Analysis of K-means Algorithm

Vivek Barwat[#1], Prof. J. M. Bhattad[*2]

[#]*Electronics & Telecommunication, R.T.M. Nagpur university*
P.C.E, Nagpur ,India
[1]vivekbarwat@yahoo.in

[*]*Second Company*
P.C.E, Nagpur ,India
[2]shivedibhattad@rediffmail.com

Abstract— **Clustering performance of the K-means greatly relies upon the correctness of the initial centroids. Usually the initial centroids for the K-means clustering are determined randomly so that the determined centroids may reach the nearest local minima, not the global optimum. This paper proposes a new approach to optimizing the designation of initial centroids for K-means clustering. This approach is inspired by the thought process of determining a set of pillars' locations in order to make a stable house or building. We consider the pillars' placement which should be located as far as possible from each other to withstand against the pressure distribution of a roof, as identical to the number of centroids amongst the data distribution. Therefore, our proposed approach in this paper designates positions of initial centroids by using the farthest accumulated distance between them. First, the accumulated distance metric between all data points and their grand mean is created. The first initial centroid which has maximum accumulated distance metric is selected from the data points. The next initial centroids are designated by modifying the accumulated distance metric between each data point and all previous initial centroids, and then, a data point which has the maximum distance is selected as a new initial centroid. This iterative process is needed so that all the initial centroids are designated. This approach also has a mechanism to avoid outlier data being chosen as the initial centroids. The experimental results show effectiveness of the proposed algorithm for improving the clustering results of K-means clustering.**

Keywords— **Clustering , Centriods , Segmentation ,Iteration**

I. INTRODUCTION

The main disadvantage of the k-means algorithm is that the number of clusters, *K*, must be supplied as a parameter. In this paper we present a simple validity measure based on the intra-cluster and inter-cluster distance measures which allows the number of clusters to be determined automatically. The basic procedure involves producing all the segmented images for 2 clusters up to *Kmax* clusters, where *Kmax* represents an upper limit on the number of clusters. Then our validity measure is calculated to determine which is the best clustering by finding the minimum value for our measure. The validity measure is tested for synthetic images for which the number of clusters in known, and is also implemented for natural images. Many criteria have been developed for determining cluster validity , all of which have a common goal to find the clustering which results in compact clusters which are well separated. The Davies-Bouldin index , for example, is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The objective is to minimize this measure as we want to minimize the within-cluster scatter and maximize the between-cluster separation. Bezdek and Pal have given a generalization of Dunn's index. Also, by considering five different measures of distance function between clusters and three different measures of cluster diameter, they obtained fifteen different values of theThe K-means algorithm generates the initial centroids randomly and fails to consider a spread out placement of them spreading within the feature space. In this case, the initial centroids may be placed so close together that some become inconsequential. Because of this, the initial centroids generated by K-means may be trapped in the local optima. We propose in this paper a method of placing the initial centroids whereby each of them has a farthest accumulated distance between them. The proposed algorithm in this paper is inspired by the thought process of determining a set of pillars' locations in order to make a stable house or building. Fig. 1 illustrates the locating of two, three, and four pillars, in order to withstand the pressure distributions of several different roof structures composed of discrete points. It is inspiring that by distributing the pillars as far as possible from each other within the pressure distribution of a roof, the pillars canAn adaptive noise removal filtering using the Wiener filter is applied for noise removal of images. The Wiener filter can be considered as one of the most fundamental noise reduction approaches and widely used for solution for image restoration problems . In our system, we use 3x3 neighborhoods of filtering size.

