RESEARCH ARTICLE

A Novel Model for Smart Breast Cancer Detection in Thermogram Images

Iman Abaspur Kazerouni*, Hossein Ghayoumi Zadeh, Javad Haddadnia

Abstract

Background: Accuracy in feature extraction is an important factor in image classification and retrieval. In this paper, a breast tissue density classification and image retrieval model is introduced for breast cancer detection based on thermographic images. The new method of thermographic image analysis for automated detection of high tumor risk areas, based on two-directional two-dimensional principal component analysis technique for feature extraction, and a support vector machine for thermographic image retrievalwas tested on 400 images. The sensitivity and specificity of the model are 100% and 98%, respectively.

Keywords: Thermographic image - hot area - support vector machine - feature extraction

Asian Pac J Cancer Prev, **15 (24)**, 10573-10576

Introduction

CBreast density has been shown to be related to the risk of developing breast cancer since women with dense breast tissue can hide lesions, causing cancer to be detected at later stages (Wolfe, 1976). Breast cancer, among women, is the second-most common cancer and the leading cause of cancer death. It has become a major health issue in the world over the past decades and its incidence has increased in recent years mostly due to increased awareness of the importance of screening and population ageing (Boquete et al., 2012). Currently, mammography is the dominant method for detection of breast cancer. However, it is still far from being perfect. The high sensitivity of screening mammography is compromised by its low specificity to benign lesions, which often appear mammographically similar to malignant lesions (Wei et al., 2009). Although breast cancer incidence has increased over the past decade, breast cancer mortality has declined among women of all ages (Bray et al., 2004; Sickles, 1997). This favorable trend in mortality reduction may relate to improvements made in breast cancer treatment (Buseman et al., 2003) and the widespread adoption of early detection technology. Early detection is crucial in the effective treatment of breast cancer. Current mammogram screening may turn up many tiny abnormalities that are either not cancerous or are slow-growing cancers that would never progress to the point of killing a woman and might never even become known to her (Thomas et al., 2013). Normally surgery, radiation therapy and chemotherapy are used to treat breast cancer. Radiations cause damage to DNA strand inside the cancer cells, which inhibits its further growth (Wafa et al., 2014). Radiations can also damage the healthy tissues, but the effect is more on cancerous cells, as the growth of cancerous cells is very rapid and they cannot repair any damage easily (Wafa et al., 2014). The high sensitivity for cancer detection allows it to be used in evaluating several aspects of breast cancer diagnosis and treatment (Sawsan, 2014).

Three of these respective technologies include digital infrared thermal imaging (DITI), electrical impedance scanning (EIS) and elastography. While these devices all use non-invasive imaging methods, which neither emit ionizing radiation nor compress the breast, they do operate under differing physiological principles (Thomas et al., 2013). DITI aims to detect localized skin temperature increases, which are thought to occur as a result of increased vascularisation, vasodilatation and recruitment of inflammatory cells to the site of a developing tumor (Wei et al., 2009). Localized differences in skin temperature are captured by infrared cameras, which produce a heat map of the breast called a thermogram.

In this paper, a novel model for classification of breast tissue and the breast image retrieval is proposed, in which the principal component analysis (PCA) and the two-dimensional principal component analysis (2D PCA) and the two-directional two-dimensional principal component analysis ((2D)2PCA) hve been used for feature extraction and dimension reduction of the thermographic images. After this step, retrieval is performed with the aid of a support vector machine (SVM) that is able to solve a variety of problems. This paper is organized as follows: Section (2) presents a compact expression for $(2D)^2$PCA technique for feature extraction. In section (3), the support vector machine (SVM) with RBF kernel for image retrieval is described. Sections (4) present results

*Department of Electrical Engineering, Hakim Sabzevari University, Sabzevar, Iran *For correspondence: imanabaspur@gmail.com*

of the proposed model and summary, respectively.

