Expert Systems With Applications 57 (2016) 204–213

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Mammogram classification using sparse-ROI: A novel representation to arbitrary shaped masses

Karteeka Pavan Kanadam*, Srinivasa Rao Chereddy

Department of Computer Applications, R.V.R. & J.C. College of Engineering, Chowdavaram, Andhra Pradesh, India

ARTICLE INFO

Article history:
Received 14 January 2015
Revised 17 March 2016
Accepted 18 March 2016
Available online 29 March 2016

Keywords:
Computer aided detection
Computer aided diagnosis
Region of interest, sparse matrix
Classification
GLCM

ABSTRACT

Masses in breast are the important radiographic signs of cancer. Developing automated detection of these masses is the main objective in the medical detection of breast cancer. Computer aided detection (CADe) and computer aided diagnosis (CADx) respectively refer to the process of identifying the region of interest (ROI) and the classification of the ROI into one of the classes of abnormalities. Computer aided system of identifying abnormalities will aid the medical doctors to decide the course of treatment for the patients. So far, the methodologies of CAD systems are based on regular shape and a common uniform size ROI as either suggested by radiologists or by domain knowledge. Here an attempt is made to model automatically, the identified irregular shaped masses (ROI) as they occur using sparse matrix and was named as sparse-ROI. This is the first attempt that considers arbitrary shape of the mass as ROI. The proposed sparse-ROI eliminates the risk of a common optimum sized window selection that fits best to all or a class of mammograms of the data base. Once the shape of the mass is detected through the new technology of sparse-ROI, diagnosis methodology (classifying a given mammogram into one of the 7 well known classes) is proposed based on the features extracted. Having extracted the features, multi-SVM is used for the classification. The performance of the classifier is studied on mammograms of the bench mark data set MIAS. For this purpose two algorithms are proposed based on the well-known statistical matrices, gray level co-occurrence matrix (GLCM) and gray level aura matrix (GLAM). The efficacy of the classifier of the two new algorithms developed is evaluated in terms of accuracy, precision, sensitivity, size and computational time. The results of the study are enterprising with reduction in computational time by 99.93% in GLCM and 75.73% in GLAM with the concomitant retention of classification accuracy of 97.2%.

1. Introduction

Breast cancer is the most common cancerous disease in women and is considered as the second main cause of death among women (Winsberg, Elkin, Macy, Bordaz, & Weymouth, 1967; Kopans, 2007; Malvezzi, Bertuccio, Levi, La Vecchia, & Negri, 2014).Various medical surveys show that one in two cancer diagnosed cases are leading to death. A longer survival time can be possible with the early detection of breast cancer. The common image modality for the efficient analysis of breast cancer is the mammogram (Bozek, Mustra, Delac, & Grgic, 2009). Mammography is the most reliable method that helps radiologists in early detection of abnormalities and treatment planning (Pisano et al., 2000). As the human interpretation of mammogram varies from one expert to another, a repetitive interpretation is required to avoid misinterpretation of breast tissues. Therefore, computer aided detection (CADe) and computer aided diagnosis (CADx) systems are being developed for efficient diagnosis (Marrocco, Molinara, D'Elia, & Tortorella, 2010; Jiang, Yao, & Wason, 2007; Verma, McLeod, & Klevansky, 2009; Haralick, Shanmugam, & Dinstein, 1973; Chang et al., 2006; Ke, Mu, & Kang, 2010).

Earlier works in literature depict that CAD systems significantly increase the accuracy of detection and diagnosis (Rouhi, Jafari, Kasaei, & Keshavarzian, 2015, Abdel-Zaher & Eldeib, 2016). However, CAD systems suffer from higher rate of false positives and false negatives (Tang, Rangayyan, Xu, Naqa, & Yang, 2009, Choi, Kim, Plataniotis, & Ro, 2016). There is a need to develop CAD systems with improved accuracy to help radiologists in an accurate detection and a better treatment planning (Braz Junior, Rocha, Marcelo, & Silva, 2013).

The process of diagnosing cancer using CAD system has three steps. First step is segmentation of region of interest (ROI),

* Corresponding author: Tel.: +0919885139413.
E-mail addresses: karteeka@yahoo.com (K.P. Kanadam), chereddy_sriny@yahoo.com (S.R. Chereddy).

http://dx.doi.org/10.1016/j.eswa.2016.03.037
0957-4174/

followed by feature extraction and classification (Rouhi & Jafari, 2016). The performance of each step affects the subsequent step and thus overall performance of the system (Mousa et al., 2014).

Mammographic images contain different types of noises, artifacts and pectoral muscles. Literature reveals that suspicious masses and pectoral muscle portions have similar pixel intensity values. The automatic CAD systems may misclassify the pectoral muscle as a mass. Therefore, the unwanted portions like pectoral muscle to be eliminate to improve the performance of the texture analysis and further mammogram classification (Beura, Majhi, & Dash, 2015).

The portion of the area in mammographic image, containing the identified suspicious mass or any abnormality is called as the Region of Interest (ROI). The efficiency of CAD systems heavily depend on finding the size and shape of ROI (Jen & Yu, 2015). ROI is commonly selected either as suggested by radiologists or by domain knowledge or by image segmentation technique in CAD systems (Chu, Min, Liu, & Lu, 2015). Usually the segmented ROIs are of different sizes and shapes, as abnormalities occur in large varieties of sizes and shapes. Hither to researchers resize each ROI into a fixed window with a common size for feature extraction. But in reality, the mass regions are always of arbitrary shapes and of different sizes. Different varieties of masses of the mammograms result in a high dimensioned feature space.