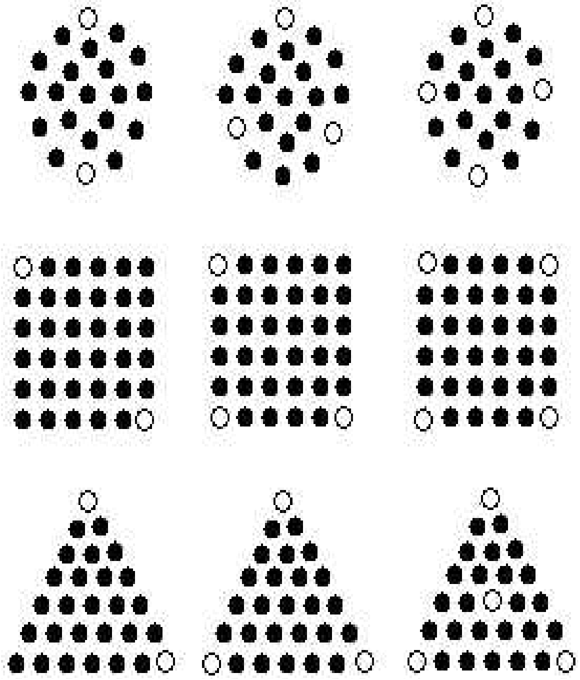


Fig. 1 Illustration of locating a set of pillars.

II . COMPUTATIONAL STEPS FOR COLOR IMAGE SEGMENTATION

Computational Steps For Color Image Segmentation is shown in fig.2.The Various steps involved are :-

A. Noise Removal

An adaptive noise removal filtering using the Wiener filter is applied for noise removal of images. The Wiener filter can be considered as one of the most fundamental noise reduction approaches and widely used for solution for image restoration problems . In our system, we use 3x3 neighborhoods of filtering size.

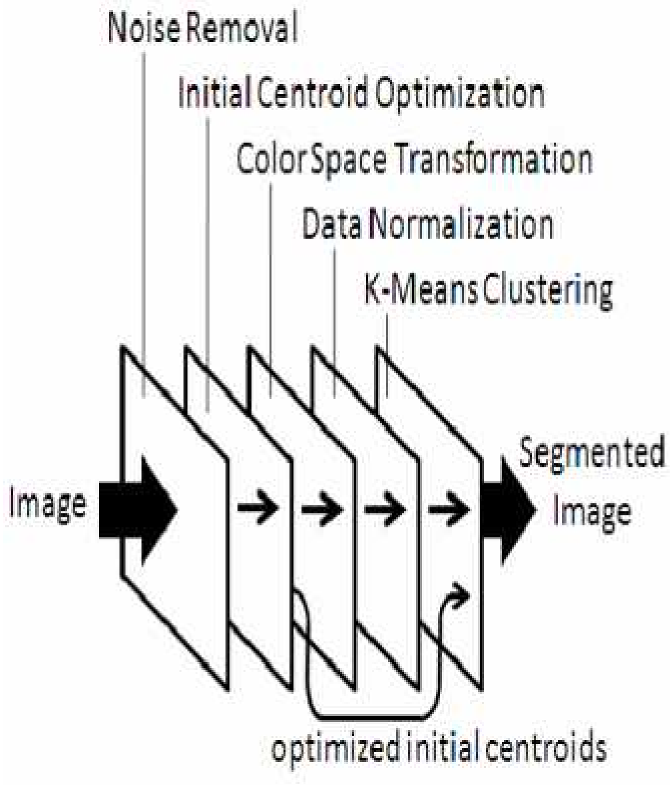

Noise Removal
Initial Centroid Optimization
Color Space Transformation
Data Normalization
K-Means Clustering
Image
Segmented Image
optimized initial centroids

Fig. 2 Computational steps of our approach for image segmentation

B. Color Space Transformation

Our image segmentation system pre-proceeds the image by transforming the color space from RGB to HSL and CIELAB color systems. HSL is well-known as an improved color space of HSV because it represents brightness much better than saturation. Beside, since the hue component in the HSL color space integrates all chromatic information, it is more powerful and successful for segmentation of color images than the primary colors . The CIELAB color system has the advantage of being approximately perceptually uniform, and it is better than the RGB color system based on the assumption of three statistically independent color attributes . The CIELAB color space is also widely-used for image restoration and segmentation [. Considering the advantages of each color system of HSL and CIELAB, in our system we utilize both of them as hybrid color systems for image segmentation. we apply clustering using the K-means algorithm and then obtain the position of final centroids. We use these final centroids as the initial centroids for the real size of the image as shown in Figure 2, and then apply the image data point clustering using K-means. This mechanism is able to improve segmentation results and make faster computation for the image segmentation.