Feature Extraction

Images can be numerically represented by a feature vector, preferentially at a low-dimensional space in which the most relevant visual aspects are emphasized. In order to describe the different patterns of parenchyma tissue within one category, the texture attribute can be used. However, the high dimensionality of a feature vector that represents texture attributes limits its computational efficiency, so it is desirable to choose a technique that combines the representation of the texture with the reduction of dimensionality, in a way to turn the retrieval algorithm more effective and computationally treatable. Principal component analysis (PCA) is a useful technique for feature extraction in the various applications as image processing, pattern recognition and, etc (Zhang and Zhou, 2005). In the PCA technique, a 2-dimensional image matrix is transformed to 1-dimensional vector's row or column. When the image matrix is very large, it is difficult to evaluate the covariance matrix accurately due to its large size and the relatively small number of training samples. The 2-dimensional principal component analysis (2DPCA) has been already proposed to solve this problem. In this new technique, the eigenvectors are calculated directly without the matrix to vector transmission. The size of the image covariance matrix is equal to image wide size in the 2DPCA model, and the size of the covariance matrix is smaller than PCA model. Dislike the standard principal component analysis (PCA) which is based on one-dimensional vector, the 2DPCA technique is based on two-dimensional matrices. The experimental results in many reports proved that the accuracy of 2DPCA is often better than PCA (Zhang and Zhou, 2005). If 2DPCA is applied in the rows of image matrices once and then it is worked in the columns of image matrices, the results of principal component extraction will be accurate. This technique is called 2-directional 2-dimensional principal component analysis ($(2D)^2PCA$). In this model; a 2DPCA is essentially used in the row direction of images, and then an alternative 2DPCA is worked in the column direction of images.

In $(2D)^2PCA$ technique, the size reduction is performed in the image rows and columns simultaneity. Assume A is an image matrix with m row and n column and $X \in R^{n \times d}$ is a conversion matrix with orthonormal columns where $n \geq d$ and therefore, $Y=AX$ is an m by d matrix.

Now let M as training samples with m by n matrices, which are shown by A_k $(k=1, 2,\ldots, M)$ and $\bar{A}$ and C as average matrix and covariance matrix respectively. Thus $\bar{A}$ can be written as:

$$\bar{A} = (1/m)\Sigma_{k=1}^{M} A_k \quad (1)$$

and C can be obtained by:

$$\bar{A} = (1/m)\Sigma_{k=1}^{M} (A_k - \bar{A})^T (A_k - \bar{A}) \quad (2)$$

Now an optimal conversion matrix can be obtained by calculation of the covariance matrix eigenvectors and selection of d eigenvectors. The eigenvectors can be found by d first eigenvalues, which are arranged from high to low values. The optimal conversion matrix is shown as X which the columns of this matrix are formed by the selected eigenvectors. It is to be noted that the size of C matrix is $n \times n$ and thus the computation of eigenvalues and eigenvectors will be fast and very optimal.

The further size reduction can be also yielded using another 2DPCA on image matrix. Thus, the covariance matrix C can be defined as:

$$\bar{A} = (1/m)\Sigma_{k=1}^{M} \Sigma_{i=1}^{m} (A_k^{(i)} - \bar{A}^{(i)})^T (A_k^{(i)} - \bar{A}^{(i)})^T \quad (3)$$

Where $A_k^{(i)}$ and $\bar{A}^{(i)}$ denote the i-th row vectors of A_k and $\bar{A}$ respectively. Equation (3) is a 2DPCA operator in the row direction of image and shows that when the average of images matrix be zero, the covariance matrix can be derived by multiplication of row vectors of images. Now another 2DPCA can be worked in the column direction of image, and Equation (3) can be rewritten as:

$$\bar{A} = (1/m)\Sigma_{k=1}^{M} \Sigma_{j=1}^{n} (A_k^{(j)} - \bar{A}^{(j)})^T (A_k^{(j)} - \bar{A}^{(j)})^T \quad (4)$$

Where $A_k^{(j)}$ and $\bar{A}^{(j)}$ denote the j-th column vectors of A_k and $\bar{A}$ respectively. In this step, q eigenvectors corresponding to q first high eigenvalues of matrix C can be obtained, and these eigenvectors are located as columns in the matrix Z which, $Z \in R^{m \times q}$. Projecting the random matrix A onto Z yields a q by n matrix $Y=Z^TA$ and also projecting the matrix A onto Z and X yields a q by d matrix $Y=Z^TAX$. The matrix C is used as the feature matrix in the proposed model.

