Indonesian Journal of Electrical Engineering and Computer Science
Vol. 10, No. 3, June 2018, pp. 1030~1035
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v10.i3.pp1030-1035

Fusion of Random Projection, Multi-resolution Features and Distance Weighted K Nearest Neighbor for Masses Detection in Mammographic Images

Viet Dung Nguyen[1], Minh Dong Le[2]
[1]Department of Biomedical Engineering, Hanoi University of Science and Technology, Vietnam
[2]Department of Computer Science, Chonnam National University, South Korea

Article Info

Article history:

Received Jan 11, 2018
Revised Mar 19, 2018
Accepted Mar 30, 2018

Keywords:

Distance weighted K nearest neighbor
Mammography
Mass detection
Multi-resolution features
Random projection

ABSTRACT

Breast cancer is the top cancer in women both in the developed and the developing world. For early detection of the disease, mammography is still the most effective method beside ultrasound and magnetic resonance imaging. Computer Aided Detection systems have been developed to aid radiologists in diagnosing breast cancer. Different methods were proposed to overcome the main drawback of producing large number of False Positives. In this paper, we presented a novel method for masses detection in mammograms. To describe masses, multi-resolution features were utilized. In feature extraction step, we calculated multi-resolution Block Difference Inverse Probability features and multi-resolution statistical features. Once the descriptors were extracted, we deployed random projection and distance weighted K Nearest Neighbor to classify the detected masses. The result is quite sanguine with sensitivity, false positive reduction and time for carrying out the algorithm

Corresponding Author:

Viet Dung Nguyen,
Department of Electronic Technology and Biomedical Engineering,
Hanoi University of Science and Technology,
No 1. Dai Co Viet Str., Hanoi, Vietnam.
Email: dung.nguyenviet1@hust.edu.vn

1. INTRODUCTION

Breast cancer is the most common cancer in women worldwide, with nearly 1.7 million new cases diagnosed in 2012 [1]. Abnormal tissue screening using X-ray mammography is currently the most effective method of early detection of the disease [2-3]. The introduction of digital mammography gave the opportunity of increasing the number of commercial Computer Aided Detection (CAD) systems, which has significantly enhanced the radiologists' ability to detect and diagnose cancer and take immediate precautions for its earliest prevention [4]. One problem with CAD systems is due to a large number of false positive (FP) marks when high sensitivity is required [5]. Too many false positives may confuse the radiologist of the most common types of cancer among women all over the world is breast cancer. Great effort has been devoted in recent years to the development of CAD which propose a lot of features to reduce false positives [6]. However, many features are not key features of masses and they make high dimensions for classification.

In this paper, we introduce novel method using moment and basic characteristic of the masses. Block Difference Inverse Probability (BDIP) and basic features are calculated in different multi-resolutions. Once the features are extracted, random projection [7] and k nearest neighbor (k NN) [8] with distance weighting are used to classify the suspicious areas into real mass or normal parenchyma.

2. PROPOSED METHOD

2.1. Database

In this study, we use mammogram database Mini- MIAS [9] to test the method presented. MIAS is the public database of Mammographic Image Analysis Society - an organization of United Kingdom research groups. This database includes 322 mammograms from 161 patients. Films taken from the United Kingdom National Breast Screening Program have been digitized to 50-micron pixel edge and presented each pixel with an 8-bit word. Every image in database always has extra information or ground truth as shown in Figure 1 from the radiologists about characteristic of background tissue, type of abnormality present, severity of abnormality, the coordinates of center and approximate radius (in pixels) of a circle enclosing the abnormality. Mini-MIAS database is a reduced type of the original MIAS database (digitized at 50-micron pixel edge) has been reduced to 200-micron pixel edge and clipped/padded so every image has size of 1024 x 1024 pixels.

