AUTOMATED INTEGRATED CLUSTERING ALGORITHM FOR MAMMOGRAPHIC MASS SEGMENTATION

K. Akila[1], L.S. Jayashree [2], A. Vasuki[3]

[1]Department Mechatronics Engineering, Kumaraguru college of Technology, Coimbatore, Tamilnadu, India
[2]Department of Computer Science and Engineering, PSG College of Technology, Coimbatore, Tamilnadu, India
[3]Department of Electronics and Communication Engineering, Kumaraguru college of Technology, Coimbatore, Tamilnadu, India

ABSTRACT

Segmentation plays an important role in mammographic image processing by facilitating the delineation of regions of interest. An automated Histon based integrated clustering algorithm is presented in this paper for the detection masses in mammographic images by integrating K-means clustering algorithm with Fuzzy C-means algorithm. Initially Histon of the input image was calculated and given as initial centroid for K-means clustering algorithm and Fuzzy C-means algorithm was applied to segment the mass. The performance of proposed algorithm was evaluated using area overlap measure. The morphological features are extracted from the segmented mass and 85% classification accuracy was obtained using SVM classifier.

Index Terms— Segmentation, Histon, K-means clustering, FCM.

I. INTRODUCTION *(HEADING 1)*

Breast cancer is the most common type of invasive cancer found among women worldwide [1, 2]. There is a good chance of recovery from death if it is detected in its initial stages and the same can be treated before it spreads to nearby parts of the body. Mammography is considered as one of the best procedures for early detection of Breast cancer. A screening mammogram is used to detect the signs of breast cancer in its initial stages even when there are no symptoms of breast problems and even before a lump can be felt. The major abnormalities that can found often in mammograms are Masses and Microcalcifications (MCs). Breast mass is a localized swelling, or lump in the breast. The most noteworthy features that indicate whether a mass may be benign or malignant are its *shape* and the character of its *margins*. The *shape* can be round, oval, lobulated, or irregular. Circumscribed oval and round masses are usually benign where malignant masses are generally irregular in shape. The *margins* can be described as circumscribed, micro lobulated, obscured (partially hidden by adjacent tissue), indistinct, or spiculated [4]. Masses are more difficult to detect than MCs because the features of a mass [3] may be obscured by or be like those of normal breast parenchyma. Thus, mass detection and segmentation remains to be a significant topic in breast cancer detection.

Segmentation refers to the process of partitioning a digital image into disjoint homogenous regions. These regions usually contain similar objects of interest which are more meaningful and easier to analyze. The Unsupervised image segmentation techniques can be classified into Region-based methods, Contour-based methods, Clustering methods [5].

Region growing method [6] is effective for segmenting the masses with blurred edge and good at overcoming the interference of adhesion tissue. Adaptive Thresholding based on multiresolution Analysis method [7] can detect suspicious lesions of different types at low false positive rates. The OTSU Threshold makes use of only the zeroth- and the first-order cumulative moments of the gray-level histogram and hence is trouble-free. It is possible to extend the method to multithreshold problems in an uncomplicated manner [8]. Patel and Sinha [9] proposed a topographic representation called the isocontour map, in which a salient region forms a dense quasi-concentric pattern of contours. The algorithm concurrently delineates the boundaries of the breast boundary, the pectoral muscle, as well as dense regions that include candidate masses. The topological and geometrical structure of the image is analyzed using an inclusion tree that is a hierarchical representation of the enclosure relationships between contours. The topographic representation is obtained from a set of isocontours at multiple distinct partition values over the intensity range of the image. The cluster based methods segment the image into diverse groups based on the similarity of the intensity of pixels. In k-Means Clustering Algorithm [10], the accuracy will be improved if it is implemented adaptively. The fuzzy C-means algorithm is used to segment fatty versus dense tissue types in the mammograms. The clustering based methods obtain better sensitivity at individual level. FAFCM requires an initial estimate of centroid values. Proper selection will generally improve accuracy and reduce the number of iteractions as well as increase the speed [13]. Although clustering algorithms do not directly incorporate spatial modeling and can therefore be sensitive to noise and intensity inhomogeneities, this lack of spatial modeling, however, can provide significant advantage for fast computation.

Several segmentation techniques have been developped by various researchers so far. The aim of the segmentation is to extract Region of Interest (RoI) containing all masses and locate the suspicious mass candidates from the RoI. In this paper, we proposed an automated histon based segmentation algorithm by integrating K-means and fuzzy c means algorithm.

