Breast Masses Identification through Pixel-Based Texture Classification*

Jordina Torrents-Barrena[1], Domenec Puig[1,**], Maria Ferre[1], Jaime Melendez[2], Lorena Diez-Presa[3], Meritxell Arenas[3], and Joan Martí[4]

[1] Department of Computer Engineering and Mathematics, University Rovira i Virgili
domenec.puig@urv.cat
[2] Department of Radiology, Radboud University Medical Center
[3] Radiotherapic Oncology Research Group. Hospital Universitari Sant Joan de Reus
[4] Computer Vision and Robotics Research Institute, University of Girona

Abstract. Mammographic image analysis plays an important role in computer-aided breast cancer diagnosis. To improve the existing knowledge, this paper proposes a new efficient pixel-based methodology for tumor vs non-tumor classification. The proposed method firstly computes a Gabor feature pool from the mammogram. This feature set is calculated through multi-sized evaluation windows applied to the probabilistic distribution moments, in order to improve the accuracy of the whole system. To deal with a high dimensional data space and a large amount of features, we apply both a linear and non-linear pixel classification stage by using Support Vector Machines (SVMs). The randomness is encoded when training each SVM using randomly sample sets and, in consequence, randomly selected features from the whole feature bank obtained in the first stage. The proposed method has been validated using real mammographic images from well-known databases and its effectiveness is demonstrated in the experimental section.

Keywords: Texture feature extraction, Gabor filters, Support Vector Machine, mammographic images, pixel-based classification.

1 Introduction

Breast cancer among middle aged women is a significant public health problem in the world. At present, there are no effective ways to prevent it, because its cause is not yet fully known. Early detection is the key for improving cancer prognosis since the death rate can be significantly reduced. Mammography has been one of the most reliable methods for detecting breast carcinomas [1, 2], in its earliest and most treatable stage, so it continues to be the primary imaging modality for breast cancer screening and diagnosis. In addition, it allows the detection of

* This work was partly supported by the Spanish Government through projects TIN2012-37171-C02-01 and TIN2012-37171-C02-02.

** Corresponding author.

H. Fujita, T. Hara, and C. Muramatsu (Eds.): IWDM 2014, LNCS 8539, pp. 581–588, 2014.

other pathologies and may suggest the tumor nature such as normal, benign or malignant.

Nowadays, reading mammograms is a very demanding job for radiologists, who visually examine the images for the presence of deformities that can be associated as cancerous changes. Mammograms are hard to interpret because of the complex tissue morphology of the breast and the number of imaging parameters that affect its acquisition. For this reason, the radiologists judgments depend on their training, experience and subjective criteria. There are several lesions that are characteristics of breast cancer such as microcalcifications, masses, architectural distortions and bilateral asymmetry. Since some lesions are often indistinguishable, because they have similar features to normal mammary tissue, automated detection and classification is even more difficult.

Furthermore, some other diseases have similar patterns to the breast cancer, which challenges the diagnosis. Manual readings may result misdiagnosis due to human errors caused by visual fatigue. To improve accuracy and efficiency of screening mammography, computer aided techniques are introduced. Therefore, CAD systems have been shown to be a helpful tool [3]. They can provide an important contribution for breast cancer control by marking suspicious regions and detecting abnormalities, to decrease the death rate among women with this disease.

This paper presents a framework for tumor vs non-tumor identification founded on a pixel-based texture classification approach, which is broadly divided in two stages. In the first stage, texture features are extracted from both tumor and normal regions by using a Gabor filter bank. In the second stage, a pixel-based texture classification strategy by using SVMs is applied [4], which provides the probability of each pixel in the mammogram to belong to a tumor region.

The rest of the paper is organized as follows. In Section 2, we present the proposed methodology. First, we describe the feature extraction step and then, the pixel-based classification algorithm. Experiments are shown and discussed in Section 3. Finally, conclusions and further tasks are given in Section 4.

