ICIC Express Letters
Part B: Applications
Volume **7**, Number **5**, May **2016**

ICIC International ©2016 ISSN 2185-2766
pp. **961–967**

MAMMOGRAPHY IMAGE CLASSIFICATION AND CLUSTERING USING SUPPORT VECTOR MACHINE AND K-MEANS

Kedkarn Chaiyakhan, Nittaya Kerdprasop and Kittisak Kerdprasop

School of Computer Engineering
Suranaree University of Technology
111 University Avenue, Nakhon Ratchasima 30000, Thailand
kedkarnc@hotmail.com; {nittaya; kerdpras}@sut.ac.th

Received October 2015; accepted January 2016

Abstract. *Mammography is an extraordinary type of low-powered x-ray process that provides detailed images of the internal structure of the breast. An early detection of breast cancer by means of mammography results in a successful treatment. Many researches show that the dense masses in the breast density are one of the strongest indicators of breast cancer developing. In this paper, we propose an approach to automatically appraise the density and contrast of breast images using gamma correction to increase the intensity of dense pixels with light intensity and vice versa to decrease the sparse intensity pixels showing dark intensity. In the segmentation process, we use region growing technique to get region of interest. We also extract three important features including texture, shape, and intensity histogram. In the classification process, we use SVM to classify tumor into two classes: malignant and benign. Moreover, we also compare the SVM classification result to the Naïve Bays and artificial neural network techniques. In clustering process, we use the k-means algorithm to cluster image into 2 categories: malignant and benign. The results of classification and clustering show that our proposed work can classify and cluster two types of mammography images after the appropriate application of gamma correction feature extraction process.*

Keywords: Image segmentation, Image classification, Image clustering, k-means, Support vector machine

1. **Introduction.** Breast cancer is a dangerous type of tumor originated from breast tissue. The most effective way to detect breast cancer is through the breast mammogram screening. However, the major limitation for mammography diagnosis is its sensitivity because interpreting mammography is a labor-intensive task for radiologists who cannot always offer stable results during interpreting. Many methodologies have thus been proposed to solve this uncertain interpretation problem by providing assistance to the advanced cancer detection and diagnosis tools.

The statistical approach has been proposed [1]. The authors provide connected density clusters taking the spatial information of the breast tissue into account. Quantitative and qualitative results show that their approach is able to correctly detect dense breasts apart from other tissue types. A methodology that is based on modeling a set of patched of either fatty or dense parenchyma using statistical analysis has been presented [2]. The two strategies, PCA and linear-discriminant analysis, are applied in the modeling process. In the work of [3], they use mixtures of Gaussian for modeling and segmenting the breast into four and five regions, respectively. However, these approaches do not take spatial information into account resulting in too many disconnected regions. Thus, the work of [4] has included a fuzzy affinity function in their proposed method, while [5] employs textural features to take the spatial distribution of the pixel and its neighborhood. Some researchers [6,7] use region growing, which is the region-based segmentation method. In the work of [8], they use region growing method based on the gradients and variances along

and inside of the boundary curve. Some researchers use edge and smoothness factors as criteria to determine initial seed points and then seeded region growing method is used to segment images based on seed regions [9].

In our proposed method, we use gamma correction to enhance the image contrast. In segmentation process, we use a well-known region growing method to find the ROI and then crop the image to consider only the tumor region. The unnecessary background has been removed in this process. After that we extract three types of feature and input digital data to the classification and clustering process. The performance of the proposed image classification approach has been evaluated by comparing the accuracy with some state of the art classification algorithms.

2. **Proposed Work.** In the proposed work, we have divided our process into five main parts: image preprocessing, segmentation, feature extraction, classification, and clustering. Figure 1 shows the framework of this research.