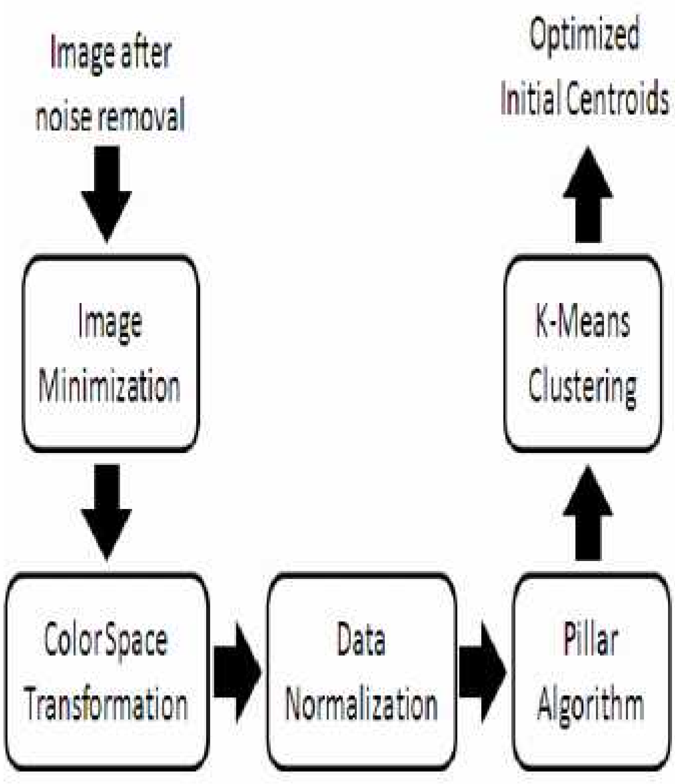


Image after noise removal
Image Minimization
Color Space Transformation
Data Normalization
Pillar Algorithm
K-Means Clustering
Optimized Initial Centroids

Fig. 3 Initial centroid optimization of K-means clustering for image segmentation.

III K-MEANS METHOD

The k-means method aims to minimize the sum of squared distances between all points and the cluster centre. This procedure consists of the following steps, as described by Tou and Gonzalez.

1. Choose K initial cluster centres z1(1), z2(1),,..., zK(1) .

2. At the k-th iterative step, distribute the samples {x} among the K clusters using the relation, for all i = 1, 2, …, K; i ¹ j; where Cj(k) denotes the set of samples whose cluster centre is zj(k). x Cj (k) if x z j (k) x zi (k) for all i = 1, 2, …, K; i ¹ j; where Cj(k) denotes the set of samples whose cluster centre is zj(k).

3. Compute the new cluster centres zj(k+1), j =1, 2, …, K such that the sum of the squared distances from all points in Cj(k) to the new cluster centre is minimized. The measure which minimizes this is simply the sample mean of Cj(k). Therefore, the new cluster centre is given by

