ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology
Volume 1, Issue 5, July 2012

Gabor Filter, PCA and SVM Based Breast Tissue Analysis and Classification

Pravin S. Hajare, Vaibhav V. Dixit

Abstract—Early signs of breast cancer are revealed in Mammographic images. This experiment focuses towards the identification of relevant, representative and more important, discriminate image features for analysis of medical images. The images are taken as an input for further processing from MIAS database after which Gabor filter is used to extract intensity features and the patches are obtained to recognize whether the mammogram image is normal benign or malign. The images are further processed using Principal Component Analysis (PCA) to reduce data dimensionality. Finally, the extracted features are classified using the proximal support vector machines as classifier. The Gabor filter is used with four orientations. Also, two different frequencies of Gabor filters are used. Finally, the recognition rate of all orientations and different frequencies are calculated and compared. Also, PCA is directly applied to the unfiltered images and these results are compared with the results of Gabor + PCA. The Gabor filter with low frequency and all orientation gives the highest recognition rate of 84.375%.

Index Terms— Breast cancer, Mammography, Gabor wavelets, PCA, SVM

I. INTRODUCTION

Mammography is at present the best available technique for early detection of breast cancer. In mammographic images early signs of breast cancer, such as bilateral asymmetry, can be revealed. Bilateral asymmetry is asymmetry of the breast parenchyma between corresponding regions in left and right breast. The most common breast abnormalities that may indicate breast cancer are masses and calcifications. Early detection and treatment are considered as the most promising approaches to reduce breast cancer mortality. Mammogram image is considered as the most reliable, low cost, and highly sensitive technique for detecting small lesions. One of the main points that should be taken under serious consideration when implementing a robust classifier for recognizing breast tissue is the selection of the appropriate features that describes and highlight the differences between the abnormal and the normal tissue in an ample way. Feature extraction is an important factor that directly affects the classification result in mammogram classification. Most systems extract features to detect and classify the abnormality as benign or malignant from the textures. A particular image type is given by mammographic images that are typically X-ray captures of breast region displaying points with high intensities density that are suspected of being potential tumors. Early diagnostic and screening is crucial for having a appearing in the mammogram images could indicate a potential presence of a benign or malignant tumor.

Manuscript received May, 2012.
Pravin S. Hajare*, Department of E & Tc, University of Pune, Sinhgad College of Engineering., (e-mail: hajarepravin24@gmail.com). Pune, India , Mobile No: 9822918392* ***Vaibhav V. Dixit****, Department of E & Tc, University of Pune, Sinhgad College of Engineering., (e-mail: vvdixit.scoe@sinhgad.edu) Pune, India Mobile No.9822777265*

II. DATABASE (MIAS)

The experimentation is done with the database images taken from Mammographic Image Analysis Society (MIAS), which contains 322 samples belonging to three different categories as normal, benign and malign. The database consists of 208 normal images, 63 benign and 51 malign cases, which are considered abnormal [12]. These database images are of 1024 x 1024 pixel size and having background information like breast contour, therefore the pre-processing of these images is required. To obtain region of interest, 140 × 140 patches are extracted from mammogram images as shown in figure 1 and figure 2. The database is with two different sets. First set is having 80% database images of whole database with known classes, normal, benign and malign. Whereas the second set is with 20% database images which are the test images and are having unknown classes. This experiment uses 258 training images and 64 testing images from mammogram database which are with all the classes, normal, benign and malign.

Figure 1. mdb001.pgm (Mammogram Image)

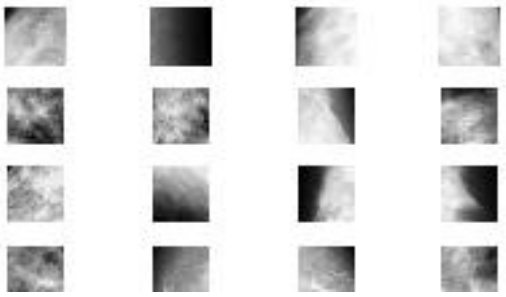

Figure 2 Patches of 140 X140 pixels extracted from Mammographic images.

To reduce the computations during the further processing the images are down sampled to the size of 30x30 pixels. The patches from image are extracted based on the intensity variations. The affected area is with light intensity. Also, the database provides the information of all three classes, normal, benign and malign.