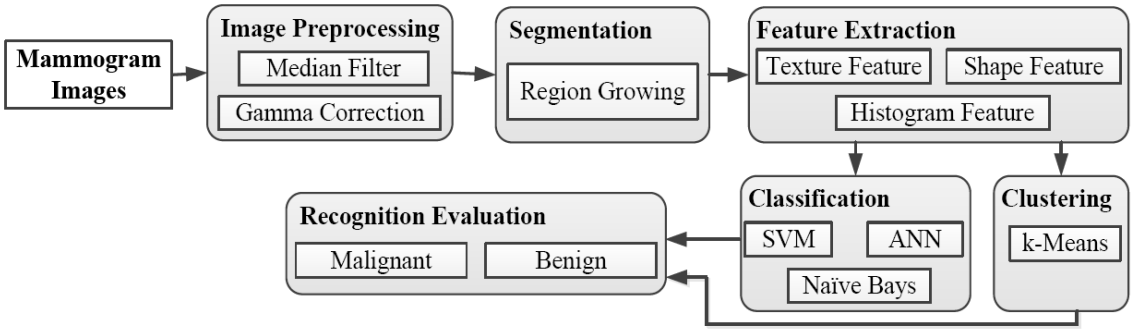


FIGURE 1. The framework of the proposed tumor recognition system

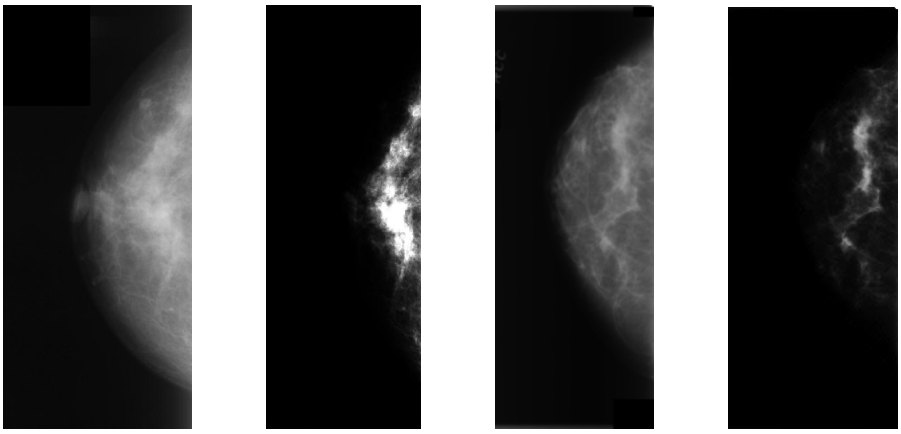
2.1. **Image preprocessing.** Mammogram images usually contain noises because of disturbances like Gaussian noise or some little darkness and brightness noise called salt and pepper noise. We use median filter to remove these noises. The output of this de-noising step is the clear images that are appropriate for further processing.

The next step of image preprocessing is image enhancement. We adjust the brightness and darkness of images using gamma correction algorithm. Figure 2 shows the original images of malignant and benign cases comparing to the improved results after applying the gamma correction technique. The gamma correction helps contrasting the tumor area from the fatty area.

2.2. **Segmentation.** This process separates the tumor areas from the background tissue in mammogram images. In this step, we apply the region growing segmentation method. Region growing is a region-based method starting with selecting seed points in the image, then propagating seeds until the specified stopping criteria are satisfied. Appropriate seed point selection is important. Therefore, in our proposed work, we select seed point using the centroid of object computed from area and position of object (centroid), as shown in Equation (1).

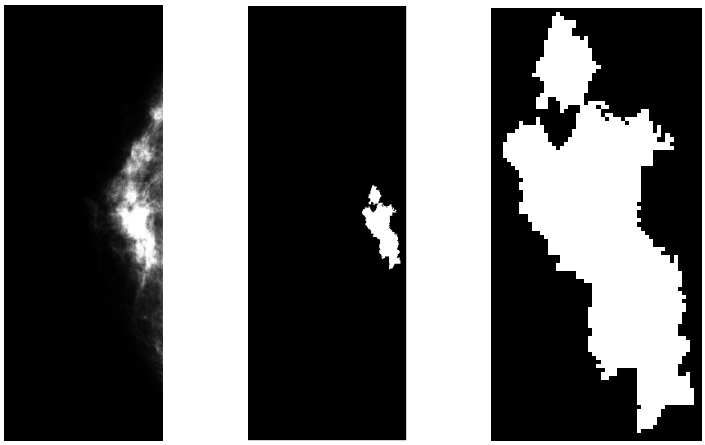
$$Centroid \quad \bar{x} = \frac{\sum_i \sum_j jW[i,j]}{Area} \qquad \bar{y} = \frac{\sum_i \sum_j iW[i,j]}{Area} \tag{1}$$

where W is the white pixel in the image, $Area$ is summation of white pixels, and i, j are the position of white pixel. After the region growing process, we will get the region of interest (ROI, white pixels) and then we crop only the ROI (Figure 3) removing background that may affect the classification and clustering process.