The rest of the paper is organized as follows: *Section* 2 describes the proposed segmentation algorithm, *Section* 3 gives the results and discussion, and *Section* 4 concludes the paper.

II. PROPOSED SEGMENTATION ALGORITHM

In segmentation using cluster based algorithms, the initial centroid plays a key role for initial membership function. For that the initial centroid values are calculated using the concept Histon for the given input image [11,12,13,14]. After that, the maximum value histon is given to the initial centroid value of K-means integrated FCM algorithm [15] to perform the mass segmentation process.

A.**Histon Construction**: A rough set is a representation of a vague concept using a pair of precise concepts called lower and upper approximations [14]. The lower approximation is a description of the universe of objects that are known with certainty, whereas the upper approximation is the description of the objects that possibly belong to the set. Based on this roughest theory, Mohabey and Ray [14] introduced a new concept of encrustation of the histogram, called histon. The histon correlates with the upper approximation of a set such that all elements belonging to this set are clarified as possibly belonging to the same segment or segments showing similar grey value. Consider I is a grayscale image, of size MXN. The histogram of the image can be computed as follows:

$$h_i(x) = \sum_{m=1}^{M} \sum_{n=1}^{N} \delta(I(m,n) - x) \quad (1)$$

for 0 ≤ x ≤ L-1. Where δ (.) is Dirac impulse function and L is the total number of intensity levels. For a PXQ neighborhood around a pixel *I (*m, n), the total distance of all the pixels in the neighborhood and the pixel *I* (m, n) is then given by

$$d_T(m,n) = \sum_{p \in P} \sum_{q \in Q} d(I(m,n), I(p,q)) \quad (2)$$

Where d (I (m, n), I (p, q)) is the Euclidean distance.

The pixels in the neighborhood fall in the sphere of the similar color if the distance dT (m, n) is less than a Threshold T_0. We define a matrix X of the size M X N such that an element X (m, n) is given by

$$\boldsymbol{X}(m,n) = \begin{cases} 1 & d_T(m,n) < T_0 \\ 0 & otherwise \end{cases} \quad (3)$$

$$H_i(x) = \sum_{m=1}^{M} \sum_{n=1}^{N} (1 + X(m,n))\delta(I(m,n) - x) \quad (4)$$

The Histon is defined as using eqn. 4, the histon is constructed by finding the distance between the pixel and their neighborhood pixel values.

Mass segmentation: K-means algorithm works faster and robust and provides best results when the data points are distinct and not overlapped. This algorithm could not handle the noisy data and outliers. Fuzzy c-means Clustering Algorithm permits one piece of data to belong to one or more cluster. The better result is obtained for lower value of termination criteria at the cost of more number of iterations. In this work k-means and fuzzy c-means are integrated to obtain the benefit of both algorithms. The k-means clustering algorithm requires a priori specification of the number of cluster centers. Randomly choosing of the cluster center cannot lead us to the best result. In this proposed segmentation algorithm, the best taken maximum values of histon are assigned as initial centroids for K-means algorithm to cluster the image data points. The new cluster centers obtained by the K-means are given to fuzzy c-means algorithm. Based on the centroid values from the K-means clustering, the FCM segment the mass from the mammographic images more effectively.

III. Results And Discussion

The Mammography Image Analysis Society (MIAS) is a research organization in the UK produced a database of digital mammography images. The database consists of 322 images of mammograms [21]. There are 208 normal, 63 benign and 51 malignant (abnormal) images. It also includes radiologist's `truth'-markings on the locations of any abnormalities that may be present. We have taken 40 images containing masses for validating the proposed algorithm. In these 40 images, 23 mammograms contain circumscribed masses and 17 images consist of spiculated masses. The segmented masses of the sample mammograms are shown in Fig.2. Our method detected all the circumscribed masses accurately. In the case of speculated masses, the mass area was segmented successfully but the radially spiculated details are not covered properly.

A. Performance Evaluation

1) ***Area Overlap measure*** *[16,17] (Jaccard similarity index):* In the process of segmentation, accuracy is the degree to which the delineation of the object corresponds to ground truth. Here we utilized the standard measure Area Overlap measure (AOM). The measure AOM is given by

$$AOM = \frac{S \cap G}{S \cup G} \quad (5)$$

Where S is the segmented area and G is the ground truth. This necessitates the creation of a ground-truth dataset for the evaluation. The ground truth is generated by manually segmenting the breast region repressenting the mass from each mammogram. The boundary of the regions is then manually traced to extract the RoI to generate a ground truth (GT) image. A value close to one means a good match between two regions. The AOM for the proposed method is greater than 0.8 for high percentage of cases which shows the reasonable overlap between segmented area and ground truth. The averaged value for AOM on 40 images is 87.6%.