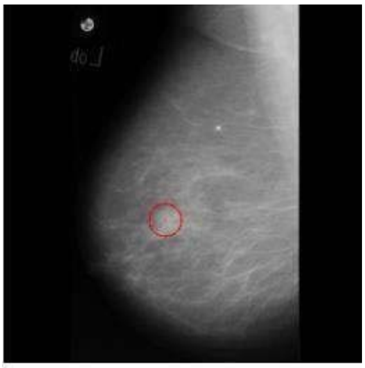

Figure 1. Red line shows ground truth in MINI-MIAS database

2.2. Preprocessing

The aim of the step is to remove unnecessary information in mammograms such as label, pectoral muscle or other noise. To separate the breast region from image label, we just threshold the image and keep the biggest threshold region. The pectoral muscle in a mammographic image appears as a predominant density region. It can affect negatively the result of detection method [10]. For this reason, the region representing the pectoral muscle should be eliminated. In the mammogram, there are also some small bright spots which have gray level approximate that of circumscribed mass. Median filtering with a window of 3x3 is applied for eliminating these spots as illustrated in Figure 2.

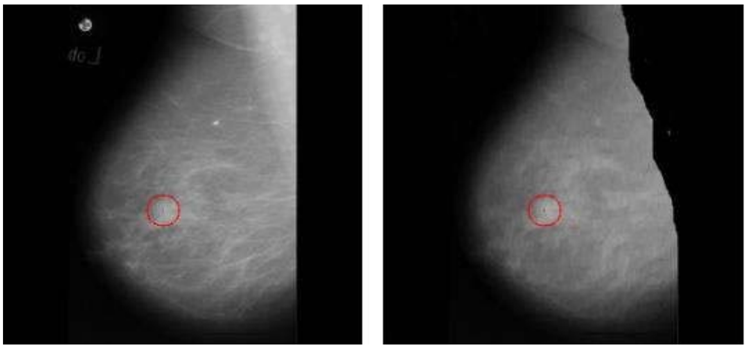

Figure 2. Original (left) and preprocessed (right) mammograms

2.3. Mass detection

In this stage, suspicious regions are extracted from the preprocessed mammogram. The radiologists should focus their attention to these extracted regions. The steps of this procedure are fully described in [11]. Shown in Detected ROIs are masked are masked as true positive ROIs (TP-ROIs) or false positive ROIs (FP-ROIs) as illustrated in Figure 3 based on the provided ground truth.

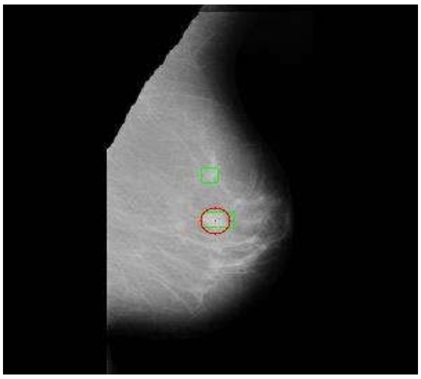

Figure 3. Detected ROIs (green) and ground truth (red)

2.4. Feature extraction

In human vision, edges and valleys [12] in an image are very important features, especially valleys are fundamental in the vision perception of an object shape [13-14]. Block Difference Inverse Probability (BDIP) is the texture feature which measures the variation in intensities of an image block. It effectively extracts edges and valleys. The larger the variations of intensity or the size of the block, the higher the value of BDIP [15]. BDIP of a block of size WxW is defined as:

$$BDIP = \frac{\frac{1}{W^2}\sum_{(i,\,j)\in B}\left[\max_{(i,\,j)\in B} I(i,\,j) - I(i,\,j)\right]}{\max_{(i,\,j)\in B} I(i,\,j)}$$

where I(i,j) denotes the intensity of a pixel (i,j) in the block B.

As the detected ROI is not in size of WxW so we subtitute the term “W2”in above equation by size or number of pixels in the ROI to calculate the BDIP feature at first resolution, which then is just simply called BDIP. Other BDIP features at different resolution are calculated as follow:

a. Divide each side of the minimal rectangular that contains the ROI by 2, 3...n to get 4, 9... n^2 blocks.
b. For each block using above equation to calculate BDIP features which are called BDIP2x2 and BDIP3x3... BDIPnxn.
c. Expectation and variation of BDIPs are used as BDIP features for each RoI. They are BDIP2x2mean, BDIP2x2var, BDIP3x3mean, BDIP3x3var,...BDIPnxnmean, BDIPnxnvar respectively.

On the other hand, we compute basics features of each ROI:

a. Mean: the average grey level
b. Var: the standard deviation of grey level
c. Max: the highest grey level
d. Min: the lowest grey level

However high or low intensity values is not absolute, input images often have different brightness. We propose two extra features for ensuring the persuasive of our algorithm

a. Ratio_1: Mean/Max
b. Ratio_2: Max/Max_I

where Max_I is the highest gray level of the whole image.