2 Proposed Methodology

The texture classification methodology proposed in this work is as follows. During an initial training stage, a set of prototype features is computed at every texture pattern of interest (normal/tumor). The training images associated with each pattern are first filtered by applying a multichannel Gabor filter bank, obtaining a cloud of texture feature vectors for every pattern [5–7]. A set of prototypes is then extracted in order to represent that cloud. During the evaluation stage of the classifier, a given test image is processed in order to identify the texture pattern corresponding to each of its pixels. This is done by first applying the multichannel Gabor filter bank to the test image. A feature vector is thus obtained for every pixel. Each vector is classified into one of the given texture patterns by a SVM-based classifier fed with the prototypes extracted during the training stage. The stages involved in this scheme are detailed below.

2.1 Texture Feature Extraction

Textural properties in an image can be used to detect different types of information such as edges, lines, spots, flat areas and other local patterns. Some of these properties can be observed in mammograms at different scales and orientations. For this reason, we define a Gabor filter bank in order to capture texture patterns in mammograms, as it has been shown to be optimal in the sense of minimizing the joint two-dimensional uncertainty in space and frequency.

Gabor Filters. A wide variety of texture feature extraction methods have been proposed in the literature. Among them, multichannel filtering techniques based on Gabor filters have received considerable attention. The texture feature extraction stage of this work is based on the optimized multichannel Gabor wavelet filters. The next paragraphs give a brief overview about them.

Gabor filters are biologically motivated convolution kernels that have enjoyed wide usage in a myriad of applications in the field of computer vision and image processing. In order to extract local spatial textural micro-patterns in mammogram ROIs, Gabor filters can be tuned with different orientations and scales, and thus provide powerful statistics which could be very useful for breast cancer detection.

A two-dimensional Gabor filter defined as a Gaussian kernel modulated by an oriented complex sinusoidal wave can be described as follows [5, 8, 6]:

$$g(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp^{-\frac{1}{2}(\frac{\tilde{x}^2}{\sigma_x^2}+\frac{\tilde{y}^2}{\sigma_y^2})} \exp^{2\pi j W\tilde{x}} \tag{1}$$

$$\tilde{x} = x\cdot\cos\theta + y\cdot\sin\theta \quad \textbf{and} \quad \tilde{y} = -x\cdot\sin\theta + y\cdot\cos\theta \tag{2}$$

where σ_x and σ_y are the scaling parameters of the filter and describe the neighborhood of a pixel where weighted summation takes place, W is the central frequency of the complex sinusoidal and $\theta \in [0,\pi)$ is the orientation of the normal to the parallel stripes of the Gabor function.

Evaluation of Texture Methods over Multisized Windows. The texture features that characterize each pixel and its surrounding neighborhood (window) are both the mean and standard deviation of the module of the Gabor wavelet coefficients. The Gabor filter bank has been configured with four scales and six orientations, and a range of frequencies between 0.05 and 0.4. The orientations and frequencies for a bank are calculated using the following equations:

$$orientation(i) = \frac{(i-1)\pi}{m} \quad \textbf{where} \quad i = 1,2,...,m \tag{3}$$

$$frequency(i) = \frac{f_{max=0.4}}{(\sqrt{2}^{i-1})} \quad \textbf{where} \quad i = 1,2,...,n \tag{4}$$

where m is the total number of orientations and n is the total number of frequencies. Therefore, every feature vector is composed by a total of 48 dimensions: $6(scales) \times 4(orientations) \times 2(mean, stdev)$.

The means and stdevs mentioned above are computed for W different window sizes. W is set to 3 in this case: 1×1, 33×33 and 51×51. Thus, W sets of feature vectors are generated for each pixel of the given texture patterns during the training stage, as well as for each pixel of the test image during the classification stage.

2.2 Supervised Pixel-Based Classification

Once the features characterizing both normal and tumor tissue have been extracted, the goal of this stage is to classify the pixels of an input test mammogram into one of the two patterns of interest (normal/tumor).

Support Vector Machine-Based Classifier. A classification problem encompasses the assignment of an unseen pattern to a predefined class, according to the characteristics of the pattern, presented in the form of a feature vector. However, a classifier needs to be trained in order to perform this task.

