A New K-mean Color Image Segmentation with Cosine Distance for Satellite Images

Modh Jigar S, Shah Brijesh, Shah Satish k

Abstract— This paper represents unsupervised method of k-means segmentation which is new adaptive technique of color-texture segmentation. With the progress in satellite images, the image segmentation technique for generating and updating geographical information are become more and more important. This algorithm first enhance the image then applying clustering based k-means segmentation technique ,using L*a*b color space and using cosine distance matrices instead of sqeuclidean distance. With this it is possible to reduce computational time and calculation for every pixel in the image .Although colors are not frequently used in image segmentation; it gives high discriminative power to the regions present in image.

Keywords-K-means segmentation, cosine distance ,Euclidean distance,

I. INTRODUCTION

Image segmentation is the most important part of the image processing based application and it affects the overall system performance [1-2]. For the remote sensing image, segmentation can be defined in terms of search of homogeneous regions and classification of these regions. The image segmentation technique can be differentiated in to following basic concepts: pixel oriented, model oriented, contour-oriented, region oriented, color oriented and hybrid. The most common features used in image segmentation are texture, shape, gray level intensity and color. Color oriented image segmentation technique is most important in image analysis and computer vision, pattern reorganization, and image interpretation. It can be use in many scientific and industrial applications such as medicine, remote sensing, microscopy, document analysis, quality control, industrial automation, content based video and image retrieval [3, 4]. The segmentation evaluation technique is generally divided in two categories: supervised and unsupervised. For satellite image segmentation, we cannot use supervised segmentation technique because of optimum segmentation [5]. The color homogeneity is based on standard deviation and while the shape homogeneity is based on compactness and smoothness of shape. The iteration is of parameters have two basic principle first we should choose a scale value as large as possible to distinguish different regions. In second we should use color criterion where it possible [6]. The work represented her is based on color image segmentation .We have selected L*a*b color space which is perceptually uniform orthogonal Cartesian coordinate system [2, 7]. The difference between two pixel in L*a*b color space is the same in the sense of human eye visual systems and this color space enables people to quantify these visual difference. The k-mean image segmentation technique is clustering based segmentation technique in which we have to find centroid of each cluster. In this paper we used Cosine distance instead of Sqeuclidean distance to find the centroid of cluster. This will reduce the computation time of implementation of algorithm. The segmentation of satellite images with this algorithm will get better result than standard algorithm.

In this paper, section II is based on k-means segmentation. Section III gives the overview of two different type distance matrices .The difference between sqeuclidean distance and cosine distance is represents here. Section IV shows the steps of this algorithm. The experiment and results are carried out in section V and finally section VI concludes the algorithm.

II. K-MEANS SEGMENTATION

K-means [8] is one of the simplest unsupervised learning algorithms that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters fixed priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in cunning way because of different location cause different result. So the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging to a given data set and associate it to the nearest centroid .When no point is pending, the first step is completed and an early group is done. At this point we need to re- calculate k new centroids of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set pints and the nearest new centroid .A loop has been generated .as a result of this loop we may notice that the k centroids change their location step by step until no more changes are done .In other word centroids do not have anymore .The mathematical equitation of k-means segmentation is,

$$J_{K-means} = \sum_{k=1}^{K} \sum_{j \in S_k} d^2(x_j, c_k)$$

Where,
k is the number of cluster evaluated (in a space defined by S_k)
x_j is the pattern j evaluated in a relation to the centroid
$d^2(x_j,c_k)$ is the distance between pattern x_j and centroid c_k
The algorithm contains following steps:

a. Place K points into the space represented by the object that are being clustered .these points represent group centroids.

Manuscript Received on June 22, 2012.

Modh Jigar S, PG student, V.T. Patel Department of Electronics and Communication Engineering,
Shah Brijesh, Associate Professor, Associate Professor, C S. Patel Institute of Technology, CHARUSAT, Changa, Gujarat, India
Shah Satish k., Professor M. S. University, Baroda, Gujarat, India

b. Assign each object to the group that has the closest centroid.
c. When all objects have been assigned, recalculate the positions of the K centroids.
d. Repeat steps 2 and 3 until the centroids no longer move .this produce a separation of the objects in to groups from which the metric to be minimized can be calculated.

