An Adaptive K-means Clustering Algorithm for Breast Image Segmentation

Bhagwati Charan Patel
Associate Professor (IT)
Shri Shankaracharya College of Engg. & Tech.,
Bhilai, India

Dr. G.R.Sinha
Professor & Head(IT))
Shri Shankaracharya College of Engg. & Tech.,
Bhilai, India

ABSTRACT
Breast cancer is one of the major causes of death among women. Small clusters of micro calcifications appearing as collection of white spots on mammograms show an early warning of breast cancer. Early detection performed on X-ray mammography is the key to improve breast cancer diagnosis. In order to increase radiologist's diagnostic performance, several computer-aided diagnosis (CAD) schemes have been developed to improve the detection of primary identification of this disease. In this paper, an attempt is made to develop an adaptive k-means clustering algorithm for breast image segmentation for the detection of micro calcifications and also a computer based decision system for early detection of breast cancer. The method was tested over several images of image databases taken from BSR APPOLO for cancer research and diagnosis, India. The algorithm works faster so that any radiologist can take a clear decision about the appearance of micro calcifications by visual inspection of digital mammograms and detection accuracy has also improved as compared to some existing works.

Keywords
K-mean;, breast image; segmentation; detection; CAD.

1. INTRODUCTION
One in eight deaths worldwide is due to cancer. Cancer is the second leading cause of death in developed countries and the third leading cause of death in developing countries. In 2009, about 562,340 Americans died of cancer, more than 1,500 people a day. Approximately 1,479,350 new cancer cases were diagnosed in 2009. In the United Sates, cancer is the second most common cause of death, and accounts for nearly 1 of every 4 deaths [1]. The chance of developing invasive breast cancer at some time in a woman's life is about 1 in 8 (12%) [2]. Breast cancer continues to be a significant public health problem in the world. Approximately 182,000 new cases of breast cancer are diagnosed and 46,000 women die of breast cancer each year in the United States [3]. Thus, the incidence and mortality of breast cancer are very high, so much so that breast cancer is the second leading cause of cancer death in women. The chance that breast cancer will be responsible for a woman's death is about 1 in 35 (about 3%) [2]. In 2009, about 40,610 women died from breast cancer in the United States [4]. Although breast cancer has very high incidence and death rate, the cause of breast cancer is still unknown. No effective way to prevent the occurrence of breast cancer exists. Although breast cancer has very high incidence and death rate, the cause of breast cancer is still unknown [1]. No effective way to prevent the occurrence of breast cancer exists. Therefore, early detection is the first crucial step towards treating breast cancer. It plays a key role in breast cancer diagnosis and treatment. This process requires image segmentation and analysis of the images. Based on the analysis, detection of breast cancer along with location of affected area is identified.

2.BREAST CANCER DETECTION METHODS
Breast cancer screening is vital to detecting breast cancer. The most common screening methods are mammography and sonography. Compared to mammography, breast ultrasound examinations have several advantages [5]. Breast ultrasound examinations can obtain any section image of breast, and observe the breast tissues in real-time and dynamically. Ultrasound imaging can depict small, early-stage malignancies of dense breasts, which is difficult for mammography to achieve. Several statistical studies on the accuracy rate of breast disease diagnosis using ultrasonic examination have been carried out [6, 7]. The ultrasound examination has a high detection rate of tumors, in particular of malignant tumors. Accuracy rate of breast disease diagnosis using ultrasonic examination depends segmentation of images.

In order to increase detection and diagnosis accuracy and save to labor, computer aided detection (CAD) systems have been developed to help radiologists to evaluate medical images and detect lesions at an early stage. In general, CAD is a procedure that employs computers to assist doctors in the interpretation of medical images [8]. A CAD system is an interdisciplinary technology combining elements of digital image processing with radiological image processing. It combines image processing techniques and experts' knowledge for greatly improved accuracy of abnormality detection. In particular, the CAD system for automated detection/classification of masses and micro classification of clusters can be very useful for breast cancer control.