(a) (b) (c) (d)

FIGURE 2. Breast tumor images: (a) original malignant case, (b) malignant image after gamma correction, (c) original benign case, (d) benign image after gamma correction



(a) (b) (c)

FIGURE 3. The result of region growing and the cropped image: (a) gamma corrected image, (b) the image after applying region growing technique, (c) cropped image

2.3. **Feature extraction.** In this work, we extract three types of features: texture, shape, and intensity histogram features.

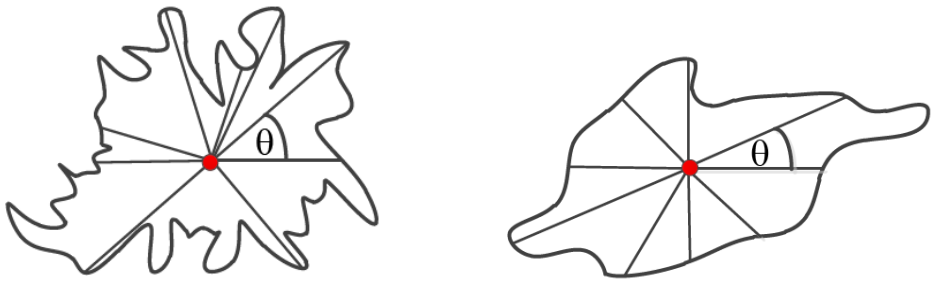
1). Texture Features

Texture is one of the important features used in identifying objects in an image. Texture features are based on the gray-level co-occurrence matrix (GLCM). The GLCM function characterizes the texture of an image by calculating how often pairs of pixels with specific values and in a specified spatial relationship occur in an image. We create a GLCM, and then extract from the matrix statistical measures such as contrast, correlation, energy, and homogeneity.

2). Shape Features

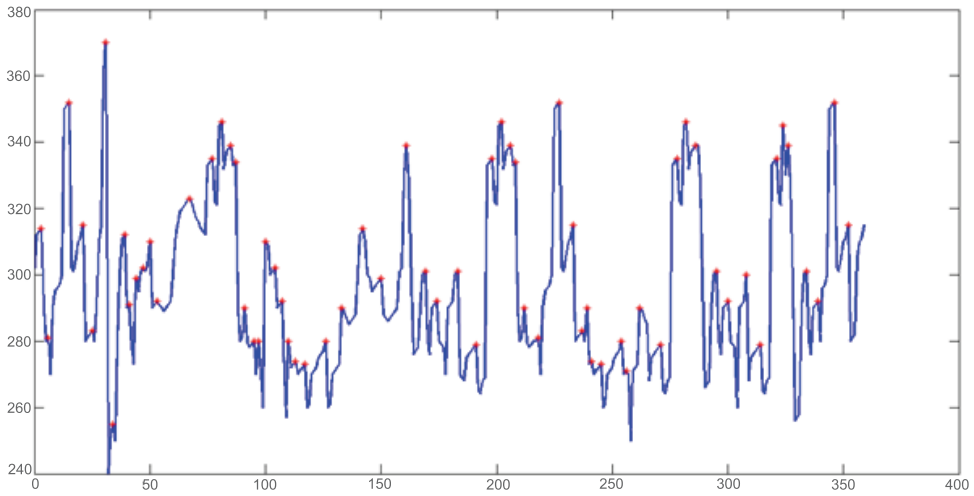
We extract shape feature using the percentage of curvature. First we drag lines from centroid to every edge pixel and then measure distance and angle from centroid to every edge pixel. After that, we plot the graph with angle along the x-axis and distance on the y-axis. From the graph, we can notice difference of curvature because of the distinct shape of malignant and benign tumors. We also do the normalization to find the percentage of

curvature. As a result, we get the different percentage of curvature between malignant and benign cases. We observe that malignant tumor shows many curves along its contour and we can get the percentage of peak in this graph. On the contrary, benign tumor has less curves than the malignant contour. Figure 4 illustrates example of curvature measurement. Figure 5 shows the different graph of curvature between malignant and benign contours.