Image Retrieval

After $(2D)^2PCA$ and features extraction, the obtained data must be classified. The proposed technique for the classification in the proposed model is the support vector machine (SVM) technique. Support vector machines (SVM) are a popular machine learning method for classification, regression, and other learning tasks. In data classification and analysis for the linear samples, SVM can select two parallel lines to separate the data with possible high accuracy. Generally speaking, when these lines cannot classify data, SVM uses a nonlinear conversion for transform data vector x to a higher-dimensional space. SVM can be used without any kernel but in the proposed model a SVM with the kernel function is used for the over-fitting error reduction.

For two data set (x_i, y_i), $i=1, 2, \ldots, l$ where $x_i \in R^n$ and $y \in \{1, -1\}^1$, the support vector machines solve the following optimal equation (Boser, Guyon, and Vapnik, 1992):

$$\min_{w, b, \eta} = (1 \div 2) w^T w + L\Sigma_{i=1}^{l} \eta_i \quad (5)$$

Subject to $y_i(w^T \varphi(x_j) + b \geq 1 - \eta_i$

Where function φ transforms training vectors x_i to higher-dimensional space. $L>0$ is the parameter of the error term. The kernel function is denoted by $K(x_i, y_j) \equiv \varphi(x_i)^T \varphi(x_j)$. In the proposed model, the radial basis function (RBF) has been used for classification:

$$K(x_i, y_j) = \exp(-\lambda \|x_i - y_j\|^2) \quad (6)$$

Where λ is the kernel parameter. The RBF kernel parameters L and λ have significant effects on accuracy of the classification process. The selection of these parameters can reduce the error percentage. The v-fold

cross-validation has been used for the parameters selection. In this approach, data is divided into v equal size subsets. Sequentially, one subset is tested using the classifier trained on the remaining v-1 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy means the percentage of correctly classified data. The v-fold cross-validation can also eliminate the over-fitting error. For image retrieval, the various parameters values were tested and the results were analyzed to select the best kernel parameters. The best values in the proposed model are $2^{3.54}$ and $2^{-2.1}$ for C and λ respectively for v=3.

Results

The proposed model has been applied to 400 thermographic images that captured and collected by Hakim Sabzevari Medical Imaging Group in the Vasei Hospital. The proposed model has been implemented using MatLab (Matrix Laboratory) and the LIBSVM library (Chang and Lin, 2001).

The feature extraction has been performed using PCA, 2DPCA and $(2D)^2$PCA techniques. Several feature matrices have been selected and compared to previous literature. Table 1 shows the results of average precision for the selected d first eigenvalues for the proposed model and PCA and 2DPCA. The average precision values for the eigenvalues based on $(2D)^2$PCA is almost higher than other three techniques. Using $(2D)^2$PCA technique, proposed model gives proper results while uses a few principal components. The first eight eigenvalues of the covariance matrix was used as features matrix because it can be observed from Table 1 that the highest value was 99.32% of average precision for the first eight principal components of $(2D)^2$CPA.

Table 2 shows a comparison among valid reported literature for a general comparison with other techniques. It can be inferred that the proposed model has higher correct classification percentage.

Table 1. The Results of Average Precision for the Selected d First Eigenvalues for the Proposed Model and PCA and 2DPCA Techniques

d	PCA	2DPCA	$(2D)^2$PCA
1	81.67%	84.13%	89.01%
2	71.78%	90.32%	92.58%
3	72.67%	86.83%	96.93%
4	69.87%	89.62%	94.14%
5	73.92%	82.66%	90.92%
6	75.17%	80.45%	92.66%
7	80.98%	88.24%	88.81%
8	77.01%	82.74%	99.32%
9	79.12%	81.04%	91.22%
10	78.33%	89.11%	93.09%
11	70.39%	90.76%	89.91%
12	71.09%	87.89%	93.31%
13	76.20%	76.93%	92.64%
14	73.52%	82.47%	90.04%
15	79.34%	84.09%	97.36%
16	79.55%	88.41%	96.72%
17	82.98%	79.37%	98.72%
18	81.37%	84.96%	97.42%
19	78.99%	86.46%	98.10%
20	74.60%	89.94%	98.87%

Table 2. A General Comparison with other Techniques

Author	This Work	Boquete (Boquete et al., 2012)	Acharya (Acharya, 2012)
Method	2DPCA and $(2D)^2$PCA and SVM	ICA	SVM
Specificity	98%	94%	90.48%
Sensitivity	100%	100%	85.71%
Year	2014	2012	2012

