International Journal of Advanced Technology in Engineering and Science
Vol. No.5, Issue No. 02, February 2017
www.ijates.com
ijates
ISSN 2348 - 7550

IMAGE SEGMENTATION USING K-MEANS CLUSTERING BASED THRESHOLDING ALGORITHM

Saravanan M[1], Kalaivani B[2], Geethamani R[3]

[1]*Associate Prof. ,Computer Science, K.R.College of Arts & Science, Kovilpatti, Tamilnadu, (India)*

[2&3]*. Assistant Prof., Computer Science, K.R.College of Arts & Science, Kovilpatti, Tamilnadu, (India)*

ABSTRACT

The proposed method tries to develop thresholding concept and K-means algorithm to obtain high performance and efficiency. Gray value thresholding is a segmentation technique commonly applied to medical images. Many procedures have been proposed to optimally select the grey value thresholds based on the intensity of the grey level image. The main objective of the present work is the attainment of proper segmentations of mammographic images which are taken from Mammographic Image Analysis Society (MIAS) digital mammogram database. In this study normal breast images and breast image with masses used as the standard input to the proposed system, are taken from Mammographic Image Analysis Society (MIAS) digital mammogram database so as to be used in tasks such as to find micro calcification. To full this purpose, we suggest that our work should follow two different techniques, namely, local thresholding technique and K-means clustering technique. Thus, two techniques are used to segment the mammographic images in order to trace the abnormal areas and to find the best one suitable for the objective.

Thresholding approaches segment scalar images by creating a binary partitioning of the image intensities. A thresholding procedure attempts to determine an intensity value, called the threshold, which separates the desired classes. The segmentation is then achieved by grouping all pixels with intensities greater than the threshold into one class and all other pixels into another class. Determination of more than one threshold value is a process called multi thresholding. By detailed theoretical study, the k-means clustering and local thresholding technique were chosen for evaluation. The performance of image segmentation using k-means clustering was evaluated for the parameters like PSNR and MSE, and for thresholding concept parameters like Processing speed and Nice segment cut are evaluated. Finally, a higher performance is achieved with k-means clustering based thresholding method.

I INTRODUCTION

An image is not just a random collection of pixels to humans; it is a meaningful arrangement of regions and objects. There also exits a variety of images: natural scenes, paintings, etc. Image segmentation is the first step in image analysis and pattern recognition. Image segmentation is the process of dividing an image into different regions such

that each region is homogeneous. There are two ways of image segmentation, (i) Thresholding concept and (ii) K-means clustering.

1.1 Thresholding concept

There are several types of thresholding concepts, they are

i. Soft thresholding - shrinks all the coefficients towards the origin.
ii. Quantile thresholding is useful to reduce quantity-based noise.
iii. Universal thresholding - all wavelet coefficients are threshold on some formulae. It may be soft or hard.

The output [1] of the thresholding operation is a binary image whose one state will indicate the foreground objects, that is, printed text, a legend, a target, defective part of a Material, etc., while the complementary state will correspond to the background. Depending on the application, the foreground can be represented by gray-level 0, that is, Black as for text, and the background by the highest luminance for document paper, that is 255 in 8-bit images, or conversely the foreground by white and the background by black. Various factors, such as non stationary and correlated noise, ambient illumination, busyness of gray levels within the object and its background, inadequate contrast, and object size not commensurate with the scene, complicate the thresholding operation. Finally, the lack of objective measures to assess the performance of various thresholding algorithms, and the difficulty of extensive testing in a task oriented environment, has been other major handicaps. In [2], they developed adaptive thresholding technique for segmenting H-α solar images to get back with a foreground segmented filaments and a non-ROI background. Based on false acceptance rate and output images, it can be concluded that our ALT technique is the best. The well defined and visible filaments could in future be considered for further studies by characterizing the features which may give us the ability to provide work for machine vision techniques. [3] included that thresholding is a useful technique for images containing solid objects with contrasting background . According to the thresholding rule, the pixels at or above the threshold fall outside the object. Thresholding techniques can be classified into bi-level and multi level thresholding. In bi-level thresholding the image is partitioned into two regions: object and background. Multi-level thresholding is applied when the image is composed of several objects with different surface characteristics, as we need several thresholds of segmentation. Threshold determination methods work only when object size is large enough to make a distinct mode in the histogram and noise is spatially unconelated.. In the first method image is subdivided into several sub-images and for each sub image histogram is computed, smoothed and a threshold is determined using interpolation. The second algorithm locates the object in the image using the intensity gradient which guides to determine an initial threshold for various areas of the image.