Multi-resolution basic features are calculated in the same manner as multi-resolution BDIP feature.

2.5. Random Projection

In mathematics and statistics, random projection is a technique used to reduce the dimensionality of a set of points which lie in Euclidean space. Random projection methods are powerful methods known for their simplicity and less erroneous output compared with other methods. According to experimental results, random projection preserve distances well, but empirical results are sparse [15]. In random projection, the original D-dimensional data is projected to a L- dimensional (L << D).

$$X_{LxN} = R_{LxD} X_{DxN}$$

where X_{LxD}, X_{DxN} denote output and input matrix and R_{LxD} is a random projection matrix.

The random matrix R can be generated using a Gaussian distribution. Achlioptas [15] has shown that the Gaussian distribution can be replaced by a much simpler distribution such as:

$$R_{i,j} = \sqrt{3}\begin{cases} +1 & \text{with probability } 1/6 \\ 0 & \text{with probability } 2/3 \\ -1 & \text{with probability } 1/6 \end{cases}$$

2.6. K Nearest Neighbor

Let T = {(x_i, y_i): i=1:N} denote the training set where x_i is the training vector in m-dimensional feature space and y_i is the corresponding class label. Given unknow x', class y' is assigned by two steps

a. First, a set of k labelled target neighbours for the x' is identified and sorted in ascending order in term of Euclidean distance to x'.
b. Second, the class label y' is predicted by major voting of it nearest neighbours.

A weighted voting scheme for kNN, which is called distance-weighted k nearest neighbor (wkNN) rule is proposed in [16]. In wkNN, the closer neighbors are weighted more heavily than the farther ones, using the distance-weighted function. Then the classification result of the query is made by the majority weighted voting a neighbor with smaller distance is weighted more heavily than one with greater distance: the nearest neighbor gets weight of 1, the furthest neighbor a weight of 0 and the other weights are scaled linearly to the interval in between.

3. RESULTS

The number of detected ROI is 1000 [11]. For each ROI, BDIP and basic features are calculated at n level. The maximal value of n is the minimal radius of a circle enclosing the abnormality provided in the Mini-MIAS database. Totally we have 2400 features. Different values of K are tested and value of K which gives highest sensitivity is selected. Figure 4 shows the performance with different K value. The selected value of K is 21 with sensitivity of 90 %.

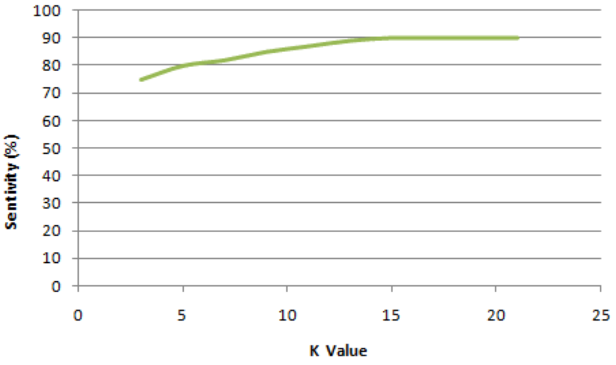

Figure 4. Original (left) and preprocessed (right) mammograms

Table 1 gives comparisons of our method to different approaches. It is obvious that our method provides higher sensitivity at lower number of false positives per image. On the other hand, we also compare the performance in terms of sensitivity, false positive per image, time of random projection and time of running between different sizes of random projection matrix. The results are given in Table 2. The result shows random projection help to reduce time of running. This tool should be effective with big data and a lot of features but in small data it can influence to other performance.

Table 1. Comparison to other approaches

Approach	Sensitivity (%)	False Positives per Image
Density slicing, texture flow field analysis	81	2.2

Multi-level threshold segmentation	80	2.3
K mean clustering	85	1
Multi-resolution features, distance weighted k nearest neighbor	90	1.04

Table 2. Performance with different size of random projection matrix

Size of matrix	Sensitivity	False positive per image	Time of random projection per image (s)	Running time per image (s)
2000x2400	89	1.1	2.1	19
1500x2400	87	1.2	1.9	17
1000x2400	85	1.4	1.6	16
Full	90	1.04		24

4. CONCLUSIONS

This study proposes a new method to detect masses in mammographic image based on combination of multi-resolution features and distance weighted K nearest neighbor algorithm. The highest sensitivity is observed with small false positive per image. Comparisons with other related works prove that our method is effective and has potential to be further investigated. When using random projection, this tool will be effective with big data. In the future, we will evaluate the method on larger set of mammograms and use different features.