A way to efficiently summarize and learn all the available information obtained from the training set is through SVMs, since they are the most advanced ones, generally, designed to solve binary classification problems. SVM formulation is based on statistical learning theory [9, 10] and has attractive generalization capabilities in linear and non-linear decision problems. The classifier maps an M-dimensional data point into a class label based on an aggregating decision function. A supervised classification task involves separating data into training and test sets. Each instance in the training set contains the class label and the features. The goal of the SVM is to produce a model, based on the training data, which predicts the target values of the test data given only the test data features. Given a training set of instance-label pairs (x_i, y_i) , $i = 1, ..., l$, where $x_i \in R^n$ and $y \in \{1, -1\}^l$, the SVM casts the classification problem into an optimization problem. The training vectors x_i are mapped into a higher or infinite dimensional space. The SVM finds a linear separating hyperplane with the maximal margin in this higher dimensional space by using what is called the kernel trick.

The four basic kernel functions are linear, polynomial, sigmoid and radial, from which we only use two of them (linear and radial). For linearly separable problems, kernel function is simply the dot product of the two given points in the input space:

$$k(x_i, x) = x_i \cdot x \tag{5}$$

However, for non-linear problems, the original input space is mapped through a non-linear function, possibly making the data linearly separable, using different suitable kernels (for computational efficiency). In our experiments, RBF (radial basis function) kernel is used as given by:

$$k(x_i, x) = \exp^{(-\gamma \|x_i - x\|^2)}, \gamma > 0 \quad (6)$$

There are two parameters now tied with the RBF kernel: γ that represents the width of the kernel function, and C (a regularization parameter) who controls the trade-off between error of SVM and margin maximization [9].

2.3 Breast Masses Identification System

The block diagram of the breast cancer identification system is detailed below (see Fig. 1). The proposed system is composed of four main stages: pre-processing, feature extraction, feature selection and classification. Various existing approaches differ in the choice of techniques for these stages. Our proposed approach for feature extraction is robust against noise (this method has been described in detail in subsection 2.1). Then, we apply a significant selection of the features extracted. In the training mode we choose all the pixels inside the tumor region, by contrast, when the classifier is in the prediction phase, a random selection is applied in order to choose both some normal pixels and also pixels affected by the disease. In this way, the next step will use both types of information.

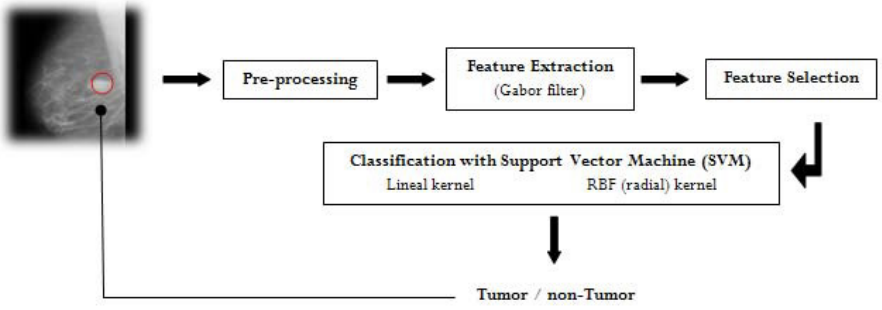

Fig. 1. Breast Masses Detection System

3 Experimentation

This section describes the materials used for the development and validation of the proposed technique, as well as the experimental results obtained through the application of the proposed methodology to a well-known mammogram database.

3.1 Materials

The algorithm proposed in this paper has been evaluated on the mammograms of the mini-MIAS database [11] that comprises 322 images of 1024×1024 pixels (e.g., Fig. 2, 1st column). Every image includes information about the existing

anomalies: it comprises the location of the lesion and the radius of the circle that roughly delimits the lesion (e.g., Fig. 2, 3rd column). We randomly selected several cases from this database which contain true and false masses (but with suspicious tissues). These ROIs are used for training and testing.