$$Zj(k+1)= 1/N_j \sum_{x€cj(k)} x, j=1,2\ldots..k$$

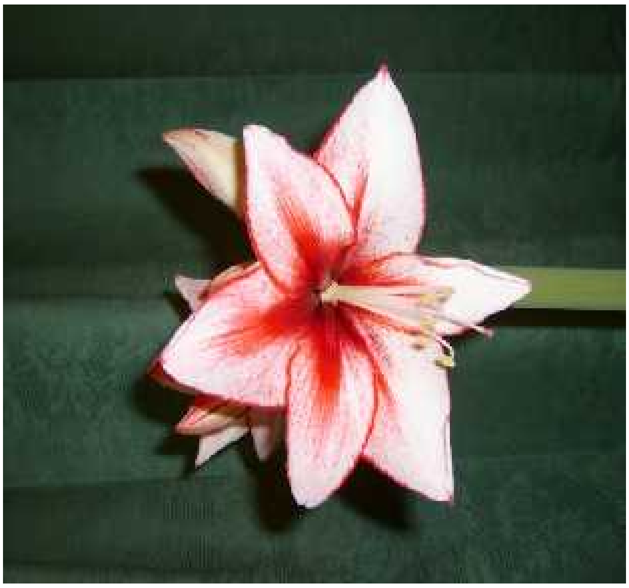
Where *Nj* is the number of samples in *Cj(k).*

4. If *zj(k+1) = zj(k)* for *j = 1, 2, …, K* then the algorithm has converged and the procedure is terminated.

IV. EXPERIMENTAL RESULTS

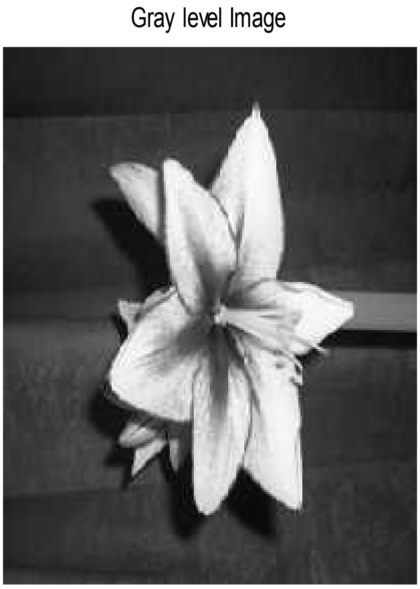
We developed the code for k-means algorithm in MATLAB. We used MATLAB R 2008a.

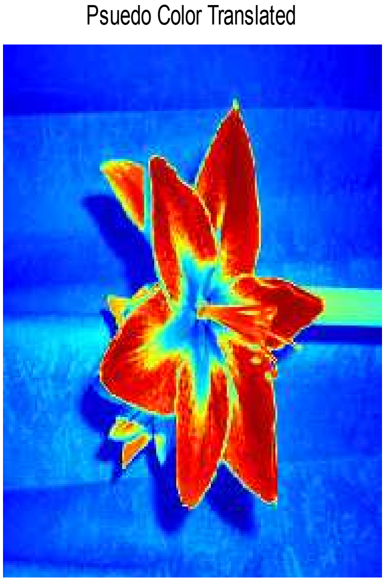
INPUT IMAGE:-



OUTPUT IMAGES:-

a) Gray Level Image





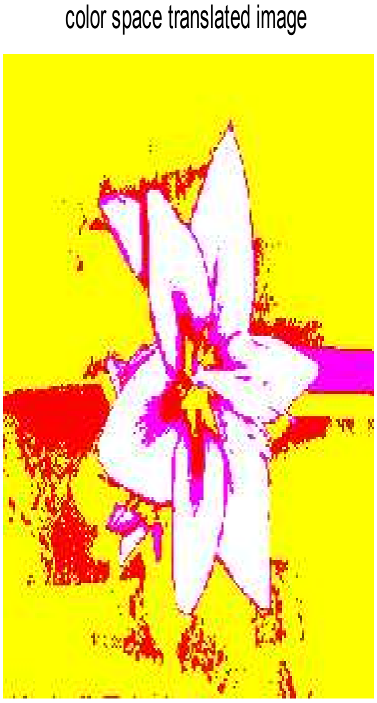
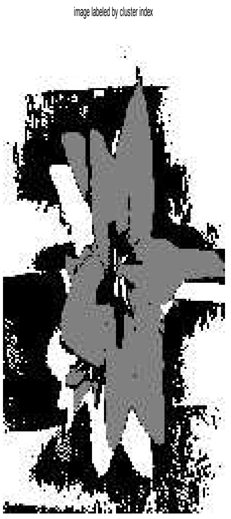
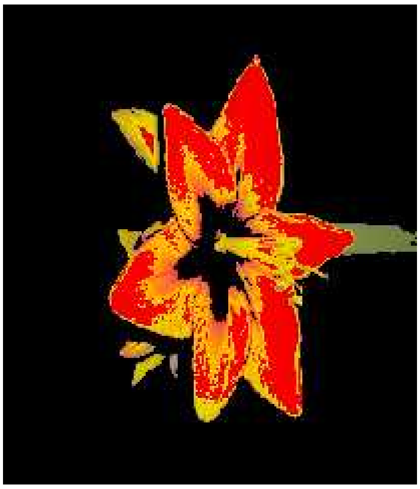