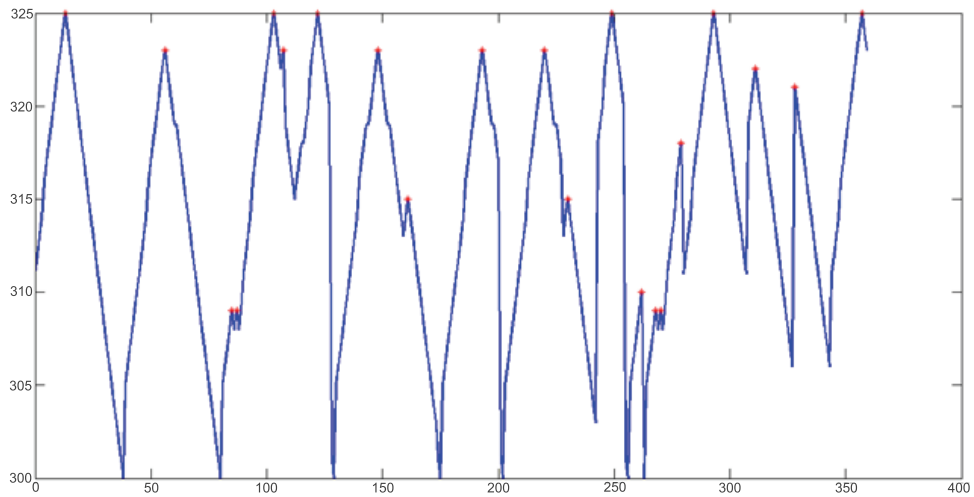


(a) (b)

FIGURE 4. Measuring the curvature: (a) malignant shape, (b) benign shape



(a)



(b)

FIGURE 5. Graph of curvature: (a) malignant contour, (b) benign contour

3). Intensity Histogram Features

The shape of the intensity histogram features provides much information to describe the properties of the image. Six statistic features obtained from the histogram are mean, variance, skewness, kurtosis, energy, and entropy. The mean is the average intensity level, whereas the variance is the variation of intensities around the mean. The skewness shows whether the histogram is symmetric. The histogram is symmetrical if the skewness is zero.

2.4. **Classification.** We use support vector machine (SVM) with radial basis function (RBF) kernel to classify the mammogram images using three features including texture, shape (percentage of curvature), and intensity histogram. In the SVM training process, we train SVM with 56 images (70% of 80 images selected from the DDSM database). In the evaluation process, the rest 24 images are used for testing. Training and testing images have been preprocessed through the same steps. We also use Naïve Bays and artificial neural network (ANN) in the classification process to compare the performance with SVM.

2.5. **Clustering.** In the clustering process, we use k-means algorithm (k = 2) to cluster the mammogram images. We also use the same three features (texture, shape, intensity histogram) as in the classification process. By means of this feature section process, we have noticed that k-means can accurately cluster images into the correct class.

3. **Experimental Results.** In this proposed work, we use data set from DDSM (digital database for screening mammography). We have selected from DDSM 80 images that include both cases of tumor, that is, malignant and benign (each case containing 40 images). This work has been implemented using MATLAB and R language. We run our experiments on a core i5/2.4 GHZ computer with 4 GB RAM. In the classification process, we compare our proposed method using SVM with Naïve Bays and ANN.

