International Journal of Hybrid Information Technology
Vol.8, No.10 (2015), pp.427-436
http://dx.doi.org/10.14257/ijhit.2015.8.10.39

A Novel Breast mass segmentation method based on patch merging and GHFCM

Shenghua Gu[1], Yunjie Chen[2], Jin Wang[2] and Jeong-Uk Kim[3]

[1] *Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology, Nanjing 210044*
[2] *College of Computer & Software, Nanjing University of Information Science & Technology, Nanjing 210044*
[3] *Department of Energy Grid, Sangmyung University, Seoul 110-743, Korea*

Abstract

Breast cancer is regarded as one of the most frequent mortality causes among women. It is very important to create a system to diagnose suspicious masses in mammograms for early breast cancer detection. In this paper, we propose an automatic breast mass segmentation method based on patch merging method and generalized hierarchical Fuzzy C Means (GHFCM). The patch merging method is used to obtain the adaptive region of interest (ROI), while the GHFCM method which is able to overcome the drawbacks of effect of image noise and Euclidean distance FCM which is sensitive to outliers is used to obtain the precisely mass segmentation results. The new method is evaluated over MiniMIAS dataset. The segmentation performance from experimentations demonstrates that our method outperforms the other compared methods.

Keywords: *Breast mass segmentation, Hierarchical distance function, Generalized mean, Patch merging, GHFCM*

1. Introduction

According to American cancer society, breast cancer is the most commonly diagnosed cancer type among women [1]. The early detection of breast cancer is the main way to increase recovery rate from disease. One of the ways to detect the breast cancer in early stages is to using mammography which is thought of as one of the most effective methods for breast cancer diagnose [2]. Recent developments in digital mammography imaging systems have aimed to better diagnosis of abnormalities in the breast and have increased the survival chance [3]. Nowadays, computer-aided diagnosis (CAD) system is widely used to assist the radiologists in breast masses detection and identification. CAD system which seem appealing to the radiologists generally consists of segmentation, feature extraction and classification stages [4]. Precisely breast masses segmentation in mammograms plays a critical role in the whole system and successfully influences the consequent stages.

In the past, many automatic or semi-automatic breast mass segmentation methods have been widely proposed. In these methods, one type of two step segmentation strategy is a very common mass segmentation pattern. These two steps contain the ROI generation which is cropped from the breast area and the mass boundary detection in ROI. Region growing method which using a similarity measure of the two adjacent pixels to group the pixels is used to generate the ROI in breast area [5-6]. However, this method is very sensitive to the noise and local optimum. In this paper, patch based method is used for ROI generation at first to overcome the effect of the noise.

After ROI generated, some boundary detection methods based on pattern recognition algorithms [7-11] can be used to segment the breast mass from ROI. In [12], a semi-automatic region growing approach is proposed based on the choice of the starting point by the radiologist. In [13], Kobatake, *et al.*, applied a modified Hough transform to

ISSN: 1738-9968 IJHIT

extract lines passing near the centre of the mass and automatically selected candidates based on the number of line-skeletons. In [15], Elter, *et al.,* proposed a contour tracing approach to extract shape of the region. The ROI of mass is transformed into polar coordinate system, and then contour of estimated mass is calculated by using a shortest path algorithm. In [15], Tao, *et al.,* proposed a classification system to identify spiculation of a mass by integrating machine learning method and graph-cut algorithm. Tolga, *et al.,* [16] proposed a breast mass contour segmentation method based on classical seed region growing with additional capability of adaptive threshold value to extract optimum contour information. Song, *et al.,* [17] use a plane fitting method based on dynamic programming optimization approach to propose a breast mass segmentation method.

As a classification method, FCM algorithm which have been widely studied and successfully applied in image clustering and segmentation is also able to detect the breast mass. However, the traditional FCM algorithm is very difficult to obtain the accurate mass segmentation results due to overlapping intensities, low contrast of images, and especially the noise perturbation. To overcome the effect of the noise, a wide variety of approaches have been proposed to incorporate spatial information in the image [18-19]. In [20], Zheng, *et al.,* proposed two algorithms, generalized FCM (GFCM) and hierarchical FCM (HFCM), then integrated them to propose GHFCM method to overcome the effect of noise and outliers. In this paper, we propose an automatic breast mass segmentation method based on patch merging method and generalized hierarchical FCM (GHFCM). The patch merging method is used to obtain the adaptive ROI, while the GHFCM method is used to obtain the precisely mass segmentation results. The new method is evaluated over MiniMIAS dataset. The segmentation performance from experimentations demonstrates that our method outperforms the other compared methods.