REFERENCES

[1] Ghoncheh M*, et al.*, "Incidence and Mortality and Epidemiology of Breast Cancer in the World," *Asian Pac J Cancer Prev*, vol. 17(S3), pp. 43-46, 2016.

[2] H. D. Cheng, *et al.*, "Approaches for automated detection and classification of masses in mammograms," *Pattern Recognition,* vol. 39(4), pp. 646-668, 2006

[3] Elahe Chaghari, *et al.*, "A Novel Approach for Tumor Detection in Mammography Images," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12(8), pp. 6211-6226, 2014.

[4] V. M. Rao, *et al.*, "How Widely Is Computer-Aided Detection Used in Screening and Diagnostic Mammography?," *Journal of the American College of Radiology,* vol. 7(10), pp. 802-805, 2010.

[5] P. Taylor, *et al.*, "Impact of computer-aided detection prompts on the sensitivity and specificity of screening mammography," *Health Technol Assess.* vol. 9(6), pp. 1-58, 2005.

[6] Fariha Nosheen, *et al., "False positive and false negative reduction in digital mammograms using binary rotation invariant and noise tolerant texture descriptor,"* 2017 International Conference on Communication Technologies (ComTech), 2017

[7] P. Punithavathi, *et al.*, "Random Projection-based Cancelable Template Generation for Sparsely Distributed Biometric Patterns," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 7(3), pp. 877-886, 2017

[8] Zhigao Zheng, *et al.*, "Time-Weighted Uncertain Nearest Neighbor Collaborative Filtering Algorithm,*" Indonesian Journal of Electrical Engineering and Computer Science*, vol. 12(8), pp. 6393-6402, 2014

[9] http://peipa.essex.ac.uk/info/mias.html

[10] N. R. Mudigonda, *et al.*, "Detection of breast masses in mammograms by density slicing and texture flow-field analysis," *IEEE Transactions on Medical Imaging*, vol. 20(12), pp. 1215-1227

[11] V. D. Nguyen, *et al., "Detection of tumor in mammographic images by hierarchy of block's features,"* 19th International Conference on Digital Signal Processing, DSP2014

[12] Y. D. Chun, *et at.*, "Image retrieval using BDIP and BVLC moments," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13(9), pp. 951-957

[13] D. E. Pearson and J. A. Robinson, *"Visual communication at very low data rates,"* Proceedings of the IEEE, vol.73(4), pp. 795-812.

[14] Y, J. Ryoo, N. C. Kim, "Valley operator for extracting sketch features: DIP," *Electronics Letters*, vol. 24(8), pp. 461-463.

[15] T. D. Nguyen, *et al.*, "Surface Extraction Using SVM-Based Texture Classification for 3D Fetal Ultrasound Imaging," 1st International Conference on Communications and Electronics (ICCE2006), 2006.

[16] D. Achlioptas, "Database-friendly random projections," *20th ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* PODS2001

[17] S. A. Dudani, "The Distance-Weighted k-Nearest-Neighbor Rule," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 6(4), pp. 325-327

BIOGRAPHIES OF AUTHORS

Viet Dung Nguyen received Doctorate degree from Hanoi University of Science and Technology, Hanoi, Vietnam, in Electronic Engineering in 2016. Dr. Nguyen is currently working as Senior Lecturer, Vice Head of the Department of Electronic Technology and Biomedical Engineering of School of Electronics and Telecommunications, Hanoi University of Science and Technology, Hanoi, Vietnam which he joined in 2000. His main research interests include biosignal and medical image analysis; medical instrumentation.

Minh Dong Le received his Engineer Degree and Master Degree of Engineering in Biomedical Engineering at Hanoi University of Science and Technology, Hanoi, Vietnam in 2014 and 2016 respectively. He is currently working as a researcher at Department of Computer Science, Chonnam National University, South Korea. His research interests are in signal processing, biomedical engineering, machine learning & pattern recognition.