The [4] insisted that Global grey value thresholding is a conceptually trivial, yet often used segmentation technique in tomography. They presented an innovative approach, called PDM (Projection Distance Minimization), to find the optimal threshold grey levels by exploiting the available projection data. Reprojection of the segmented image and subsequent comparison with the measured projection data yields an objective criterion for the quality of segmentation. They approached at minimizing the projection distance. The experimental results show that if the

original object consists of only a few different materials, PDM results in a small difference between the original object and the reconstruction.

The thresholding methods were categorized [5] in six groups' according to the information they are exploiting. These categories are:

1. Histogram shape-based methods, where, for example, the peaks, valleys and curvatures of the smoothed histogram are analyzed
2. Clustering-based methods, where the gray-level samples are clustered in two parts as background and foreground (object) or alternately are modeled as a mixture of two Gaussians
3. Entropy-based methods result in algorithms that use the entropy of the foreground and background regions, the cross-entropy between the original and binarized image, etc.
4. Object attribute-based methods search a measure of similarity between the gray-level and the binarized images, such as fuzzy shape similarity, edge coincidence, etc.
5. The spatial methods use higher-order probability distribution and/or correlation between pixels
6. Local methods adapt the threshold value on each pixel to the local image characteristics.

1.2 K-Means Clustering

Clustering is a classification technique. Given a vector of N measurements describing each pixel or group of pixels (i.e., region) in an image, a similarity of the measurement vectors and therefore their clustering in the N-dimensional measurement space implies similarity of the corresponding pixels or pixel groups. Therefore, clustering in measurement space may be an indicator of similarity of image regions, and may be used for segmentation purposes.

Most popular clustering algorithms suffer from two major drawbacks: First, the number of clusters is predefined, which makes them inadequate for batch processing of huge image databases. Secondly, the clusters are represented by their centroid and built using an Euclidean distance therefore inducing generally an hyperspheric cluster shape, which makes them unable to capture the real structure of the data. This is especially true in the case of color clustering where clusters are arbitrarily shaped.

K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating globular clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative. Clustering in high dimensions, [6], has been an open problem for many years. Recent research has shown that it may be preferable to use dimensionality reduction techniques before clustering, and then use a low-dimensional clustering algorithm such as k-means, rather than clustering in the high dimension directly. The author shows that using a simple, inexpensive linear projection preserves many of the properties of data (such as cluster distances), while making it easier to find the clusters. Thus there is a need for good-quality, fast clustering algorithms for low-dimensional data. Our work is a step in this direction. Additionally, recent image segmentation algorithms such as normalized cut are based on eigenvector computations on distance matrices. These "spectral" clustering algorithms

still use k-means as a post-processing step to find the actual segmentation and they require k to be specified. Thus we expect G-means will be useful in combination with spectral clustering.

Regarding mammograms, [7] ,the use of computational tools to aid detection and diagnosis of breast masses has grown and gained increasing acceptance in recent years, as a kind of ”second readers” of medical images. These tools have been contributing to increase the early detection rates for breast cancer. They presented a methodology for detection of masses in digital screening mammograms, which can also be used in the development of a CAD tool. Such methodology used for both purposes is subdivided into preprocessing, segmentation through K-means algorithm, reduction of mass candidates, and classification of segmented structures into mass or non-mass, based on co-occurrence matrix, shape descriptors (eccentricity, circularity and convexity) and Support Vector Machine classification. The results indicate that the use of these techniques in the detection of masses is promising, since it achieves accuracy rates of over 85%. This will lead to a natural development of a CAD system capable of assisting health professionals in the painstaking task of tracing mammograms in search of mass abnormalities.

1.2.1 K-Means Algorithm Properties

- There are always K clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the center of clusters.