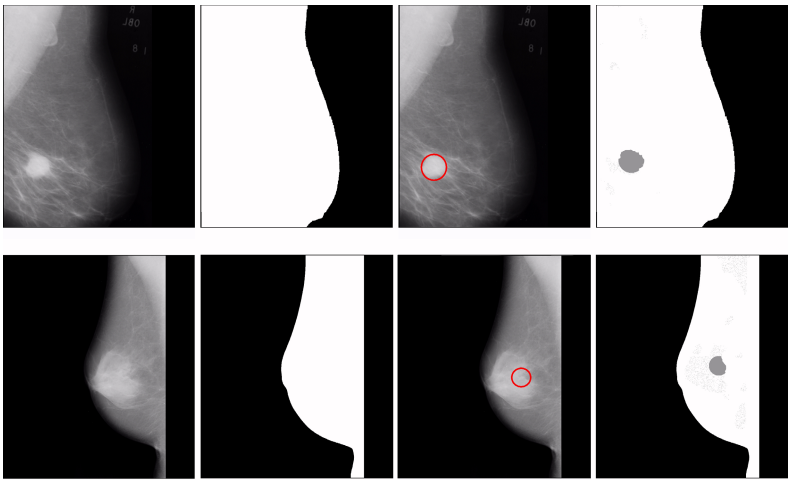

Fig. 2. Classification example: original test image (1st column), region of interest (2nd column), ground-truth: a red circle delimits the tumor region (3rd column), tumor pixels identification (4th column)

Our method has been implemented in Matlab by taking advantage of its high performance to develop computer vision and image processing software. In addition, the *LibSVM* library [4] has been used to implement the SMO algorithm for kernelized SVMs, supporting classification and regression.

3.2 Experimental Results

The convenient values of the SVM parameters to reach good accuracy ratios for discrimination between tumor and normal regions were found by means of an iterative procedure. Furthermore, our algorithm removes the background regions in the mammogram and focusses the tumor search in the breast region (e.g., Fig. 2, 2nd column). Finally, each pixel in the test mammogram is classified as belonging to tumor or normal region (e.g., grey pixels have been classified as belonging to a tumor in Fig. 2, 4th column).

Commonly used evaluation measures of the predictive ability of the breast cancer detection systems are *sensitivity* (a measure of true positive rate) and *specificity* (a measure of true negative rate) under *ROC curve*. In addition, the *F1* score also helps to determinate the effectiveness of our pixel-based classifier. We adopt these performance measures to evaluate the proposed system.

First of all, *F1* score can measure the discrimination capability of the classifier between the cancerous and normal regions (the closer the *F1* to 1, the better the tumor identification). The *F1* score is defined as: $F1 = 2TP/2TP + FN + FP$, where *TP*, *FN* and *FP* are the number of true positives, false negatives and false positives respectively. The average results corresponding to the mini-MIAS database [11] classification are shown in Table 1. Notice that, especially the SVM classifier based on the radial kernel produces good identification ratios for the tumor pixels (TP ratio), and relatively low false detections (FP ratio).

A second analysis is based on the *ROC curve* and a complete sensitivity/specificity report, a fundamental tool for diagnostic test evaluation. Fig. 3 shows the *ROC curves* corresponding to both SVM classifiers, where the true positive rate (Sensitivity) is plotted in function of the false positive rate (Specificity) for different cut-off points of a parameter. Each point on the *ROC curve* represents a sensitivity/specificity pair corresponding to a particular decision threshold. The area under the *ROC curve* (AUC) is a measure of how well a parameter can distinguish between two diagnostic groups (tumor/non-tumor).

Table 1. Quality scores corresponding to different configurations of the SVM classifier. All the ratios are shown between 0 and 1.