The advantages of our method are can be explained in three aspects. First, using patch merging method as the initial step to provide ROI can improve efficiency since the number of operated units is observably reduced. Second, GHFCM clustering process in ROI for detecting breast mass is much faster than that in the whole image since we avoid the abundant calculation of pixels which is outside the ROI. Finally, by combining the general FCM (GFCM) which is robust to the noise with the spatial constraints and hierarchical FCM (HFCM) which is robust to the outliers with the a general and flexible distance function to introduce a generalized hierarchical FCM (GHFCM), the breast mass segmentation results can be more accurately obtained in ROI generated in the first step.

2. Our Method

2.1 ROI Generation using ISODATA Clustering and Patch Merging

In mammograms, there are two different parts: one is breast tissue region which containing the breast mass and another one is non-breast tissue region which takes up a large region in the image. To effectively segment the breast mass, we need to firstly preprocess the mammogram and select a ROI which only contain the whole breast mass and breast normal tissue. In this subsection, we introduce a patch merging method to generate ROI.

Patches taken inside images, are at the very heart of many image processing applications [21], such as texture synthesis [22], image inpainting [23], image restoration [24]. The advantage of using patches to process the image is that it takes of very important image feature and is robust to the noise. Let $N(x)$ be image patch of the pixel x, the image can be divided into lots of similar patches with same size. Then we transform the patch $N(x)$ to a vector form $\mathrm{V}(x)$ and use ISODATA algorithm [25-26] to classify the vectors from all the image patches. ISODATA method is a method which added division of a cluster, and processing of fusion to the K-means method. The procedure of the ISODATA method is shown as follows:

Algorithm 1: ISODATA clustering

```
1. Choose randomly $K = K_0$ initial mean vectors $\{\mathrm{m}_1, \mathrm{m}_2, \cdots, \mathrm{m}_k\}$ from the dataset.
2. Assign each patch $V(x)$ of data point $x$ to the cluster with closest mean:
       $x \in \omega_i$ if $d(V(x), m_i) = \min\{d(V(x), m_1), \cdots, d(V(x), m_k)\}$
3. Discard clusters containing too few members, i. e., if $n_j < n_{\min}$, then discard $\omega_j$ and
   reassign its members to other clusters. $K \leftarrow K-1$.
4. For each cluster $\omega_j \, (j = 1, \cdots, K)$, update the mean vector
       $m_j = \frac{1}{n_j} \sum_{x \in \omega_j} V(x)$
   and the covariance matrix:
       $\Sigma_j = \frac{1}{n_j} \sum_{x \in \omega_j} (V(x) - m_j)(V(x) - m_j)^T$
   The diagonal elements are the variance $\sigma_1^2, \cdots, \sigma_N^2$ along the $N$ dimensions.
5. If $K \le K_0/2$ (too few clusters), go to Step 6 for splitting;
   else if $K \ge 2K_0$ (too many clusters), go to Step 7 for merging;
6. For each cluster $\omega_j \; (j = 1, \cdots, K)$, find the greatest covariance $\sigma_m^2 = \max\{\sigma_1^2, \cdots, \sigma_N^2\}$
     If $\sigma_m^2 > \sigma_{\max}^2$ and $n_j > 2n_{\min}$, then split into two new cluster centers
       $m_j^+ = m_j + \sigma_m$ , $m_j^- = m_j - \sigma_m$
   Alternatively, carry out PCA to find the variance corresponding to the greatest eigenvalue
   $\lambda_{\max}$ and split the cluster along the direction along the corresponding eigenvector.
   Set $K \leftarrow K+1$
   Go to Step 8.
7. (merge) Compute the $K(K-1)/2$ pairwise Bhattacharyya distance between every two
   cluster mean vectors:
   $d_B(\omega_i, \omega_j) = \frac{1}{4}(m_i - m_j)^T \left[\frac{\Sigma_i + \Sigma_j}{2}\right]^{-1} (m_i - m_j) + \log\left[\left|\frac{\Sigma_i + \Sigma_j}{2}\right| \Big/ (|\Sigma_i, \Sigma_j|)^{1/2}\right], \; 1 \le i, j \le K, \; i > j$
   For each of distances satisfying $d_B(\omega_i, \omega_j) < d_{\min}$, merge of the corresponding clusters to
   form a new one:
   $m_i = \frac{1}{n_i + n_j}[n_i m_i + n_j m_j]$
   Delete $m_j$, set $K \leftarrow K-1$
8. Terminate if maximum number of iterations is reached. Otherwise go to Step 2.
```