II SEGMENTATION USING THRESHOLDING CONCEPT

2.1 Thresholding Process

Thresholding is a useful technique for images containing solid objects with contrasting background. According to the thresholding rule, [3], the pixels at or above the threshold, fall outside the object. Thresholding techniques can be classified into bi-level and multi level thresholding. In bi-level thresholding the image is partitioned into two regions: object and background. Multi-level thresholding is applied when the image is composed of several objects with different surface characteristics, as we need several thresholds of segmentation. Threshold determination methods work only when object size is large enough to make a distinct mode in the histogram and noise is spatially uncorrelated

2.1.1. Algorithm for Thresholding process

The following steps are used for thresholding process [8].

i. An initial threshold (T) is chosen; this can be done randomly or according to any other method desired.

ii. The image is segmented into object and background pixels as described above, creating two sets:

   a. G1 = {f(m,n):f(m,n) > T} (object pixels)

b. G2 = {f(m,n):f(m,n) ≤ T} (background pixels) (note, f(m,n) is the value of the pixel located in the mth column, nth row)

iii. The average of each set is computed.

a. m1 = average value of G1

b. m2 = average value of G2

iv. A new threshold is created that is the average of m1 and m2

a. T' = (m1 + m2)/2

v. Go back to step two, now using the new threshold computed in step four, keep repeating until the new threshold matches the one before it (i.e. until convergence has been reached).

The result of the image segmentation was shown in Fig.1(f)-(j) by the thresholding concepts.

III. K-MEANS CLUSTERING PROCESS

K-means is one of the most popular adaptive technique for texture segmentation that is a generalization of the k-means algorithm [9]. It is simple and fairly fast. K-means is initialized from some random or approximate solution. Each iteration assigns each point to its nearest cluster and then points belonging to the same cluster are averaged to get new cluster centroids. Each iteration successively improves cluster centroids until they become stable.

3.1 Steps of K-means clustering Algorithm

1. Initialization – generate the starting condition by defining the number of clusters and randomly select the initial cluster centers.
2. Generate a new partition by assigning each data point to the nearest cluster center.
3. Recalculate the centers for clusters receiving new data points and for clusters losing data points.
4. Repeat the steps 2 and 3 until a distance convergence criterion is met.

The *K*-means clustering is a partitioning method for grouping objects so that the within-group variance is minimized. By minimizing dissimilarity of each subset locally, the algorithm will globally yield an optimal dissimilarity of all subsets.

The algorithm, as applied to image threshold, is given by the following steps:

1. Initialize the (*K*) class centers. For simplicity, un equal-distance method is used to define the initial class centers:

$$Center_i^0 = GL_{min} + [(i-i/2)\,(GL_{max} - GL_{min})/k)] \quad \text{---------- (1)}$$
$$i = 1, 2 \ldots k$$

Where $Center_i^0$ is the initial class center for the i th class, GL_{max} and GL_{min} are the maximum and minimum of the gray value *GL* in the sample space, respectively.

2. Assign each point to its closest class center. The criterion to assign a point to a class is based on the Euclidean distance in the feature (*GL*) space using:

$$Distance_{i,j} = \text{abs}\,(GL_j - Center_i) \qquad \text{---------- (2)}$$

$$i = 1, 2 \ldots K; j = 1, 2 \ldots N.$$

Where $Distance_{i,j}$ is the distance from the *j* th point to the *i* th class, and *N* is the total number of points in the sample space.

3. Calculate the (*K*) new class centers from the mean of the points that are assigned to it. The new class centers are calculated by

$$Center_i^m = 1/N_i \sum_{i=1}^{N_i} GL_j \qquad \text{---------- (3)}$$

$$j=1, 2 \ldots K.$$

Where N_i is the total number of points that are assigned to the *i*th class in step 2.

4. Repeat step 2 if any class centers change, otherwise end the circulation.

5. The threshold value is defined as the average of the *K*th class center and the (*K*-1) th class center:

$$Threshold = 1/2(center_k + center_{k-1}). \qquad \text{---------- (4)}$$

The result of the image segmentation was shown in Fig.1 (k)-(o) by the *K*-means clustering (k=7).

IV. PERFORMANCE EVALUATION

Performance Evaluation is mainly used to compare different techniques under Image processing. In this project the threshold concept and k-means clustering algorithms are compared. The parameters like PSNR and MSE values are calculated only for k-means clustering. The segmented images obtained by thresholding concept is based on object one and background. The grey level value with 0 represents the white that is object and the grey value with 1 represents the colour black that is background of the images. Table 1 shows the performance of the thresholding concept, evaluated by few parameters.

