DETECTION OF LESIONS IN MAMMOGRAMS USING LEARNED DICTIONARIES AND SPARSE REPRESENTATIONS

J. Herredsvela, K. Engan, T. O. Gulsrud, and K. Skretting

University of Stavanger, Stavanger, Norway

jostein.herredsvela@uis.no

Abstract: A novel method for detecting lesions in digital mammograms is presented. The method relies on image texture segmentation based on sparse representation of image blocks using learned dictionaries. Two different dictionaries are used to separate the suspicious tissue from the background tissue. By using a vector augmentation technique, the brighter regions are emphasized. The proposed scheme is trained and tested on mammograms from the MIAS public database. In 14 mammograms containing 15 circumscribed lesions the method was capable of detecting all but one lesion.

Introduction

Mammographic screening is widely considered to be the most reliable method for the early detection of breast cancer. To this end, large-scale mammographic screening programs are currently running in a large number of countries. In Norway, all women between 50 and 69 years of age are invited to participate in the national screening program. The probability of dying of breast cancer has been reduced by 30 % for women participating in the screening program [1].

Mammograms are x-ray projections of the breast tissue onto a detector array or a film plate. Tumors often consist of dense tissue, and thus absorb most of the incident x-rays. They can therefore often be seen as bright regions in the mammograms.

The screening programs generate a vast number of mammograms which are to be carefully examined, usually by two radiologists. This is a costly and time consuming process. A major concern is the number of false negative errors, i.e. cases in which a mammogram containing a malignant tumor is classified as normal. It is seen that between 10 and 30 % of cancers are missed during routine screening [2].

The above points have led to a large interest in Computer Aided Detection (CAD) of breast cancer. Research started in the early 1990's, and lately a few commercial systems have become available. Most CAD systems are intended to give the radiologists a second opinion of the suspicious regions in mammograms. One problem with CAD systems is that "perfect" detection (i.e. detection of all tumors present) in practice leads to false positives (FPs), i.e. regions of normal tissue detected (and classified as tumors). Too many false positives may confuse the radiologist.

There are several factors that make breast cancer detection in mammograms a very difficult task in image analysis. There is a large variation in the appearance of both normal breast tissue and of cancerous tissue. Some breasts have very dense or glandular parenchymal tissue that is radiopaque, while other breasts are mostly fatty and therefore radiolucent. There are several types of breast abnormalities that are visible in mammograms: Asymmetry between the breasts, calcifications, increase in breast tissue density, masses/lesions, and architectural distortions. CAD performance for microcalcifications is on an acceptable level, but the performance for lesions of various types is poorer.

The lesion class includes circumscribed lesions/masses, which are compact with lobular or circular/oval shape, and spiculated/stellate lesions which consist of a central mass with radiating spicules in some or many directions. In the present work we have so far focused on the circumscribed or well-defined lesions.

The last years have seen little progress in CAD performance. Most methods found in the literature have been based on extracting various textural or morphological features from the image pixels and combining these in various ways.

We believe much can be gained by considering the pixel values directly, i.e. without extracting any features. The work presented in this article is to the best of the authors' knowledge the first attempt to use learned dictionaries for texture classification of digital mammograms.

Materials and Methods

Four training mammograms were used for learning the dictionaries used in the classification. These and the 14 test mammograms were taken from the MIAS database provided by the Mammographic Image Analysis Society (MIAS) in the UK [3]. Each

ISSN: 1727-1983

mammogram contains at least one circumscribed lesion. The mammograms were downsampled to $1/16 \times 1/16$ of their original size before classification.

The proposed method segments/classifies the input mammogram into suspicious and non-suspicious regions (i.e. normal breast tissue) using two learned dictionaries.

Below we give a summary of the theory of learned dictionaries.

Any N-dimensional vector can be written as a linear combination of $K \geq N$ vectors that span the space. A good approximation to an N-dimensional signal vector $\mathbf{x}$ can often be obtained by linearly combining only a few of these K vectors. Mathematically, $\mathbf{x} = \mathbf{F}\mathbf{w} + \mathbf{n}$, where $\mathbf{F}$ is a learned frame/dictionary [4] in the form of an $N \times K$ ($K \geq N$) matrix, and where $\mathbf{w}$ is a sparse coefficient vector. $\mathbf{n}$ represents the approximation error. The columns of $\mathbf{F}$ are *dictionary vectors*. We emphasize that $\mathbf{F}$ must be *learned*, contrary to the case of orthogonal expansions. A properly trained dictionary results in a small representation error even using a very sparse representation, if the test vectors are reasonably similar to the vectors used when learning the dictionary. In general the error is larger for vectors that differ much from the the vectors used when learning it.