After the ISODATA clustering finished in the mammograms, we use the patch which contains the seed point and its neighborhood patches to construct a ROI to improve the segmentation efficiency. The patch is considered to be the neighborhood of the patch which contains the seed point if the distance between its center and the seed point is small than an acceptable limits and its label is same as seed point label. After that, a rectangle which exactly surround all above patches is selected as ROI for following breast mass segmentation. Figure 1 demonstrates the ROI generation results of different mammograms. The advantage of our ROI selected method is that the ROI is adaptive to the breast mass.

2.2 Breast Mass Segmentation in ROI using GHFCM

After ROI generation, we can segment the breast mass in ROI instead of the whole image domain to improve the efficiency. The clustering algorithms such as FCM or GMM can be use to perform this task. However, without the spatial prior knowledge, these clustering algorithms are still weak in imaging noise, outliers and other imaging artifacts. To overcome these drawbacks, a GHFCM which is proposed in[20] is applied to segment the breast mass in ROI. In [20], the generalized mean is incorporated into FCM to propose GFCM at first. Then, the distance function is estimated by a sub-FCM to propose HFCM. At last, GFCM and HFCM are combined to introduce Generalized Hierarchical FCM (GHFCM) to overcome the effect of image noise and outliers.

2.2.1 GFCM

In GFCM, the distance function of FCM is improved by using local generalized mean and the new object function is presented as follows:

$$J_m^{GFCM} = \sum_{i=1}^{N}\sum_{j=1}^{J} u_{ij}^m \sum_{c \in N_i} w_c d_{cj} \tag{1}$$

where w_c is the weighted factor to control the influence of the neighborhood pixels depending on their distance from the central pixel i . The strength of w_c should decrease as the distance between pixel c and i increase. In [20], Gaussian function is used to construct w_c :

$$w_c = 1\big/(2\pi\delta^2)^{1/2} \exp(-d_{ci}^2 / 2\delta^2) \tag{2}$$

By applying local weighted generalized mean on membership, the modified membership u_{ij} can be calculated as:

$$u_{ij} = \sum_{c \in N_i} w_c u_{cj} \Bigg/ \sum_{h=1}^{J}\sum_{c \in N_i} w_c u_{ch} \tag{3}$$

From Eq. (3), the new membership incorporating image spatial information according to the help of local weighted generalized mean, makes the algorithm is more robust to image noise.

2.2.2 HFCM

To overcome the outliers, the distance function in traditional FCM is assumed to be a sub-fuzzy model and generate the HFCM. The object function of HFCM is given as:

$$J_{m,n}^{HFCM} = \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K} u_{ij}^m v_{ijk}^n \bar{d}_{ijk} \tag{4}$$

It can be seen that at the second level of the hierarchy, information is provided about the data along with their class labels. It is noted that we have J classes; K subclasses correspond to each class of the first level. Using Euclidean distance $\bar{d}_{ijk} = \left\| y_i - \mu_{jk} \right\|^2$, and the presentation form of μ_{jk} is given as

$$\mu_{jk} = \sum_{i=1}^{N} u_{ij}^m v_{ijk}^n y_i \Bigg/ \sum_{i=1}^{N} u_{ij}^m v_{ijk}^n \tag{5}$$

By assuming each cluster to contain several sub-clusters, HFCM is robust to the outliers.