III. COSINE AND SQEUCLIDEAN MATRICES

The distance or the similarity should be defined before clustering. It will measure the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clustering embedded in data. The selection of an appropriate distance measure is also crucial for cluster analysis in the k-means segmentation [9].The distances or similarity can be based on a single dimension or multiple dimensions, with each dimension representing a rule or condition for grouping object. The most straightforward way of computing distance between objects in multiple spaces is to compute Euclidean distances. To qualify a distance metrics a measure d must satisfy following conditions.
Let x and y be any object in a set and *d(x, y)* be the distance between x and y.

a. The distance between any two points must be non-negative, that is *d(x, y)>0.*
b. The distance between two objects must be zero if and only if the two objects are identical, that is, *d(x, y) =0* if and only if *x=y.*
c. Distance must be symmetric ,that is distance from x to y is the same as the distance from y to x, i.e. $d(x,y) = d(y,x)$
d. The measure must satisfy the triangle inequality that is $d(x,z) \le d(x,y) + d(y,z)$

A. Euclidean distance metric:

This is probably the most commonly chosen type of distance. It is simply geometric distance in the multidimensional space. It is computed as:

$$d_{(x,y)} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

(1)

Euclidean distance is widely used in clustering problems .The Euclidean distance satisfy all above four conditions so that it is a true metric. It is also default distance used in k-means algorithm. Measuring the distance between two centroids *x* and *y* can be define from equation 1.

B. Squared Euclidean distance metric:

The Squared Euclidean distance metric uses the same equation as the Euclidean distance metric, but does not take the square root. As a result clustering with the Euclidean squared distance metric is faster than clustering with the regular Euclidean distance.

$$d_{(x,y)} = \sum_{i=1}^{n} (x_i - y_i) \qquad (2)$$

C. Cosine distance metric:

The cosine similarity as quantified the cosine angle between vectors. Cosine similarity is one of the most popular similarity measure applied to text documents, such as in information retrieval applications [10] and clustering too [11]. The cosine of two vectors can be easily derived by using the Euclidean dot product formula. Here given two vectors attributes, *X* and *Y*, the cosine similarity θ is represented using a dot product and magnitude as

$$similarity = \cos(\theta) = \frac{X \cdot Y}{\|X\|\|Y\|}$$

The resulting similarity ranges from 0 to 1 where 1 is the exactly same and 0 usually indicating independence and in between values indicating intermediate similarity or dissimilarity.

$$d(x,y) = 1 - \left(\frac{\cos^{-1}(similarity)}{\pi} \right)$$

The cosine similarity has been used for the angular distance, the advantage of angular similarity coefficient is that, when used as a difference coefficient it means by subtracting it from 1, the resulting function is proper distance metric [12].

IV. MODIFIED ALGORITHM

As we have seen two different type of distance in nearest neighbor search of K-means algorithm. The K-means algorithm has default distance is Euclidean distance, but in this proposed method we apply the cosine similarity or distance. The advantages and comparison we will see in next section. Here first we change RGB color space in to the L*a*b color space [7].the basic steps of the algorithm are as follows:

Step 1: Read the image
Read the satellite image which is to be segmented .the image should be .JPG or any other format.

Step 2: Convert the image from RGB color space to L*a*b color space
If we ignore the brightness in image than there are three colors we can see: white, blue and pink. The L*a*b color space (also known as CIELAB or CIE L*a*b) enables us to quantify these visual difference [13]. The L*a*b color space is derived from the CIE XYZ tristimulus values. The L*a*b space consist of luminosity layer ‘L*’, chromaticity ‘a*’indicating where color falls along the red-green axis, and chromaticity ‘b*’ indicating where the color falls along the blue-yellow axis. All of the color information is in the ‘a*’ and ‘b*’ layers. We can measure the difference between two colors using cosine distance. Convert the image in to L*a*b color space.

Step 3: Classify the colors in *a*b space using K-means Clustering

Clustering is the way to separate group of objects. K-means clustering treats each object having a location in space. It finds partition such that objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. K-means clustering requires that you specify the number of clusters to be partitioned and a distance metric to quantify how close two objects are to each other. Here instead of using Euclidean distance we use Cosine angle distance.

Step 4: Label every pixel in the image using the result the from K-means.

For every object in our input k-means returns an index corresponding to a cluster. Label every pixel in the image with its cluster index.

Step 5: Create images that segment the images by color.

Using pixel labels, we have to separate objects in image by color, which will result in five images.