A typical CAD system shown in Fig.1 depicts that the preprocessing module, mammograms will be digitized in order to be processed by computer. Since more than one-third of a mammogram is dark breast background that comprised with noise and only provides very little information [8, 9], it is better to eliminate this unwanted information. The region of interests (ROIs) that contains possible micro calcifications (MCCs) is selected. However, some of detected pixels in ROIs may contain noise or breast tissue, so in order to extract the genuine MCCs,

contrast enhancement and segmentation process are really important. The purpose of contrast enhancement is to improve the low contrast of calcified pixels while segmentation will segment the detected MCCs from the breast region. Lastly, the segmented images are subjected to radiologist for diagnosis process to classify the MCCs into benign, malignant, suspicious and normal.

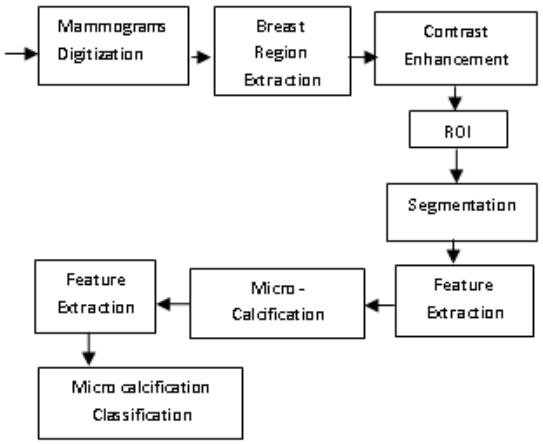

Figure 1. A CAD system for detection of micro calcification in breast images.

3.PROPOSED METHOD

A breast cancer CAD scheme separates suspicious regions that may contain masses from the background parenchyma – the tissue characteristic of an organ, as distinguished from associated connective or supporting tissues [2-6]. In other words, such schemes partition the mammogram into several nonintersecting regions and extract regions of interest (ROIs) and suspicious mass candidates from the ultrasound image. While a suspicious area is darker than its surroundings, it has a similar density, a regular shape of variable size. Thus, image segmentation is essential to maintaining the sensitivity and accuracy of the entire mass detection and classification system.

We have proposed an adaptive K-means segmentation method for detection of micro calcifications in digital mammograms. In the present work, we have made an attempt to improve the performance of existing K-means approach by varying various values of certain parameters discussed in the algorithm [11-13].

The K-means algorithm is an iterative technique that is used to partition an image into *K* clusters. In statistics and machine learning, ***k*-means clustering** is a method of cluster analysis which aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean. The basic algorithm is:

- Pick *K* cluster centers, either randomly or based on some heuristic;
- Assign each pixel in the image to the cluster that minimizes the distance between the pixel and the cluster center;
- Re-compute the cluster centers by averaging all of the pixels in the cluster

Repeat last two steps until convergence is attained (e.g. no pixels change clusters.

Given a set of observations ($\mathbf{x}_1$, $\mathbf{x}_2 \ldots \mathbf{x}_n$), where each observation is a *d*-dimensional real vector, *k*-means clustering aims to partition the *n* observations into *k* sets ($k < n$) $\mathbf{S} = \{S_1, S_2 \ldots S_k\}$ so as to minimize the within-cluster sum of squares (WCSS):

$$\arg\min_{\mathbf{S}} \sum_{i=1}^{k} \sum_{\mathbf{x}_j \in S_i} \|\mathbf{x}_j - \boldsymbol{\mu}_i\|^2 \quad (1)$$

Where $\boldsymbol{\mu}_i$ is the mean of points in S_i.

The most common algorithm uses an iterative refinement technique. Due to its ubiquity it is often called the ***k*-means algorithm**; it is also referred to as **Lloyd's algorithm**, particularly in the computer science community. Given an initial set of *k* means $\mathbf{m}_1^{(1)},\ldots,\mathbf{m}_k^{(1)}$, which may be specified randomly or by some heuristic, the algorithm proceeds by alternating between two steps[14].