The high dimensionality of feature space degrades the performance of the classification. A larger sized common template for ROIs, not only lowers performance, but also computationally expensive (Garcia-Manso, Garcia-Orellana, Gonzalez-Velasco, Gallardo-Caballero, & Macias Macias, 2013). Researchers select different window sizes in different experiments depending on the kind of study, database and malignancy class of mammogram, based on the domain knowledge (Hussain, 2014). An optimum size for a class of ROIs is selected based on the results of different ROI sizes in different experiments (Abdel-Nasser, Rashwan, Puig, & Moreno, 2015).

This paper attempts to address these issues, by proposing a new data model using sparse matrix, sparse-ROI. The proposed sparse-ROI represents automatically the arbitrary shaped masses as they occur, eliminating the risk of a common optimum sized window selection that fits best to all or a class of mammograms of the data base. The method considers only mass region by ignoring all the pixels not contributing to area of mass. Sparse-ROI does not require any domain knowledge. Eventually the method reduces the feature space and computational time drastically by considering only mass region. The proposed Sparse-ROI method is described by two new algorithms GLCM_sparse-ROI and GLAM_sparse-ROI developed based on well-known statistical matrices, Gray Level Co-occurrence Matrix (GLCM) and Gray Level Aura Matrix (GLAM), as the existing works were fit to window shape. Further, thirteen textural features are extracted from sparse-ROI and SVM is used to classify the mammogram. The efficacy of the classifier of the two new algorithms developed is evaluated in terms of accuracy, precision, sensitivity, size and computational time, on mammograms of a benchmark database called MIAS. The results of the study are enterprising with reduction in computational time by 99.93% in GLCM and 75.73% in GLAM with the concomitant retention of classification accuracy of 97.2%.

The remaining part of the paper is organized as follows: Section 2 explains the related work, Section 3 presents the main contributions of the paper, Section 4 devotes to the proposed sparse-ROI based methodology, Section 5 discusses the experimental results and discussions and the conclusions and future directions are presented in Section 6.

2. Related work

Major works in literature posed mammogram analysis problem as a multi-modal classification framework. The size and the shape of the breast mass are two significant factors in prediction of class of the mass (Anitha & Peter, 2015). Based on size, shape and location of the mass, the masses can be classified into calcifications (CALC), circumscribed (CIRC), speculated (SPIC), ill-defined (MISC), architectural distortions (ARCH), asymmetry (ASYM) and non-mass (NORM) categories. The CAD systems are developed based on image processing techniques followed by a feature extraction for classification of masses. The features are extracted from ROI of mammogram. Hence, in most of the CAD systems, a cropping operation is applied to extract ROI excluding the unwanted portions of the image. The ROI is a regular rectangular window of common size either suggested by an expert or chosen from the empirical domain knowledge.

A wide variety of methodologies are developed for the problem of mammogram classification. Among the existing, the two major categories for describing ROIs are geometry and texture. This paper focuses on texture. In geometric analysis, mammograms are classified based on shape of the mass. The shape of the mass can be characterized using features like area, perimeter, circularity, density etc. (Dong et al., 2015, Sampaio, Moraes, Silva, Paiva, & Gattass, 2011).

A large number of studies employ textural features to classify mammograms. Texture is spatial relation of distribution of pixel intensities. Various textural descriptors are local binary patterns, statistical matrices, wavelets etc. Kim, Park, Song, and Park (1998), developed a statistical texture-analysis method, called the Surrounding Region Dependence method (SRDM), for detecting clustered micro calcifications in digitized mammograms. A back propagation neural network is used as a classifier. The observed sensitivity was more than 90%. In this work the authors acquired ROI of size 128×128 pixels (Kim & Park, 1999).

Nakayama, Uchiyama, Yamamoto, Watanabe, and Namba (2006) have used filter bank based on the concept of the Hessian matrix for micro calcification cluster detection. They have used a 115×115 pixels sized ROI. Walker, Volk, Smith, and Miller (2009) have used a parallel multi chromosomal cartesian genetic programming algorithm for characterization and classification of micro calcifications and have applied on Lawrence Livermore National Laboratory database. The ROIs are cropped to a uniform size of 128×128 pixels from the database. The mammograms which were identified as normal can be developed as malignant masses in near future. Sameti, Ward, Morgan-Parkes, and Palcic (2009) have developed a feature extraction system to detect the signs of cancer development in last screenings prior to detection of a malignant mass and concluded that the 72% of flagged regions can turn as malignant masses. For experimentation they have chosen a 256×256 size as common for all ROIs.

Markov random field model and the deterministic fractal model are used for clustered micro calcification classification. Here a 88×88 size is taken as the common size for each ROI of MIAS database (Yu & Huang, 2010). Breast density plays an important role in characterization of masses. Richard and Bierme (2010) extracted manually the ROI of size 512×512 for the analysis of texture anisotropy based on directional patterns and proved anisotropic fractional Brownian fields are better-suited than the commonly used fractional Brownian fields to the modeling of mammogram textures.

In 2011, Quintanilla-Dominguez et al. (2011), have determined micro calcifications using fuzzy possibilistic clustering algorithm and applied on MIAS ROIs. Each ROI is of size 256×256 pixels. Ramos, do Nascimento, and Pereira (2012) extracted textural features using two multi resolution techniques, wavelet and