized as follows. Section 2 introduces to Image Classification. In section 3 k-means clustering algorithm is discussed. Section 4 discusses the proposed algorithm. Section 5 Results of both the algorithm and their comparisons. Finally, the conclusion is given in Section 6.

2. IMAGE CLASSIFICATION
To classify features in an image by using the elements of visual interpretation, an analystidentifies the homogeneous regions that represent various features or land cover classes of interest.As a result an image is obtained having an array of pixels,each of which belongs to a particular theme.Image classification can be performed using two different approaches :Supervised and unsupervised.

In supervised classification, the first step is to specify the information class on the image.To do so, the analyst categorises the ,homogeneous representative samples of different features of interest from the surface referred to as signatures. An algorithm or program is then used to eventually map the remainder of the image.For this each pixel in the image is compared with these signatures and labelled as the class it closely resembles digitally. Here first information classes is recognised based on which the spectral classes that represents them is determined.

In unsupervised classification[3],the algorithm automatically groups the pixels with similar spectral characteristics such as mean,Standard deviation etc, into unique clusters based on some statistically determined criteria.The analyst then relabels and combines the spectral clusters into information classes.

There are several strategies to represent the clusters of data in spectral space.Some popular methods[5] are one pass approach,Sequencintial clustering,Statistical Clustering, K-means Clustering Iterative Self Organising Data Analysis Techniques Clustering, RGB Clustering.

3. K-MEANS CLUSTERINGALGORITHM

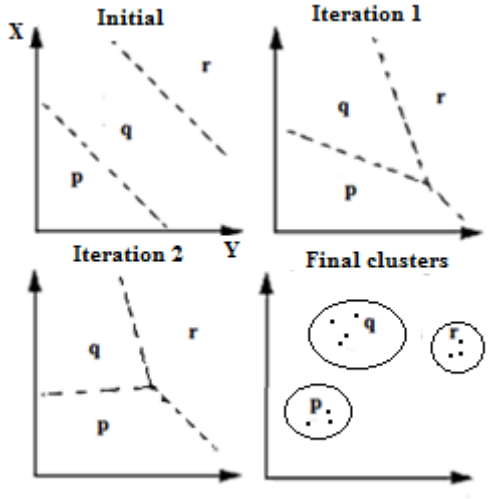

Figure 1. The cluster formation sequence

The efficiency of clustering is determined by the Clustering algorithms[2] used for unsupervised classification of remote sensing data .One of the clustering algorithms used to determine the natural spectral groupings present in a data set is K-means.

First the desired number of clusters to be located in the data is accepted from the analyst. The algorithm is typically initiated with arbitrarily located seeds for cluster means. Each pixel in the image is then assigned to the closestcluster meanThe revised mean vector are than computed for each cluster. The iterations on these two operations are continued until some criteria is met, such as cluster means do not change between iterations [10].

As K-means approach is repetitive, it is computationally expensive. However, it is a suitable for unsupervised training areas. The typical cluster formation sequence is shown in Figure 1

4. K-MEANS ALGORITHM WITH COLOR BASED THRESHOLDING

The efficiency of the pure color-based K-means algorithm can be improved by combining the algorithm by color-based thresholding. The proposed algorithm is carried out in following steps:

1. The very first step is extracting our color bands from the original image into separate 2D arrays, one for each component (Red, Blue and Green).
2. The next step is to compute red, green and blue histogram. Then all axes are set to be of same height and width (averaging of histogram), this makes them easy to compare them.
3. Decide the low and high thresholds for each color band. Choose a value which may suit the image (trial and error method is used to compute the best possible output).
4. Apply these thresholds on their respective color band. Then ANDing the masks to find where all 3 are "true".
5. Then we will have the mask of only those parts of image whose threshold has been applied.
6. It may happen that any one band mask may become 0, thus if it becomes all 0's then set them to 1. Otherwise the entire image will be ANDed with zero and output will be black image.
7. Now use this object mask to mask out the input image. Again concatenate the masked color bands to RGB image. Now here we get the extracted object.
8. But since we don't have exact red objects in image, the red mask can't be applied directly so to obtain land portion of image, the original image is subtracted from the blue and green subparts obtained as result of thresholding operation. The image thus obtained is pure land image. Now these results are given to clustering algorithm.
9. There in water bodies accuracy is perfect so that cluster is untouched, so display it as it as.
10. The other two parts of image viz. forest and land bodies are taken from clustering and displayed as output.

5. ACCURACYASSESSMENT

It determines the correctness [3] of the classified image. In digital Image processing, the degree to which information in classified image matches to information in original image is termed as accuracy. The standard form for reporting this is confusion matrix or error matrix or contingency table. It is a

n x n matrix, where n is the number of classes. Here, the columns represent the actual classes of information on the input map, rows represent the classes as identified on the classified image and the diagonal elements indicate the number of samples for which the classification results matches with the reference data.

Using these information two accuracies can be determined

1. Users accuracy: Is the ratio of the number of correctly classified feature and its row total.

2. Producers accuracy: Is the ratio of the number of correctly classified feature and its corresponding row total

The most commonly used accuracy is the overall accuracy

$$W = \frac{\sum_{i=1}^{c} Di}{N} * 100$$

Where Di is the diagonal element

N is the total number of samples

W is the overall accuracy in percentage

6. EXPERIMENTAL RESULTS

The input image is preprocessed using high pass filter for noise removal , followed by Color sharpening which uses a predefined mask that enhances the color intensity of each pixel so that the color separations are distinct and clear. Fig.2shows input image and image after preprocessing.