2.2.3 GHFCM

The generalized hierarchical FCM (GHFCM) is generated by combining HFCM and GFCM, and its objective function can be given as

$$\begin{aligned} J_m^{GHFCM} &= \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K} u_{ij}^m v_{ijk}^n \sum_{c\in N_i} w_c d_{cjk} \\ &= \sum_{i=1}^{N}\sum_{j=1}^{J}\sum_{k=1}^{K}\sum_{c\in N_i} u_{ij}^m v_{ijk}^n w_c \left\| y_c - \mu_{jk} \right\|^2 \end{aligned} \tag{6}$$

The membership u_{ij} and sub-membership v_{ijk} satisfies the constraint $\sum_{j=1}^{J} u_{ij} = 1$ and $\sum_{k=1}^{K} v_{ijk} = 1$, respectively. By applying the optimization way similar to the standard FCM, the parameters in GHFCM can be calculated iteratively as

$$u_{ij} = \left(\sum_{k=1}^{K}\sum_{c\in N_i} w_c v_{ijk}^n d_{cjk} \right)^{1/(1-m)} \Bigg/ \sum_{h=1}^{J} \left(\sum_{k=1}^{K}\sum_{c\in N_i} w_c v_{ijk}^n d_{cjk} \right)^{1/(1-m)} \tag{7}$$

$$v_{ijk} = \left(\sum_{c\in N_i} w_c u_{ij}^m d_{cjk} \right)^{1/(1-n)} \Bigg/ \sum_{k=1}^{K} \left(\sum_{c\in N_i} w_c u_{ij}^m d_{cjk} \right)^{1/(1-n)} \tag{8}$$

The cluster center μ_{jk} is evaluated as

$$\mu_{jk} = \sum_{i=1}^{N}\sum_{c\in N_i} u_{ij}^m v_{ijk}^n y_c \Bigg/ \sum_{i=1}^{N} u_{ij}^m v_{ijk}^n \tag{9}$$

According to [20], using local weighted generalized mean to incorporate spatial information and cluster information, the modified membership and sub-membership can be re-calculated as:

$$u_{ij} = \sum_{c\in N_i} w_c u_{cj} \Bigg/ \sum_{h=1}^{J}\sum_{c\in N_i} w_c u_{ch} \tag{10}$$

$$v_{ijk} = \sum_{c\in N_i} w_c v_{cjk} \Bigg/ \sum_{h=1}^{K}\sum_{c\in N_i} w_c v_{cjh} \tag{11}$$

For breast mass segmentation in ROI, we set $J = 2$ to represent breast tissue and breast mass in ROI and $K = 2$ to be the number of each sub-cluster. The integrated GHFCM algorithm is given as:

Algorithm 2: GHFCM Algorithm

```
1. Fix the cluster number $J = 2$, the sub-cluster number $K = 2$, initialize fuzzy membership, sub-membership and then select initial cluster center.
2. Set the loop counter $l = 0$.
3. Update the new cluster center using Eq. (9)
4. Update the fuzzy membership function using Eq. (7) and Eq. (10).
5. Update the fuzzy membership function using Eq. (8) and Eq. (11).
6. Terminate the iterations if the object function converges; otherwise, increase the iteration ($l = l + 1$) and repeat Steps 3 through 6.
```

3. Implementation and Experiment Results

In this section, we experimentally evaluate our proposed GHFCM in a set of synthetic images and real images. The neighborhood window size of GHFCM is set as 5×5.

Figure 1 demonstrates the outputs at first step of our method for three real mammograms. The first column shows the original mammogram with breast mass. The second column shows the corresponding ROI generated by ISODATA clustering and patch merging. The advantage of our ROI selected method is that the ROI is adaptive to the breast mass. The ROI which is used to replace the whole image domain for breast mass segmentation can improve the segmentation efficiency.