Assign each observation to the cluster with the closest mean by

$$S_i^{(t)} = \left\{ \mathbf{x}_j : \left\| \mathbf{x}_j - \mathbf{m}_i^{(t)} \right\| \le \left\| \mathbf{x}_j - \mathbf{m}_{i^*}^{(t)} \right\| \text{ for all } i^* = 1, \ldots, k \right\} \quad (2)$$

Calculate the new means to be the centroid of the observations in the cluster.

$$\mathbf{m}_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{\mathbf{x}_j \in S_i^{(t)}} \mathbf{x}_j \quad (3)$$

We have modified the algorithm as follows:

The histogram is summary graph showing a count of data points falling in various ranges. The effect is rough approximation of the frequency distribution of data. The group of data is called classes, and in context of histogram they are known as bins, because one can think of them as containers that accumulate data and fill up at a rate equal to the frequency of that data class. The shape of the histogram sometimes is particularly sensitive to the number of bins. If the bins are too wide, important information might get omitted. By reducing the number of bins and increasing the number of classes in the K-means algorithm, the detection accuracy is found to be increasing. Quantization in terms of color histograms refers to the process of reducing the number of bins by taking colors that are very similar to each other and putting them in the same bin. By default the maximum number of bins one can obtain using the histogram function is 256. For the purpose of saving time when trying to compare color histograms, one can quantize the number of bins. Obviously quantization reduces the information regarding the content of images but as was mentioned this is the tradeoff when one wants to reduce processing time.

4.RESULTS AND DISCUSSION

A database of 150 breast images was formed. All of the real time breast images were collected from a reputed cancer diagnostic and research center (BSR APPOLO hospital for cancer diagnosis and research). Some of the images were subjected to color segmentation process using MATLAB 7.3 and P-IV. The results for different values of number of bins and classes have been discussed. Figure 2 shows an original image from the image database. Results for Bins =5 and varying the values of classes are shown in Figure 3- Figure 6. It can been seen that benign and malignant elements in the breast image became more clear i.e. by increasing the number of classes keeping constant value of Bins, visual appearance and classification of micro calcification get improved.

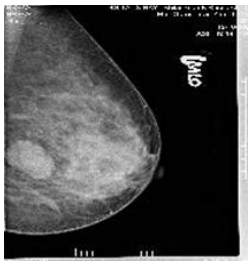

Figure 2. Original bresat image from the image databse.

Figure 3. Results for Bins=5,Classes=5.

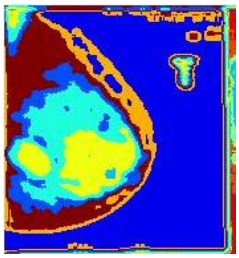

Figure 4. Results for Bins=5 Classes =10.

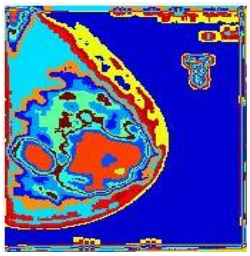

Figure 5. Results for Bins=5,Classes=15.

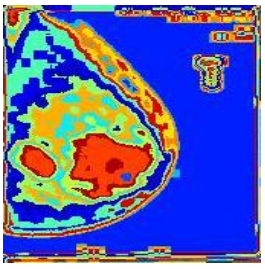

Figure 6. Results for Bins=5,Classes=20.

Figure 7- Figure 9 show the results for constant value of number of classes and increasing the number of Bins. This can be observed that the affeced regions are more accurately located i.e. the identification of affected area with malignant effects gets more prominent.

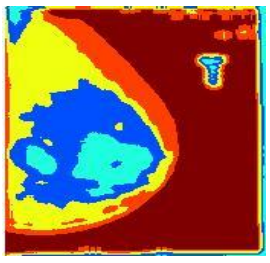

Figure 7. Results for Bins=10 Classes =5.

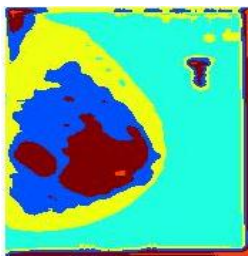

Figure 8. Results for Bins=15,Classes=5.

Figure 9. Bins=20,Classes=5.

Finally, the detection accuracy was estimated and compared the performance with previous similar research works emphasising the detection accuracy values. The results obtained are also in suport of anticipation with the findings and diagnosis by a senior radiologist of BSR APPOLO centre of cancer research. The accuracy of detection has increased. Table 1 shows the detection accuracy of proposed and exixting work.

TABLE 1. DETECTION ACCURACY OF MICRO CALIFICATION.