It can be noticed from the classification results summarized in Table 1 that the accuracy on recognizing the benign and malignant images of the SVM (with RBF kernel) shows the highest rate at 88.75%. In other two classification algorithms using Naïve Bays and ANN, the accuracy are 82.50% and 86.25%, respectively. We can conclude from this result that our proposed work using three types of feature and SVM classification has a higher accuracy than Naïve Bays and ANN. We also show in Figure 6 the area under curve (AUC) of the three classifiers: SVM, Naïve Bays, and ANN. As a result, SVM, Naïve Bays, and ANN show AUC value as 0.87, 0.83, and 0.85, respectively. The AUC closer to 1 is the better.

Table 1. Classification results for three learning algorithms

	Accuracy (%)	AUC
SVM (with RBF kernel)	**88.75**	**0.87**
Naïve Bays	82.50	0.83
ANN (artificial neural network)	86.25	0.85

From the result of clustering process using k-means with k = 2 (according to the two classes of images: benign and malignant) which is illustrated in Table 2, we obtain the image recognition accuracy as high as 90.00%. This means that k-means clustering can cluster the data to their actual class accurately. This good clustering result may be due to the effect of image preprocessing steps and the proper setting of cluster number.

Figure 7(a) demonstrates the plot of two cluster components: malignant and benign cases. The two-dimensional clustering plot of the two clusters and lines show the distance between clusters. Clustering shows a good result because it can clearly separate two

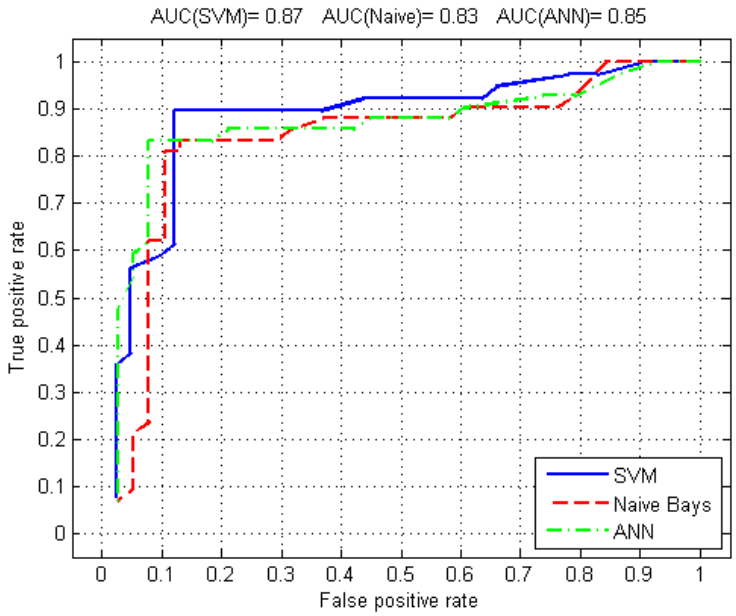
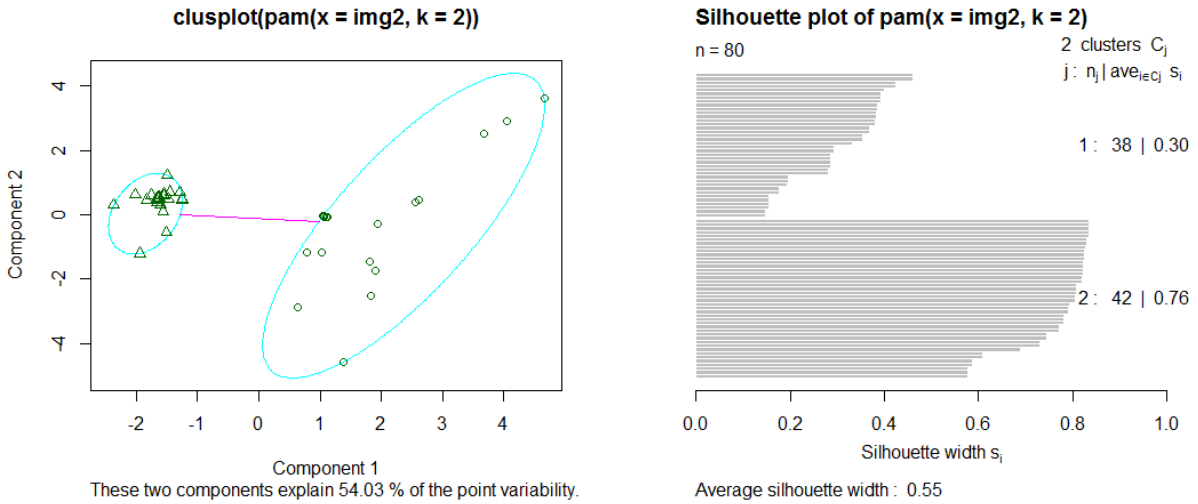