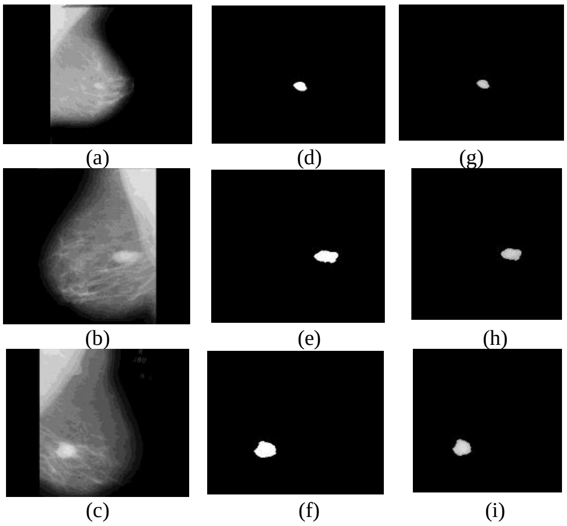

Fig.2 a,b,c, the original mammographic images, d,e,f final segmented ROI, and g,h,i ROI extracted from the mammogram

2) *Mammographic Risk Classification*

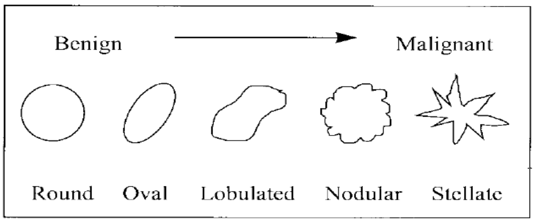

Fig.3 Morphological spectrum of Masses [6]

To perform a quantitative evaluation on mammographic image segmentation, we conducted mammographic risk classification from the segmented images using SVM classifier to assess the proposed segmentation. The results also show that overall good classifycation results incorrect segmentation can lead to misclassification. Feature extraction plays a fundamental role in many pattern recognition tasks. In general, Circumscribed oval and round masses are usually benign. An irregular shape suggests a greater likelihood of malignancy (Fig.3). Hence the morphological features are the noteworthy features to classify the masses. Based on the criteria shape and margin, the salient features Area, Perimeter, Maximum Radius, Minimum Radius, Entropy, Eccentricity, Circularity, Elongatedness, Shape Index and Standard deviation are extracted from the segmented masses [18,19,20].

The WEKA toolbox version 3.7 (Waikato Environment for Knowledge Analysis – Weka) was used to classify the features using the machine learning algorithm support vector machine. The Support Vector Machine categorizes the features extracted from the segmented masses into two classes, benign and malignant.

We have taken 28 images from benign category and 12 malignant. From the table.1, we can notice that out of 28 benign masses, 23 are correctly classified and 5 are misclassified as malignant. Similarly, out of 12 malignant masses, 11 are correctly classified, while one image is misclassified as benign. Various statistical measures such as Sensitivity, Specificity and accuracy are used as performance measures to validate the classification results given in table.2. Sensitivity gives the true positive rate and specificity gives the true negative rate. High values of sensitivity and specificity implies that the classification results obtained are statistically good. The classification accuracy achieved by the proposed method is 85%.

TABLE I. CLASSIFICATION RESULT

Mass type	Classified as Benign	Classified as Malignant
Benign	23	5
Malignant	1	11
Overall Classification rate	85%	

TABLE II. PERFORMANCE MEASURE

Measure	Experimental Value
$Sensitivity = \frac{TP}{TP+FN}$	0.8214
$Specificity = \frac{TN}{TN+FP}$	0.9166
$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$	0.85

IV. CONCLUSION

In the field of medical diagnosis, Image segmentation plays a significant role in medical image. In this paper, we proposed an automated integrated clustering algorithm for mammographic image segmentation. The initial cluster centers are decided by Histon values. Integration of K-means clustering with Fuzzy C-means algorithm provides the benefits of both methods in terms of minimal computation time and accuracy. The performance of the proposed technique has been demonstrated using area overlap measure and classification accuracy. The high values of AOM and classification accuracy proves the efficiency of the proposed method.