Classifier	Window-size	TP	FP	TN	FN	*F1*	*Overall Accuracy*
SVM (linear)	1x1	0.88	0.66	0.34	0.12	***0.66***	***0.80***
	33x33	0.83	0.66	0.34	0.17	***0.57***	***0.77***
	51x51	0.76	0.67	0.33	0.24	***0.54***	***0.73***
SVM (radial)	1x1	0.91	0.56	0.44	0.09	***0.59***	***0.79***
	33x33	0.89	0.59	0.40	0.11	***0.58***	***0.79***
	51x51	0.93	0.33	0.67	0.07	***0.33***	***0.72***

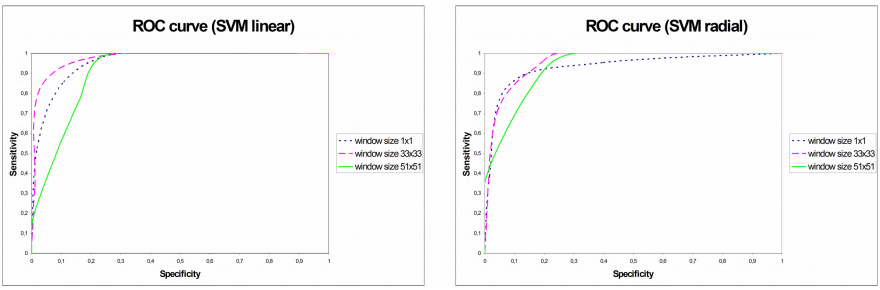

Fig. 3. ROC curves for the different configurations of the SVM classifier

4 Conclusions

This paper proposes a new pixel-based texture classification method for tumor region identification in mammograms. The proposed method firstly computes Gabor based features from the mammograms by means of multi-sized evaluation windows applied to the probabilistic distribution moments. Then, the identification of tumor regions is performed through a pixel-based classification scheme by using SVMs, which is able to deal with a high dimensional data space and a large amount of features. Promising results have been obtained for the identification of tumor regions on the mammograms of the mini-MIAS database. Further work will consist of combining new statistical texture features extracted from the Gabor filters and applying optimization methods to determine the optimal parameters of the SVM.

References

1. Yufeng, Z.: Breast Cancer Detection with Gabor Features from Digital Mammograms. Algorithms 3, 44–62 (2010)
2. Ioan, B., Gacsadi, A.: Directional Features for Automatic Tumor Classification of Mammogram Images. Biomedical Signal Processing and Control 6(4), 370–378 (2011)
3. Tang, J., Rangayyan, R.M., Xu, J., El Naqa, I.: Computer-Aided Detection and Diagnosis of Breast Cancer with Mammography: Recent Advances. IEEE Trans. Information Technology in Biomedicine 13(2), 236–251 (2009)
4. Chang, C.C., Lin, C.J.: LIBSVM: A Library for Support Vector Machines. ACM Transactions on Intelligent Systems and Technology 2(3), Article 27 (2011)
5. Manjunath, B.S., Ma, W.Y.: Texture Features for Browsing and Retrieval of Image Data. IEEE Trans. on Pattern Analysis and Machine Intelligence 18(8), 837–842 (1996)
6. Hussain, M., Khan, S., Muhammad, G., Berbar, M., Bebis, G.: Mass Detection in Digital Mammograms Using Gabor Filter Bank. In: Proc. IET Image Processing, July 2-3, pp. 1–6 (2012)
7. Hussain, M., Khan, S., Muhammad, G., Ahmad, I., Bebis, G.: Effective Extraction of Gabor Features for False Positive Reduction and Mass Classification in Mammography. Appl. Math. Inf. Sci. 6(1), 29–33 (2012)
8. Grigorescu, S., Petkov, N., Kruizinga, P.: Comparison of texture features based on Gabor filters. IEEE Trans. on Image Processing 11(10), 1160–1167 (2002)
9. Wu, T., Lin, C.J., Weng, R.C.: Probability estimates for Multiclass Classification by Pairwise Coupling. Journal of Machine Learning Research 5, 975–1005 (2003)
10. Wang, D., Shi, L., Heng, P.A.: Automatic Detection of Breast Cancers in Mammograms using Structured Support Vector Machines. Neurocomputing 72, 3296–3302 (2009)
11. The mini-MIAS database of mammograms: `http://peipa.essex.ac.uk/info/mias.html`