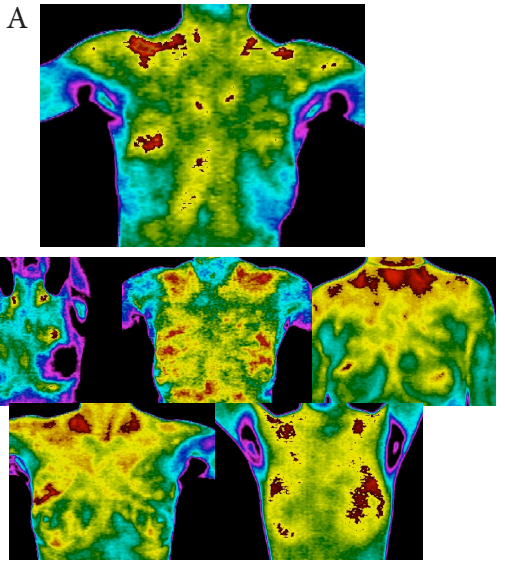

Figure 1. (A) An Example Query Image from Cancerous Breast Category and (B) the Retrieved Images Based on the Proposed Model

Figure 1 shows a query image from a breast with cancer and retrieved images based on the proposed model. All the retrieved images are similar to the query image and all of them are cancerous breasts.

Discussion

In this paper, a breast tissue density classification and image retrieval model has been studied. We present a model for the data reduction based on two-Directional two-dimensional principal component analysis ($(2D)^2$PCA) technique and the extracted features and data is used in a support vector machine (SVM) with the radial basis function (RBF) for classification and thermographic image retrieval. The 3-fold cross-validation has been used for the parameters selection in SVM to avoid the over-fitting error in the data classification. The various parameters values were tested and the results were analyzed to select the best kernel parameters. The sensitivity and specificity of the model are 100% and 98%, respectively.

References

Acharya U, Ng E, Jen-Hong T, Vinitha S (2012). Thermography based breast cancer detection using texture features and support vector machine. *J Med Syst*, **36** ,1503-10.

Boquete L, Ortega S, Miguel-Jiménez J, et al (2012) . Automated detection of breast cancer in thermal infrared images, based on independent component analysis. *J Med Syst*, **36**, 103-111.

Boser B, Guyon I, and Vapnik V (1992) .A training algorithm for optimal margin classifiers', in proceedings of the fifth annual workshop on computational learning theory. *ACM*

Press, **24**, 144-152.

Bray F, McCarron P, and Parkin D (2004) .The changing global patterns of female breast cancer incidence and mortality. *Breast Cancer Res*, **6**, 229-39.

Buseman S, Mouchawar J, Calonge N, and Byers T (2003) .Mammography screening matters for young women with breast carcinoma. *Cancer*, **97**, 352-8.

Chang C. and Lin C (2001) .LIBSVM: a library for support vector machines. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.

Sawsan AT (2014). Breast magnetic resonance imaging indications in current practice. *Asian Pac J Cancer Prev*, **15**, 569- 75.

Sickles E (1997). Breast cancer screening outcomes in women ages 40-49: Clinical experience with service screening using modern mammography. *J Natl Cancer Inst Monographs*, **22**, 99-104.

Thomas D, Cameron D, Mundy L, Hiller J (2013). A systematic review of elastography, electrical impedance scanning and digital infrared thermography for breast cancer screening and diagnosis. *Breast Cancer Res Treat*, **137**, 665-76.

Wafa M., Bilal A, Ijaz J, Tanweer K, et al (2014). Breast cancer: major risk factors and recent developments in treatment. *Asian Pac J Cancer Prev*, **15**, 3353- 8.

Wei L, Yang Y and Nishikawa R (2009). Micro-calcification classification assisted by content-based image retrieval for breast cancer diagnosis. *Pattern Recognit*, **42**,1126-32.

Wolfe K (1976). Breast patterns as an index of risk for developing breast cancer. *Am J Roentgenol*, **34**, 1130-9.

Zhang D and Zhou Z (2005). $(2D)^2$PCA: Two-directional Two-dimensional PCA for efficient face representation and recognition. *Neurocomputing*, **69**, 224-231.