Type of Micro calcification	Micro calcification detection accuracy	
	Hieken et al [6], Saarenmaa, et al. [7]	Proposed adaptive K-means method
Benign Hyperplasia	84.5%	89.6%
Benign tumor	79.0%	77%
Malignant Tumor	88.5%	91%

5.CONCLUSIONS

In this paper, an attempt is made to implement k-means clustering algorithm for breast image segmentation for the detection of micro calcifications and also a computer based decision system for early detection of breast cancer in modified way. We have developed a computer aided decision system for the detection of micro calcifications in mammogram images the system is capable of detecting micro calcifications by visual inspection of digital mammograms. The feature selection is based on the number, color and shape of objects present in the image. The number of Bins values, number of Classes, sizes of the objects is considered as appropriate features for retrieval of image information. The detection accuracy was estimated and compared with existing works and it has been reported that the accuracy is improved if K-means algorithm is implemented adaptively. The results are found to be satisfactory when subjected to radiologists for their validation.

6.ACKNOWLEDGMENT

The authors extend their sincere thanks to Dr. Dilip Soni, a senior radiologist at APPOLO BSR centre for cancer research and diagnosis; for providing necessary support and guidance throughout the research work.

7.REFERENCES

[1] Garcia, M., Jemal, 2007. A., Ward, E., Center, M., Hao, Y., Siegel, R., and Thun, M. Global Cancer Facts & Figures 2007, American Cancer Society.

[2] Network of Strength. Breast Cancer Statistics. 2009. http://www. networkofstrength.org/i nformation/bcnews/stats.php.

[3] Cheng, H.D., Cai, X., Chen, X., Hu, L., and Lou, X. 2003 Computer-aided detection and classification of microcalcifications in mammograms: a survey. Pattern Recognition 36, vol. 12, p. 2967-2991.

[4] American Cancer Society. Jan. 2010. What Are the Key Statistics for Breast Cancer? http://www.cancer.org/docroot/cri/content/cri_2_4_1x_what_are_the_key_statistics_for_breast_cancer_5.asp.

[5] Laine, H.,Rainio, J., Arko, H. and Tukeva, T. 1995. Comparison of breast structure and findings by X-ray mammography, ultrasound, cytology and histology: a retrospective study. European Journal of Ultrasound 2, vol. 2 p. 107-115

[6] Hieken, T., Harrison, J. Herreros, J., and Velasco, J.2001. Correlating sonography, mammography, and pathology in the assessment of breast cancer size. American Journal of Surgery 182, vol. 4 , p. 351-354.

[7] Saarenmaa, I., Salminen, T., Geiger, U. Heikkinen, P., Hyv rinen, S., Isola, J., Kataja, V., Kokko, M., Kokko, R., and Kumpulainen, E.2001. The effect of age and density of the breast on the sensitivity of breast cancer diagnostic by mammography and ultasonography.Breast Cancer Research and Treatment 67, vol. 2 , p. 117-123.

[8]Wikipedia.Medicalimaging.2010.
http://en.wikipedia.org/wiki/Medical/imageprocessingultrasound

[9] R. Mousa, Q. Munib, and A. Moussa,2005.Breast Cancer Diagnosis System based on Wavelet Analysis and Fuzzy-Neural, Expert Systems with Applications, vol. 28, pp. 713-723.

[10] S. K. Lee, C-S. Lo, C-M. Wang, and P-C. Chung, 2000 .A Computer-Aided Design Mammography Screening System for Detection and Classification of Microcalcifications." International Journal of Medical Informatics, vol. 60, pp. 29-57.

[11] Jianbo Shi & Jitendra Malik (1997) Normalized Cuts and Image Segmentation, Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp. 731-737.

[12] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, & A. Y.Wu (2002) An efficien k-means clustering algorithm: Analysis and implementation Proc. IEEE Conf. Computer Vision and Pattern Recognition, pp.881-892.

[13] Lloyd, S. P. (1957).Least square quantization in PCM. Bell Telephone Laboratories Paper. Published in journal much later: Lloyd., S. P. (1982). Least squares quantization in PCM . IEEE Transactions on Information Theory, vol. **28** (2), p. 129–137. http://www.cs.toronto.edu/~roweis/csc2515-2006/readings/lloyd57.pdf.

[14] Kanungo, T.; Mount, D. M.; Netanyahu, N. S.; Piatko, C. D.; Silverman, R.; Wu, A. Y. 2002. "An efficient k-means clustering algorithm: Analysis and implementation". IEEE Trans. Pattern Analysis and Machine Intelligence **24**, p. 881–892. http://www.cs.umd.edu/~mount/Papers/pami02.pdf.