FIGURE 6. Area under curve of three classifiers

TABLE 2. Clustering result using k-means

	Benign	**Malignant**
Cluster 1 (Benign)	**35**	3
Cluster 2 (Malignant)	5	**37**



(a) (b)

FIGURE 7. k-means clustering results: (a) two components of clustering plot, (b) silhouette plot when k = 2

clusters, corresponding to the correct two classes. From the silhouette plot in Figure 7(b), the width of clusters, S_i, are 0.30 and 0.76. The average silhouette width is 0.55.

4. **Conclusions.** Mammography classification using support vector machine with image enhancement and three types of extracted features that we proposed in our framework is the main contribution of this paper. Mammography images are obtained from the well-known DDSM database. Image enhancement using gamma correction can improve

contrast of mammogram images to be seen clearly. We extract the region of interest (ROI) using region growing that can help the cropping of only the tumor object and at the same time eliminate the unnecessary background. After the ROI extraction, the three types of image features including texture, shape, and intensity histogram can be constructed. The processed images are then sent as input to the classification process using SVM with RBF kernel. The classification accuracy of SVM (88.75%) is higher than the ANN (86.25%) and Naïve Bays (82.50%) classifiers.

We also apply exactly the same image preprocessing steps but change from the classification algorithms to be the k-means clustering. We have found that k-means can cluster the mammography images correctly. It clusters images into a group of malignant and benign cases with the accuracy as high as 90.00%.

Acknowledgment. The authors would like to express grateful thanks to the reviewers for their useful comments for improving the content and readability of the paper. The first author has been supported by grant from Rajamangala University of Technology Isan.

REFERENCES

[1] A. Oliver, X. Llado, E. Perez, J. Pont, E. Denton, J. Freixener and J. Marti, A statistical approach for breast density segmentation, *Journal of Digital Imaging*, vol.23, no.5, pp.527-537, 2010.

[2] D. Brzakovic, N. Vujovic, M. Neskovic, P. Brzakovic and K. Fogarty, An approach to automated detection of tumors in mammograms, *IEEE Transactions on Medical Image*, vol.9, no.3, pp.233-241, 1990.

[3] S. R. Aylward, B. H. Hemminger and E. D. Pisano, Mixture modeling for digital mammogram display and analysis, *International Workshop in Digital Mammography*, pp.305-312, 1998.

[4] P. K. Saha, J. K. Udupa, E. F. Conant, P. Chakraborty and D. Sullivan, Breast tissue density quantification via digitized mammograms, *IEEE Transactions on Medical Image*, vol.20, no.8, pp.792-803, 2001.

[5] R. Zwiggelaar and E. Denton, Optimal segmentation of mammographic images, *International Workshop in Digital Mammography*, pp.751-757, 2004.

[6] C. H. Wei, S. Y. Chen and X. Liu, Mammogram retrieval on similar mass lesions, *Computer Methods and Programs in Biomedicine*, vol.106, no.3, pp.234-248, 2012.

[7] R. Adam and L. Bischof, Seeded region growing, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.16, pp.641-647, 1994.

[8] W. Deng, W. Xiao, H. Deng and J. Liu, MRI brain tumor segmentation with region growing method based on the gradients and variances along and inside of the boundary curve, *International Conference on Biomedical Engineering and Informatics*, vol.1, pp.393-396, 2010.

[9] C. Huang, Q. Liu and X. Li, Color image segmentation by seeded region growing and region merging, *International Conference on Fuzzy Systems and Knowledge Discovery*, vol.2, pp.533-536, 2010.