TABLE 1 Segmentation using Threshold concept

Method	Manually chosen value	Segment similar colour	Nice segment cut	Processing time
Thresholding concept	Possible	Satisfactory	Poor	slow

The Peak Signal to Noise Ratio (PSNR) is calculated by using the following formula.

$$\text{PSNR} = 10\log_{10}\left(\frac{255^2}{\text{MSE}}\right) \quad \text{----------(5)}$$

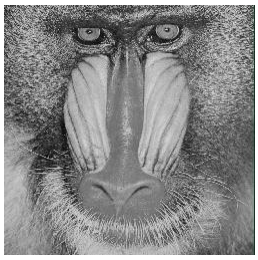
The PSNR values for k-means clustering are shown in Table 2. The graphical representation is shown in Fig. 2.

TABLE 2. PSNR Values

Image	PSNR
Baboon	16.24
Bridge	11.45
Cameraman	14.54
Clown	14.20
Peppers	15.10

TABLE 3. MSE Values

Image	MSE
Baboon	1543.63
Bridge	4653.24
Cameraman	2284.76
Clown	2468.98
Peppers	2008.58


(a) Baboon


(f)

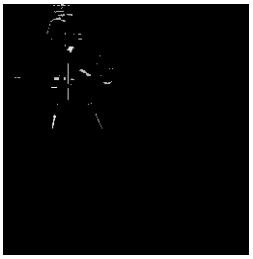

(k)

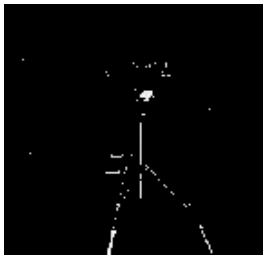

(b) Bridge


(g)


(l)


(c) Cameramen


(h)


(m)


(d) Clown


(i)


(n)


(e)Peppers


(j)


(o)

Figure 1 (a)-(e) Original Images, (f)-(j) Segmented images using threshold concept, (k)-(o) Segmented images using k-means clustering concept

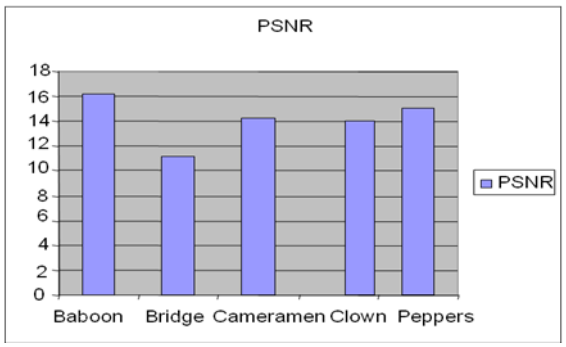


Figure 2. PSNR values for k-means clustering

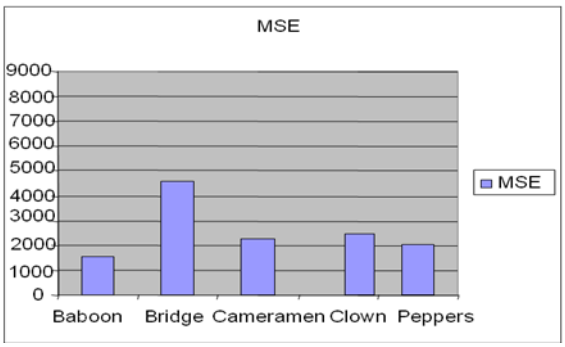


Figure 3. MSE values for k-means clustering

The Mean Square Error is calculated by using the following formula.

$$MSE = \frac{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(\hat{x}_{i,j} - x_{i,j})^2}{\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(x_{i,j})^2}, \qquad \text{----------(6)}$$

The MSE values for k-means clustering are shown in Table 3. The graphical representation is shown in Fig.3.