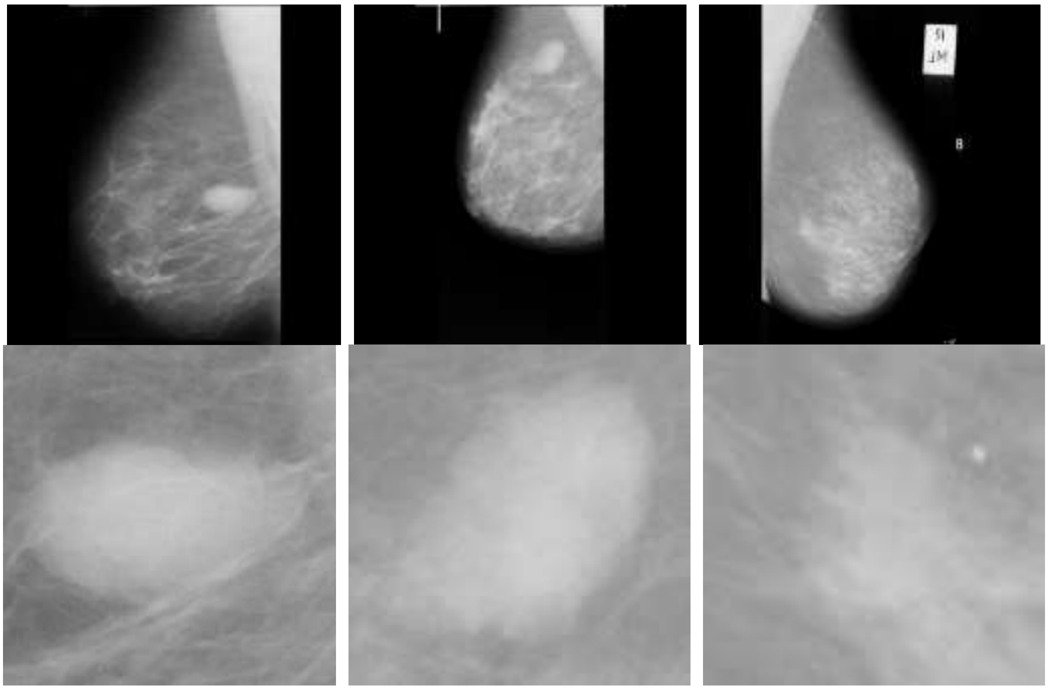
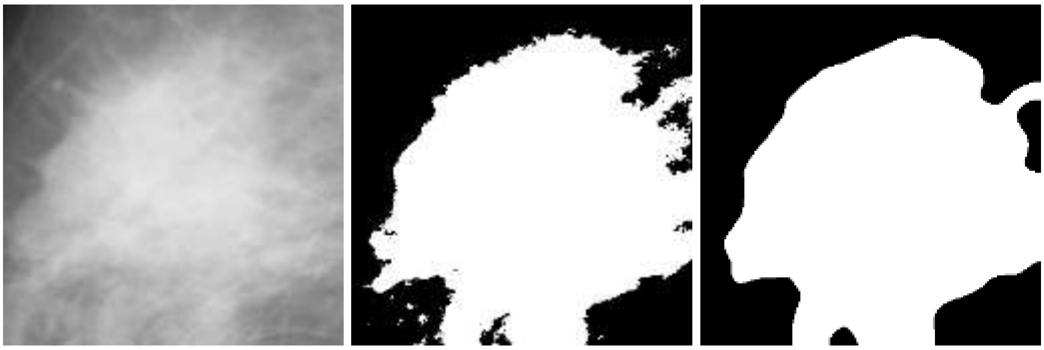


Figure 1. ROI Generation using ISODATA Clustering and Patch Merging

Figure 2 demonstrates the segmentation results of our method and the comparison with Gaussian mixture model (GMM). The first column demonstrates the ROIs of four mammograms from MiniMIAS dataset with different mass shapes. The second column shows the segmentation results obtained by GMM. From these results, we can see that the GMM model which is sensitive to the noise and outliers, cannot obtain a smooth boundary of breast mass. The third columns show the segmentation results of our method. These segmentation results indicate that our method which integrates the spatial constrain and hierarchical clustering strategy is robust to noise and outliers and is able to obtain a very smooth boundary of breast mass. For quantitative analysis, we calculate the Dice score of each method on eight original mammograms, the related results are shown in Figure 3. We can see that the Dice value of each segmentation result obtained by our method is much higher than that of result obtained by GMM, which indicates the superior performance of our segmentation method for breast mass segmentation.