V. EXEPERIMENT RESULTS

The experiments on the satellite images have been carried out in MATLAB v7.10.K-means segmentation is a clustering based segmentation algorithm. In clustering based segmentation changing in the distance metric will change the output. Euclidean distance is the default distance used in the algorithm, replacing it with the cosine distance gives better segmented areas in the satellite images. In the fig.2 we can see the original satellite images. Fig 3 shows the cluster index images by cosine distance metric of fig.2. Now compare it with fig 4, which are cluster index images by Sqeuclidean distance. We can see that segmentation of areas are good in fig .3 than in fig 4.An advantage of using cosine distance instead of sqeuclidean distance is reduce computational time. The time consumption difference of both the algorithm is shown in table 2, which shows the average time duration of 4 images implemented on both algorithm. Table 1 shows the average time duration of 4 satellite images. The main application of this algorithm is shown in fig 5, where different segmented areas are clustered .In satellite image there will be different type of objects like land, sea, desert, clouds, and mountains. Using this algorithm all the areas are clustered in different images. Here 5 clusters are used to differentiate all areas in image.Fig.5.1 is the original satellite image. Fig. 5.2 shows the clouds in cluster 1 while fig 5.3 and 5.5 shows the land areas in cluster 2 and 4.Where area shown in fig 5.5 is more green land area. Finally area of water shown in fig 5.4 and 5.6

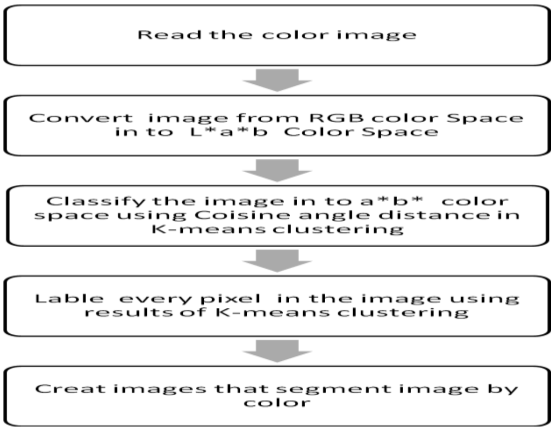

Fig.1 Algorithm.

Table-1

Image	Image Size	CPU Time-1 (sec.)	CPU Time-2 (sec.)	CPU Time-3 (sec.)	CPU Time-4 (sec.)	Average CPU Time (sec.)
Sat-1	1024×768	5.531	4.599	5.598	4.929	5.16
Sat-2	1912×1894	23.966	26.657	20.306	17.811	22.18
Sat-3	530×375	1.968	1.672	1.726	1.722	1.77
Sat-4	1024×768	4.338	7.066	6.663	3.364	5.33


Fig.2.1 Satellite Image-1


Fig. 2.2 Satellite Image-2


Fig. 2.3 Satellite Image-3

Fig. 2 Original Images

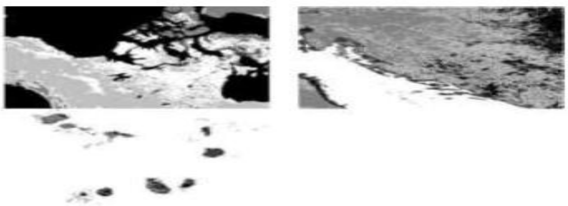
Fig. 3 Images labeled by clustered index using cosine distance

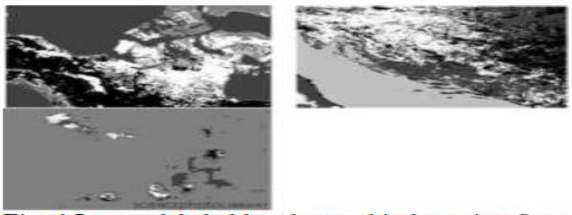
Fig. 4 Images labeled by clustered index using Sqeuclidean distance

Table-2

Image	Image Size	Average CPU Time With Sqeuclidean distance based Algorithm (Sec.)	Average CPU Time With Cosine distance based Algorithm (Sec.)	Accuracy (%)
Sat-1	1024×768	6.32	5.16	18.44%
Sat-2	1912×1894	28.95	22.18	23.41 %
Sat-3	530×375	2.24	1.77	21.29 %
Sat-4	1024×768	7.27	5.33	26.74%