TABLE 4 Comparison between threshold concept and k-means clustering

Method	Manually chosen value	Segment similar color	Nice segment cut	Processing time
Thresholding concept	Possible	Satisfactory	Poor	Slow
K-means clustering	Possible	Good	Good	Fast

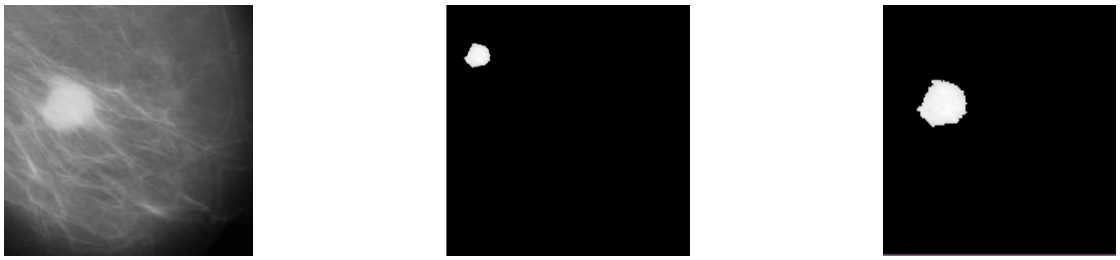
V. APPLICATION

Breast cancer is the most frequent neoplasm in women in western countries and is the leading cause of cancer deaths among women age forty to forty-nine. It is second only to lung cancer as a cause of cancer death among women. Breast cancer accounts for one of every three cancer diagnoses in women.

5. 1 Methodology

The digital mammographic images are given to image preprocessing and filtering steps before they are segmented. The underlying principle of preprocessing is to enlarge the intensity difference between objects and background and to produce reliable representations of breast tissue structures. The preprocessed mammogram is filtered by Gaussian Low pass filter. Then the mammographic images are given as input and it is segmented by both thresholding technique and k-means clustering.

The original mammographic images, shown in fig. 4(a) and fig. 5(a), are used for the experiments. The segmented images using thresholding concept are shown in fig. 4(b) and 5(b) and that of k-means clustering concepts are shown in fig.4(c) and 5(c).



(a) **(b)** **(c)**

Figure 4. Original mammographic image & its segmented images using (b) Threshold, (c) K-means clustering

(a) **(b)** **(c)**

Figure 5. Original mammographic image & its segmented images using (b) Threshold, (c) K-means clustering

Thus the original mammographic images were segmented using thresholding concept and k-means clustering concept to find the micro calcifications in order to find the some abnormal tissues in the breast. The detection is based on the intensity value of the pixel in the mammographic images.

VI. CONCLUSION

In this paper, a comparative study of different image segmentation techniques was performed. The k-means clustering and threshold techniques were chosen for evaluation. Using these two techniques, the performance for different images (also mammographic images) were evaluated. Finally, a higher performance is achieved by k-means clustering when compared with thresholding method. So this k-means clustering technique was proposed for segmentation in mammographic images. In future this k-means clustering is to be applied for obtaining better performance in order to trace more accurate micro calcifications in mammographic images.

REFERENCES

[1] Jan Meyer and Ray Land "*Threshold Concepts and Troublesome Knowledge*" Enhancing Teaching-Learning Environments in Undergraduate Courses Project, Higher and Community Education, School of Education, University of Edinburgh, Paterson's Land, Holyrood Road, Edinburgh EH8 8AQ. Occasional Report 4, May 2003.

[2] Ibrahim A. Atouma, Rami S. Qahwaji, Tufan Colak and Zakir H. Ahmed "Adaptive thresholding technique *for solar filament segmentation".*

[3] Kavitha Nagarajan "*Adaptive Clustering For Segmentation And Classification"* Texas Tech University, May, 2000.

[4] K. J. Batenburg and J. Sijbers "*Optimal Threshold Selection For Tomogram Segmentation By Projection Distance Minimization",* Belgium.

[5] Mehmet Sezgin and Bu¨ lent Sankur "*Survey over image thresholding techniques and quantitative performance evaluation***"** Journal of Electronic Imaging 13(1), 146–165 (January 2004).

[6] Greg Hamerly, Charles Elkan "*Learning the k in k-means*" University of California,San Diego.

[7] Leonardo de Oliveira Martins, Geraldo Braz Junior, Aristofanes Correa Silva, Anselmo Cardoso de Paiva+, and Marcelo Gattass " *Detection of Masses in Digital Mammograms using K-means* and Support Vector Machine" July 2009.

[8] Gonzalez, Rafael C. & Woods, Richard E. (2002). Thresholding. In Digital Image Processing, pp. 595–611. Pearson Education.

[9] Nesrine Chehata , Nicolas David , Frédéric Bretar" *Lidar Data Classification Using Hierarchical K-Means Clustering"* The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences. Vol. XXXVII. Part B3b. Beijing 2008.