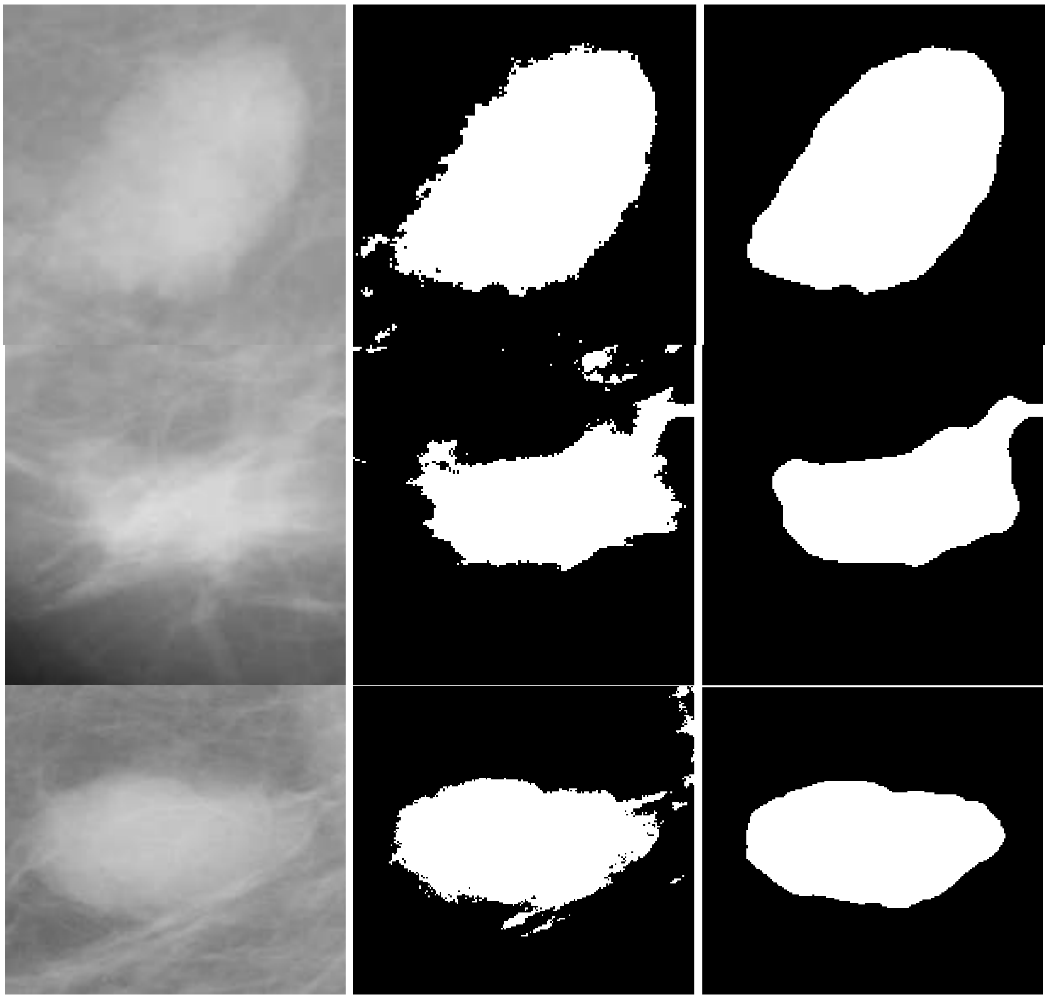


Figure 2. Comparison Results of Breast Mass Segmentation

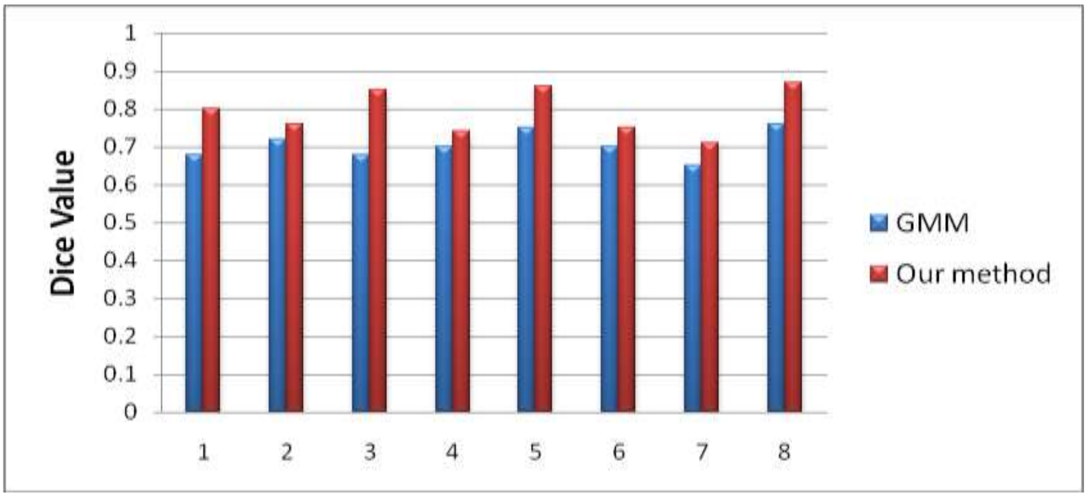


Figure 3. Dice Values Comparison of GMM and our Method

4. Conclusions

In this paper, we propose an automatic breast mass segmentation method based on patch merging method and GHFCM. The ISODATA clustering and patch merging method is used to obtain the adaptive ROI and initial segmentation results, while the GHFCM method which is able to overcome the drawbacks of effect of image noise and Euclidean distance FCM which is sensitive to outliers is used to obtain the precisely mass segmentation results. The new method is evaluated over MiniMIAS dataset. The segmentation performance from experimentations demonstrates that our method outperforms the other compared methods.

ACKNOWLEDGEMENTS

This paper is a revised and expanded version of a paper entitled “Breast mass segmentation in mammograms by using patch merging and GHFCM” presented at CIA 2015, Yeosu, Korea, May 21-23, 2015. This work was supported by the Industrial Strategic Technology Development Program (10041740) funded by the Ministry of Trade, Industry and Energy (MOTIE) Korea. It was also supported by Natural Science Foundation of Jiangsu Province (No.BY2014007-04). Prof. Jeong-Uk Kim is the corresponding author.

References

[1] American Cancer S: Breast cancer facts & figures 2007-2008. American Cancer Society Atlanta, GA; 2007.

[2] Liu J, Chen J, Liu X, et al. Mass segmentation using a combined method for cancer detection[J]. BMC systems biology, 2011, 5(Suppl 3): S6.