Fig. 5.1 Original image

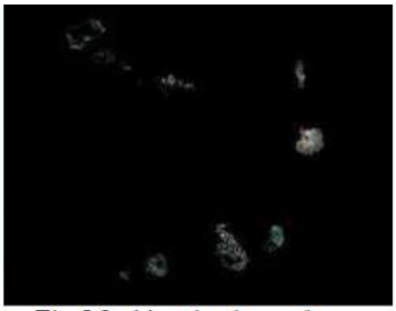
Fig 5.2 object in cluster 1

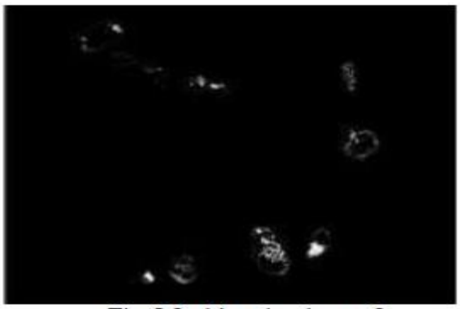
Fig 5.3 object in cluster 2


Fig 5.4 object in cluster 3

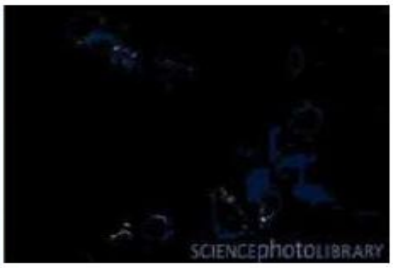

Fig 5.5 object in cluster 4

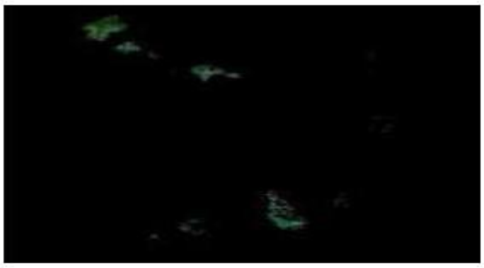
Fig 5.6 object in cluster 5

Fig.5. Images of all clusters

VI. CONCLUSION

Using the cosine distance based image segmentation, it is possible to reduce computational time and cost. The color based segmentation of satellite images give a high discriminative power of regions present in image. This type of segmentation can be used to differentiate areas in satellite images and it can be used in many geographical application based software. However color based K-means segmentation can be used with other segmentation algorithm as like marker controlled watershed algorithm for further improvement.

REFERENCES

1. He, Xiaoling; Hodgson, W. Jeffrey,”*Research Image Processing Technology hot issue*”, IEEE Transactions on Intelligent Transportation Systems, **3**(4)**:** Dec., 244-251, 2002.
2. Amanpreet Kaur, et al, “Color image Segmentation Using K-Means Clustering Algorithm” ,International journal on emerging technologies 1(2):18-20,2010.
3. Ricardo Dutra da Silva et al ”Satellite Image Segmentation Using Wavelet transform Based On color and Texture Features”, ISVC 2008,part II,LNCS 5359,pp 113-122,2008
4. H C Chen et al,”Visible Color Difference-Based Quantitative of Color Segmentation “, IEEE proceeding, Vis image signal Process Vol.153 No. 5 , pp 598-609.oct 2006.
5. Ahmed Darwish ,et al, ”Image Segmentation for the Purpose of Object-Based Classification”,IEEE, pp 2039-2041,2003
6. Anil Z Chitade, “Color Based Image Segmentation using K-mean Clustering”, International journal of Engineering Science and technology, Vol. 2(10), 5319-5325,2010.
7. W.Woelker,”Image Segmentation Based on an Adaptive 3D Analysis of thecae-L*a*b* Color Space.” proc. of SPIE’96-Visual Communication and Image Processing ’96, Vol.2727, pp.1197-1203, 1996.
8. J. B. Macqueen, “Some Methods for Classification and Analysis of Multivariate Observations”, proceedings of 5-th Berkeley symposium on mathematical statistics and probability, Berkeley, university of California press, 1:281-297,1967.
9. Anna Huang, ”Similarity Measures for Text Document Clustering” The university of waikato, Hamilton ,
10. R. B. Yates and B. R. Neto.”Modern Information Retrieval.”, Addison-Wesley , newyork,1999.
11. A. K. Jain, M. N. Murty, and P. J. Flynn.”Data clustering: a review.”ACM computing surveys (CSUR) , 31 (3) :264-323, 1999.
12. P. N. Tan, M. Stenbach & V. kumar, ”Introduction to Data Mining” ,Addison-Wesely,ch.-8,2005.
13. L.jin and D.Li”A switching Vector median Based on the CIELAB color space for Color Image Restoration”, Signal Processing, Vol. 87, pp 1345, 2007.