Fig2: (a) original input image (b) Output obtained from preprocessing of input image

The K means Clustering algorithm is applied and Classification obtained is shown in Fig.3 Output obtained with proposed algorithm is shown in Fig.4

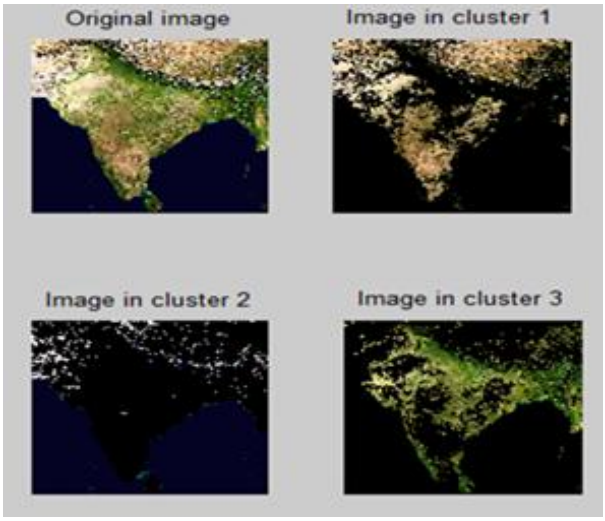

Fig3: Output image obtained from pure K-means clustering algorithm.

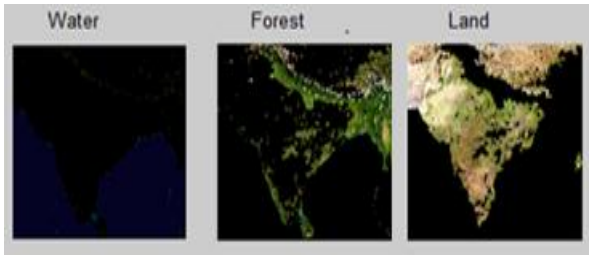

Fig.4: Output obtained from proposed algorithm.

The observation matrix also called as confusion matrix obtained in shown in Table 1 and Table 2 respectively

Table1. Confusion Matrix for K-means Clustering

	Water	Forest	Land	Total
Water	366486	0	0	366486
Forest	0	202456	35338	237794
Land	0	36438	240858	277296
Total	366486	238895	276196	881576

Table2. Confusion Matrix for K-Means clustering with color-based thresholding

	Water	Forest	Land	Total
Water	361963	3 2 5	0	362288
Forest	0	351320	10030	361350
Land	0	15135	142803	157938
Total	361963	366780	152833	881576

The algorithms are evaluated by using image quality metrics like overall accuracy, user's accuracy, producer's accuracy, average accuracy of user and producer. The proposed algorithm is compared with color based K-Means clustering algorithm using the parameters as shown in Table 3

Table 3: Comparison of parameters

Parameters (Accuracy)		**K-Means Clustering**	**Proposed Algorithm**
Overall Accuracy		91.85	97.10
User's Accuracy	Water	100	100
	Forest	84.74	95.78
	Land	87.20	93.43
Producer's Accuracy	Water	100	99.9
	Forest	85.13	97.22
	Land	86.85	90.41
Average	User	90.64	96.40
	Producer	90.66	95.84

7. CONCLUSION

K –means clustering algorithm being an unsupervised classification approach is suitable when classes are to be determined by spectral distinctions that are inherent in the data .It is observed that by using color based thresholding both the user accuracy and producers accuracy and hence overall accuracy is enhanced.

8. REFERENCES

[1] Quanfang Wang, Haiwen Zhang, Hangzhou Sun, "New Logic for Large-scale Land Cover Classification Based on Remote Sensing", Geoinformatics17th international conference,2009,ISBN 978-1-4244-4562-2,Pg 1-5

[2] Anil Z Chitadeand Dr. S K. Katiyar,"Color based image segmentation using K-Means Clustering", Department of civil Engineering, International Journal of Engineering Science and Technology, Vol. 2(10), 2010, 5319-5325

[3] A. M. Chandra and S. K . Ghosh "Remote Sensing and Geographical Information systems", Narosa publishing House, New Delhi ,2007

[4] Quanhua Zhao, Weidong Song, Guohua Sun, " The Recognition of Land Cover with Remote Sensing Image Based on Improved BP Neutral Network",Intyernational

Conference on Mutimedia technology,2010,ISBN 978-1-4244-787-2

[5] Basudeb Bhatta " Remote Sensing and GIS " Oxford University Press,2008

[6] Gonzales R and Woods R. "Digital /image Processing"Prentice hall, 2nd Edition

[7] Ashok Kumar Nagawat, Manoj Gupta, Papendra Kumar and Suresh Kumar," Performance Comparison of Median and Wiener Filter in Image De-noising", International Journal of Computer Applications ,12(2010) 0975-8887.

[8] Gong xuejing, ci linlin,yao kangze, 'Two parallel strategies of split-merge algorithm for image segmentation', International Conference on Wavelet Analysis and Pattern Recognition, Beijing, China, 2-4 Nov. 2007

[9] R. Weih, Jr.1 and D. White, Jr," Land-Use/Land-Cover Characterization Using an Object-Based Classifier for theBuffalo River Sub-Basin in North-Central Arkansas", Journal of the Arkansas Academy of Science, Vol. 62, 2008

[10] Lillesand and Kieffer, Remote Sensing, Wiley Edition 2000