III. Feature Extraction

In order to provide accurate recognition, feature patterns must be extracted. Only the significant features must be encoded. In this experimentation, the method used to extract the intensity features is the Gabor filter with low and high frequencies and also with four different orientations. Three low frequencies and three high frequencies with four orientations give 12 combinations of Gabor filter. Thus, the mammographic image is passed through 12 Gabor filters and magnitudes of all are represented.

A. Gabor Wavelets

A 2-D Gabor function is a Gaussian modulated by a sinusoid. It is a non orthogonal wavelet. Gabor filters exhibits the properties as the elementary functions are suitable for modeling simple cells in visual cortex [11] and gives optimal joint resolution in both space and frequency, suggesting simultaneously analysis in both domains. The definition of complex Gabor filter is defined as the product of a Gaussian kernel with a complex sinusoid. A 2D Gabor wavelet transform is defined as the convolution of the image I (z).

$$J_k(z) = \int \int I(z') \psi_k (z - z') dz' \quad (1)$$

with a family of Gabor filters:
ψk

$$\Psi_k(z) = k^T k/\sigma^2 \exp((-k^T k/2\sigma^2) * z^T z) (\exp(ik^T z) - \exp(-\sigma^2/2)) \quad (2)$$

Where, z = x, y and k is characteristic wave vector:

$$K = (k_v \cos\varphi_\mu \quad k_v \sin\varphi_\mu)^T \quad (3)$$

With,

$$Kv = 2 - v + \frac{2}{2}\pi, \varphi_\mu = \mu\frac{\pi}{8},$$

$$v = 0, 1, 2, 3, 4, \quad \mu = 0, \pi/4, \pi/2, 3\pi/4 \quad (4)$$

The results obtained by extracting the features with Gabor filters are as shown in figure 3 and 4. Fig. 3 shows the magnitude response of features with low frequency Gabor filter bank (v=2, 3, 4 and μ=0, π/4, π/2, 3π/4) whereas fig. 4 shows the magnitude response of high frequency filter bank (v=0, 1, 2 and μ=0, π/4, π/2, 3π/4).

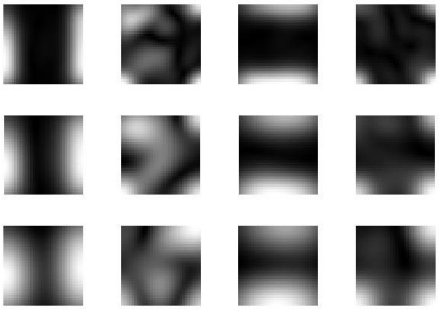

Figure 3. Magnitude of Gabor representation for one MIAS sample convolved with 12 Gabor filters Low frequency, v=2, 3, 4 and μ=0, π/4, π/2, 3π/4

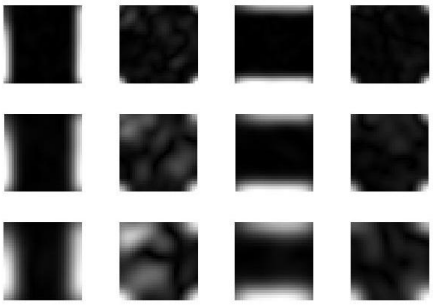

Figure 4. Magnitude of Gabor representation for one MIAS sample convolved with 12 Gabor filters High frequency, v=0, 1, 2 and μ=0, π/4, π/2, 3π/4

IV. Principal Component Analysis

PCA involves the calculation of the Eigen value decomposition of a data covariance matrix or singular value decomposition of a data matrix, usually after mean centering the data for each attribute. The results of a PCA are usually discussed in terms of component scores and loadings. PCA is the simplest of the true eigenvector-based multivariate analyses. Often, its operation can be thought of as revealing the internal structure of the data in a way which best explains the variance in the data. If a multivariate dataset is visualized as a set of coordinates in a high-dimensional data space, PCA supplies the user with a lower-dimensional picture, a "shadow" of this object when viewed from its most informative viewpoint. PCA is closely related to factor analysis; indeed, some statistical packages deliberately conflate the two techniques. True factor analysis makes different assumptions about the underlying structure and solves eigenvectors of a slightly different matrix. PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by any projection of the data comes to lie on the first coordinate the second greatest variance on the second coordinate, and so on [4]. PCA is theoretically the optimum transform for given data in least square terms. For a data matrix, XT, with zero empirical mean (the empirical mean of the distribution has been subtracted from the data set), where each row represents a

different repetition of the experiment, and each column gives the results from a particular probe.