Image Labeled By Cluster Index



Truly segmented image using K-means



V. CONCLUSION

By incorporating the validity measure based on the intra-cluster and inter-cluster distance measures, the number of clusters present in an image can be determined automatically. The validity measure roposed here works well with synthetic images, producing a minimum value for the expected number of clusters. Although there is a tendency to select smaller cluster numbers for natural images, this is due to the inter-cluster distance being much greater and greatly affecting the validity measure. We overcome this by looking for a local maximum for the validity measure and then by finding the minimum value after the local maximum. By using this modified rule, the smallest number of clusters that can be selected is four. This is not real problem because natural colour images can be expected to have more than two or three clusters. The modified rule still allows the optimal number of clusters to be selected for the synthetic images, except for the image with only two clusters as two clusters cannot be selected by this modified rule. The Davies-Bouldin index and Dunn's indexes could not detect the correct number of clusters detect the correct number of clusters for all the natural images, performed more consistently for all of the natural images, producing good segmentation results. As minor modifications to this algorithm, we could use median cluster centre representation instead of mean cluster centre representation, and we could use absolute distance instead of Euclidean distance to calculate the distance between a pixel and its cluster centre or between cluster centres. We could also use any of the number of different colour spacesavailable. As an improvement we could also incorporate the context of the image, as a given pixel is expected to be highly correlated with its neighbouring pixels values. This could be achieved by taking into account the values of the neighbouring pixels. This method is not restricted to colour images. It can be easily extended to cope with any dimensionality, so this method may also be used for multispectral images. Similarly, there is no reason why this method cannot be used for grey scale images, which have only one dimension.

REFERENCES :-

[1] N.R. Pal and S.K. Pal, A review on image segmentation techniques, Pattern Recognition, vol. 26, pp. 1277-1294, 1993.

[2] K.S. Fu and J.K. Mui, A survey on image segmentation, Pattern Recognition, vol. 13, pp. 3-16, 1981.

[3] R.M. Haralick and L.G. Shapiro, Survey image Segmentation techniques, Comput. Vision Graphics Image Process., vol. 29, pp. 100-132,1985

.

[4] G.A. Growe, "Comparing Algorithms and Clustering Data: Components of The Data Mining Process," thesis, department of Computer Science and Information Systems, Grand Valley State University, 1999.

[5] J.M. Penã, J.A. Lozano, P. Larrañaga, "An empirical comparison of the initilization methods for the K-means algorithm," Pattern Recognition Lett., 20, 1027-1040, 1999.

[6] S. Ray, R.H. Turi, "Determination of number of clusters in K-means clustering and application in colthe image segmentation," Proc. 4th ICAPRDT, pp.137-143, 1999.ss

`

[7] J.L. Marroquin, F. Girosi, "Some Extensions of the K-Means Algorithm for Image Segmentation and Pattern Classification", Technical Report, MIT Artificial Intelligence Laboratory, 1993.

[8] K. Atsushi, N. Masayuki, "K-Means Algorithm Using Texture Directionality for Natural Image Segmentation", IEICE technical report. Image engineering, 97 (467), pp.17-22, 1998.

[9] A. Murli, L. D'Amore, V.D. Simone, "The Wiener Filter and Regularization Methods for Image Restoration Problems", Proc. The 10th International Conference on Image Analysis and Processing, pp.