[3] Berber T, Alpkocak A, Balci P, et al. Breast mass contour segmentation algorithm in digital mammograms[J]. Computer methods and programs in biomedicine, 2013, 110(2): 150-159.

[4] Chan HP, Sahiner B, Helvie MA, Petrick N, Roubidoux MA, Wilson TE, Adler DD, Paramagul C, Newman JS, Sanjay-Gopal S: Improvement of radiologists’ characterization of mammographic masses by using computer-aided diagnosis: an ROC study. Radiology 1999, 212(3):817-827.

[5] Liu, X., & Tang, J. (2013). Mass classification in mammograms using selected geometry and texture features, and a new SVM-based feature selection method. IEEE Systems Journal, 8(3), 910–920.

[6] Rouhi R, Jafari M, Kasaei S, et al. Benign and malignant breast tumors classification based on region growing and CNN segmentation[J]. Expert Systems with Applications, 2015, 42(3): 990-1002.

[7] Hong R, Wang M, Gao Y, et al. Image annotation by multiple-instance learning with discriminative feature mapping and selection[J]. Cybernetics, IEEE Transactions on, 2014, 44(5): 669-680.

[8] Hong R, Pan J, Hao S, et al. Image quality assessment based on matching pursuit[J]. Information Sciences, 2014, 273: 196-211.

[9] Hong R, Cao W, Pang J, et al. Directional projection based image fusion quality metric[J]. Information Sciences, 2014, 281: 611-619.

[10] Hong R, Tang L, Hu J, et al. Advertising object in web videos[J]. Neurocomputing, 2013, 119: 118-124.

[11] Hong R, Wang M, Li G, et al. Multimedia question answering[J]. IEEE MultiMedia, 2012, 19(4): 72-78.

[12] Huo Z, Giger M L, Vyborny C J, et al. Analysis of spiculation in the computerized classification of mammographic masses[J]. Medical Physics, 1995, 22(10): 1569-1579.