Given a set of points in Euclidean space, the first principal component (the eigenvector with the largest Eigen value) corresponds to a line that passes through the mean and minimizes sum squared error with those points. The second principal component corresponds to the same concept after all correlation with the first principal component has been subtracted out from the points. Each Eigen value indicates the portion of the variance that is correlated with each eigenvector. Thus, the sum of all the Eigen values is equal to the sum squared distance of the points with their mean divided by the number of dimensions. PCA essentially rotates the set of points around their mean in order to align with the first few principal components. This moves as much of the variance as possible (using a linear transformation) into the first few dimensions. The values in the remaining dimensions, therefore, tend to be highly correlated and may be dropped with minimal loss of information. In this way, the PCA is used in this experiment for dimensionality reduction. Fig. 5 shows 10 Eigen images resulted from PCA. Thus, the 30 X 30 down sampled image is applied to PCA and total 900 dimensions are represented with 10 Eigen vectors. Gabor + PCA feature vectors that are actually used for classification, are formed by projecting the zero mean data into the PCA eigenvectors $V_r, i.e\ F^{kc}_{GaborPCA} = V_r^T X^{kc}$ where V_r is the r – rank. PCA is projection matrix and T represents the transpose operator. Employing PCA projection, the dimension is reduced from p to r, where $r << p$. The experiments were run for $r = \{5, 10, 20, 30 \ldots 150\}$.

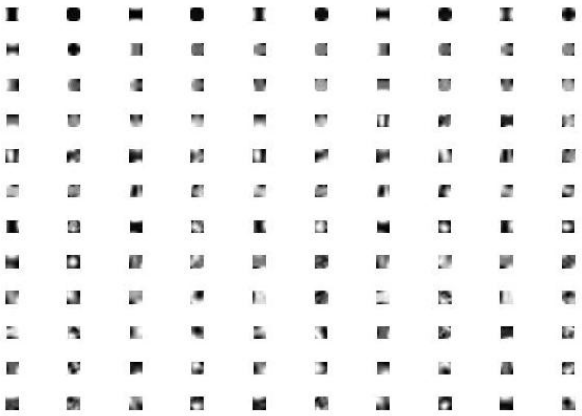

Figure 5. Each row depicts 10 Eigen images obtained by applying PCA for the Matrix Xkc and columns contain concatenated Gabor convolution results corresponding to 4 orientations and 3 frequencies (low frequency range).

V. Support Vector Machine

Support vector machines (SVMs) are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The standard SVM is a non-probabilistic binary linear classifier, i.e. it predicts, for each given input, which of two possible classes the input is a member of. Since an SVM is a classifier, then given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that predicts whether a new example falls into one category or the other. Intuitively, an SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

Multiclass SVM aims to assign labels to instances by using support vector machines, where the labels are drawn from a finite set of several elements [9]. The dominating approach for doing so is to reduce the single multiclass problem into multiple binary classification problems. Each of the problems yields a binary classifier, which is assumed to produce an output function that gives relatively large values for examples from the positive class and relatively small values for examples belonging to the negative class. Two common methods to build such binary classifiers are where each classifier distinguishes between (i) one of the labels to the rest (one-versus-all) or (ii) between every pair of classes (one-versus-one). Classification of new instances for one-versus-all case is done by a winner-takes-all strategy, in which the classifier with the highest output function assigns the class (it is important that the output functions be calibrated to produce comparable scores). For the one-versus-one approach, classification is done by a max-wins voting strategy, in which every classifier assigns the instance to one of the two classes, then the vote for the assigned class is increased by one vote, and finally the class with most votes determines the instance classification.

Here, in this experiment the SVM is trained with the images from training dataset whose classes are known. Total 258 training images are taken from dataset. Dataset has 64 testing images which are classified with SVM as benign, normal or malign.

TABLE 1. RECOGNITION RATE EXPRESSED IN % WITH DIFFERENT ORIENTATIONS. THE RANK (r) CORRESPONDING TO THE BEST RR IS GIVEN IN PARENTHESIS.