REFERENCES

[1] Breast cancer Statistics, http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/breast-cancer-statistics

[2] Tang, J., R.M. Rangayyan, J. Xu, I. El Naqa and Y. Yang, Computer-aided detection and diagnosis of breast cancer with mammography: recent advances, IEEE Transactions on Information Technology in Biomedicine 13(2): 236-251 (2009).

[3] Mencattini, A., M. Salmeri, G. Rabottino and S. Salicone, Metrological characterization of a CADx system for the classification of breast masses in mammograms. IEEE Transactions on Instrumentation and Measurement 59(11): 2792-2799 (2010).

[4] Sampat, M.P., M. K. Markey and A. C. Bovik, Computer-aided detection and diagnosis in mammography, Handbook of image and video processing 2(1): 1195-1217 (2005).

[5] Fu, K.S. and J.K. Mui, A survey on image segmentation. Pattern recognition 13(1): 3-16 (1981).

[6] Cao, Y., X. Hao, X. Zhu, and S. Xia, An adaptive region growing algorithm for breast masses in mammograms. Frontiers of Electrical and Electronic Engineering in China 5(2): 128-136 (2010).

[7] Oliver, A., J. Freixenet, J. Marti, E. Pérez, J. Pont, E.R. Denton and R. Zwiggelaar, A review of automatic mass detection and segmentation in mammographic images. Medical image analysis 14(2): 87-110 (2010).

[8] Hu, K., X. Gao and F. Li, Detection of suspicious lesions by adaptive thresholding based on multiresolution analysis in mammograms. Instrumentation and Measurement. IEEE Transactions 60(2): 462-472 (2011).

[9] Patel, B.C. and G.R. Sinha, An adaptive k-means clustering algorithm for breast image segmentation. International Journal of Computer Applications 10(4): 35-38 (2010).

[10] Panda, R.N., K.P. Bijay and M.R. Patro, Feature Extraction for Classification of Microcalcifications and Mass Lesions in Mammograms. International Journal of Computer Science and Network Security 9(5): 255 – 265 (2009).

[11] Boss, R.S., K. Thangavel and D.A. Daniel, Mammogram image segmentation using rough clustering. Int. J. Res. Engin. Technol. Pp.66-77 (2013).

[12] Basha, S.S. and K.S. Prasad, Automatic detection of breast cancer mass in mammograms using

morphological operators and fuzzy c--means clustering. Journal of Theoretical and Applied Information Technology 5(6): (2009).

[13] Nayak, J., M. Nanda, K. Nayak, B. Naik and H.S. Behera, An improved firefly fuzzy c-means (FAF CM) algorithm for clustering real world data sets/ Advanced Computing, Networking and Informatics 1: 339-348 (2014).

[14] Mohabey, A. and A.K. Ray, Rough set theory based segmentation of color images. Fuzzy Information Processing Society, NAFIPS. 19th International Conference of the North American Pp. 338-342 (2000).

[15] Abdel-Maksoud, E., M. Elmogy and R. Al-Awadi, Brain tumor segmentation based on a hybrid clustering technique. Egyptian Informatics Journal 16(1): 71-81 (2015).

[16] Oliver, A., X. Lladó, E. Pérez, J. Pont, E.R. Denton, J. Freixenet and J. Martí, A statistical approach for breast density segmentation. Journal of Digital Imaging 23(5): 527-537 (2010).

[17] Meghanathan, N., N. Chaki and D. Nagamalai, Advances in Computer Science and Information Technology. Computer Science and Engineering: Second International Conference CCSIT 2012, Bangalore, India, January 2-4, 2012. Proceedings Vol. 86 (2012).

[18] Surendiran, B. and A. Vadivel, Mammogram mass classification using various geometric shape and margin features for early detection of breast cancer. International Journal of Medical Engineering and Informatics 4(1): 36-54 (2012).

[19] Djaroudib, K., A.T. Ahmed, and A. Zidani, Textural Approach for Mass Abnormality Segmentation in Mammographic Images. arXiv preprint arXiv 1412-1506 (2014).

[20] He W., E.R. Denton, K. Stafford and R. Zwiggelaar, Mammographic image segmentation and risk classification based on mammographic parenchymal patterns and geometric moments. Biomedical Signal Processing and Control 6(3): 321-329 (2014).

[21] Database: http://peipa.essex.ac.uk/info/mias.html