[13] Kobatake H, Yoshinaga Y. Detection of spicules on mammogram based on skeleton analysis[J]. Medical Imaging, IEEE Transactions on, 1996, 15(3): 235-245.

[14] Lou S L, Lin H D, Lin K P, et al. Automatic breast region extraction from digital mammograms for PACS and telemammography applications[J]. Computerized Medical Imaging and Graphics, 2000, 24(4): 205-220.

[15] Tao Y, Lo S C B, Freedman M T, et al. Multilevel learning-based segmentation of ill-defined and spiculated masses in mammograms[J]. Medical physics, 2010, 37(11): 5993-6002.

[16] Berber T, Alpkocak A, Balci P, et al. Breast mass contour segmentation algorithm in digital mammograms[J]. Computer methods and programs in biomedicine, 2013, 110(2): 150-159.

[17] Song E, Jiang L, Jin R, et al. Breast mass segmentation in mammography using plane fitting and dynamic programming[J]. Academic radiology, 2009, 16(7): 826-835.

[18] Chatzis S P, Varvarigou T A. A fuzzy clustering approach toward hidden Markov random field models for enhanced spatially constrained image segmentation[J]. Fuzzy Systems, IEEE Transactions on, 2008, 16(5): 1351-1361.

[19] Qamar U. A dissimilarity measure based Fuzzy c-means (FCM) clustering algorithm[J]. Journal of Intelligent and Fuzzy Systems, 2014, 26(1): 229-238.

[20] Zheng Y, Jeon B, Xu D, et al. Image segmentation by generalized hierarchical fuzzy C-means algorithm[J]. Journal of Intelligent and Fuzzy Systems, 2015, 28:961-973.

[21] Salmon J, Strozecki Y. From patches to pixels in Non-Local methods: Weighted-average reprojection[C]//ICIP. 2010, 26: 121.

[22] Efros A A, Leung T K. Texture synthesis by non-parametric sampling[C]//Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on. IEEE, 1999, 2: 1033-1038.

[23] Criminisi A, Perez P, Toyama K. Object removal by exemplar-based inpainting[C]//Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on. IEEE, 2003, 2: II-721-II-728 vol. 2.

[24] Buades A, Coll B, Morel J M. Nonlocal image and movie denoising[J]. International journal of computer vision, 2008, 76(2): 123-139.

[25] Jain A K, Murty M N, Flynn P J. Data clustering: a review[J]. ACM computing surveys (CSUR), 1999, 31(3): 264-323.

[26] Tarabalka Y, Benediktsson J A, Chanussot J. Spectral–spatial classification of hyperspectral imagery based on partitional clustering techniques[J]. Geoscience and Remote Sensing, IEEE Transactions on, 2009, 47(8): 2973-2987.

[27] Jin Wang, Jeong-Uk Kim, Lei Shu, Yu Niu and Sungyoung Lee, A distance-based energy aware routing algorithm for wireless sensor networks, Sensors, Vol.10, No.10, 2010, pp.9493-9511.

[28] Jin Wang, Yue Yin, Jianwei Zhang, Sungyoung Lee, and R. Simon Sherratt, Mobility based energy efficient and multi-sink algorithms for consumer home networks, IEEE Transactions on Consumer Electronics, Vol.59, No.1, Feb. 2013, pp.77-84.

[29] Richang Hong, Jianxin Pan, Shijie Hao, Meng Wang, Feng Xue, Xindong Wu: Image quality assessment based on matching pursuit. Inf. Sci., 2014, pp.196-211.

[30] Richang Hong, Wenyi Cao, Jianxin Pang, Jianguo Jiang: Directional projection based image fusion quality metric. Inf. Sci., 2014, pp.611-619.

Authors

Shenghua Gu, Research Assistant of Jiangsu Key Laboratory of Big Data Analysis Technology, Nanjing University of Information Science and Technology. His main research interests include image processing, pattern recognition.

Email: gushenghuamath@163.com

Yunjie Chen, Associate Professor at the School of math and Statistic, Nanjing University of Information Science and Technology. His research interests include image processing, computer vision, medical imaging, and applied mathematics.