Features	Frequency range	Orientation	Recognition Rate (%)
Gabor filter and PCA	Low frequency range	0	65.62 (60)
		π/4	67.18 (20)
		π/2	68.75 (5)
		3π/4	64.06 (5)
		All	84.375 (10)
Gabor filter and PCA	High frequency range	0	65.62 (10)
		π/4	64.06 (5)
		π/2	71.87 (40)
		3π/4	64.06 (5)
		All	68.75 (20)
PCA	-	-	65.62 (30)

VI. CONCLUSION

The system works on two filter banks, low frequency and high frequency. Initially, the patches of 140 x 140 are extracted from mammographic images. The images are passed through 12 different Gabor filters. The features are obtained by convolving patches representing tumor or tumor-free areas with several Gabor filters and are employed for recognition purpose. The large dimension images are then down sampled to the size of 30 X 30 pixels. Also, these give large number of dimensions so applied to PCA to reduce the dimensionality. The results of different frequency ranges of Gabor filter coupled with PCA and different orientations are tabulated in table 1. Also, the result obtained by using the PCA directly is given. From all these results, Gabor features seem to posses more discriminative power than PCA features as Gabor filter with low frequency and all orientations gives the highest recognition rate of 84.375% among all.

REFERENCES

[1] Anna N. Karahaliou, Ioannis S. Boniatis, Spyros G. Skiadopoulos, Filippos N. Sakellaropoulos, Nikolaos S. Arikidis, Eleni A. Likaki, George S. Panayiotakis, and Lena I. Costaridou , "Breast Cancer Diagnosis:Analyzing Texture of Tissue Surrounding Microcalcifications ", 2008 *IEEE Transactions on Information Technology in Biomedicine, VOL. 12, NO. 6, November*

[2] Lori M. Bruce, Ravikiran Kalluri," An analysis of the contribution of scale in mammographic mass classification", *Proceedings - 19th International Conference - IEEE/EMBS Oct. 30 - Nov. 2, 1997 Chicago, IL. USA*

[3] Ibrahima Faye, Brahim Belhaouari Samir, Mohamed M. M. Eltoukhy," Digital Mammograms Classification Using a Wavelet Based", *2009 Second International Conference on Computer and Electrical Engineering Feature Extraction Method*

[4] Jie Luo, Bo Hu, Xie-Ting Ling, and Ruey-Wen Liu," Principal Independent Component Analysis ", *IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 10, NO. 4, JULY 1999*

[5] R. M. Rangayyan, R. J. Ferrari, J. E. L. Desautels, A. F. Fr`ere, "Directional analysis of images with Gabor wavelets", *In: Proc. of XIII Brazilian Symposium on Computer Graphics and Image Processing, SIBGRAPI*, pp. 170–177, 2000.

[6] R. N. Strickland, H. Hahn, "Wavelet Transforms for Detecting Microcalcifications in Mammograms", *IEEE Trans. on Medical Imaging, 15*, pp.218–229, 1996.

[7] H. S. Sheshadri, A. Kandaswamy,"Breast Tissue Classification Using Statistical Feature Extraction Of Mammograms", *Medical Imaging and Information Sciences, 23(3)*, pp. 105–107, 2006.

[8] Y. Sun, C. F. Babbs, E. J. Delp, "Normal Mammogram Classification based on Regional Analysis", *The 2002 45th Midwest Symposium on Circuits and Systems, 2*, pp. 375–378, 2002.

[9] L. Wei, Y. Yang, R. M. Nishikawa, Y. Jiang, "A Study on Several Machine-Learning Methods for Classification of Malignant and Benign Clustered Miycrocalcifications", *IEEE Trans. on Medical Imaging, 24(3)*, pp. 371–380, 2005.

[10] O. R. Zaiane, M. L. Antonie, A. Coman, "Mammography Classification by an Association Rule-based Classifier", *In Third International ACM SIGKDD workshop on multimedia data mining (MDM/KDD'2002) in conjunction with eighth ACM SIGKDD*, pp. 62–69, 2002.

[11] T. Lee, "Image Representation using 2d Gabor Wavelets," *IEEE Trans.Pattern Analysis and Machine Intelligence, vol. 18, no. 10*, pp. 959–971,1996.

[12] Mammographic Image Analysis Society, http://www.wiau.man.ac.uk/services/MIAS/MIAS web.html

[13] http://www.who.int/mediacentre/factsheets/fs297/en/ index.html

Pravin S. Hajare B.E. E & Tc, M.E Digital systems (Pursuing) from Sinhgad college of Engineering, Pune, M. S. India, has teaching experience of 8 years in Pune university M.S. India.

Vaibhav V. Dixit Assistant professor, Department of Engineering., Sinhgad college of Engineering., Pune M.S. India, Pursuing his PhD in Image processing, has working experience of 16 years.