Given a collection of L vectors that are to be approximated using $\mathbf{F}$, an $N \times L$ data matrix $\mathbf{X}$ can be formed, of which the columns are the collection of vectors, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_L]$. Then the representation can be written

$$\mathbf{X} = \mathbf{F}\mathbf{W} + \mathbf{N},$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_L]$.

The *training matrix* from which $\mathbf{F}$ is to be learned is denoted $\mathbf{Y}$. In order to be able to make a good approximation to $\mathbf{X}$ it is important that $\mathbf{Y}$ has more or less the same "qualities" as $\mathbf{X}$. Learning $\mathbf{F}$ implies minimizing the representation error $||\mathbf{Y} - \mathbf{F}\mathbf{W}||$ subject to $\mathbf{W}$ being sparse [4]. The sparsity constraint is that only s dictionary vectors may be used in the representation of each of the columns in $\mathbf{X}$. In this work the Method of Optimal Directions (MOD) is used for learning [4]. The algorithm starts with a user supplied initial dictionary $\mathbf{F}^{(0)}$ and then improves it by iteratively repeating two main steps:

1. $\mathbf{W}^{(i)}$ is found by vector selection (ORMP is used in this work) using dictionary $\mathbf{F}^{(i)}$, where the objective function to be minimized is $J(\mathbf{W}) = \|\mathbf{Y} - \mathbf{F}^{(i)}\mathbf{W}\|^2$.

2. $\mathbf{F}^{(i+1)}$ is found from $\mathbf{Y}$ and $\mathbf{W}^{(i)}$, where the objective function is $J(\mathbf{F}) = \|\mathbf{Y} - \mathbf{F}\mathbf{W}^{(i)}\|^2$. This gives:

$$\mathbf{F}^{(i+1)} = \mathbf{Y}(\mathbf{W}^{(i)})^T(\mathbf{W}^{(i)}(\mathbf{W}^{(i)})^T)^{-1}$$

Then we increment i and go to step 1.

i is the iteration number. The first step is suboptimal due to the use of practical vector selection algorithms, while the second step finds the $\mathbf{F}$ that minimizes the objective function.The dictionary $\mathbf{F}$ is now *learned*, and it may now be used to approximate or represent any column vector in $\mathbf{X}$ using a linear combination of s of the dictionary vectors.

Sparse representations of image blocks by learned dictionaries are used to classify each mammogram into suspicious regions and non-suspicious regions. A dictionary-based approach has been successfully used in classification of various textures in [5]. Many segmentation/classification algorithms partition an image into regions that are similar according to a predefined set of criteria. We learn one dictionary for the lesion class and one for the normal tissue class. Our criterion for e.g. a lesion region is that the image blocks within it are represented better using the lesion class dictionary than the normal class dictionary.

The first task is to obtain good training sets for both classes. As already mentioned the four training mammograms are downsampled to $1/16 \times 1/16$ of their original size. True lesion regions in the training images are found by performing a watershed segmentation [6] followed by an inspection to ensure that only the true lesion is extracted. A large number, L, of overlapping image blocks of dimension $n \times n$ from the correct lesion regions are reshaped into training vectors of dimension $N = n \times n$ and collectively used as an $N \times L$ training matrix $\mathbf{Y}_1'$ for the lesion class dictionary. Since the breast tissue as seen in mammograms often have a dominant direction all image blocks are rotated 90°, 180°, and 270°, prior to reshaping them into training vectors. Ending up with dictionaries with directional qualities is thus avoided. The unrotated blocks are used as well. In addition to blocks from the interior of each region, image blocks extending a few pixels outside the region boundaries of the true lesions are used. We augment the training matrix with an extra row containing L elements of value z. The new matrix is denoted $\mathbf{Y}_1$:

$$\mathbf{Y}_1 = \begin{bmatrix} \mathbf{z} \\ \cdots \\ \mathbf{Y}_1' \end{bmatrix},$$

where $\mathbf{z}$ is a constant vector with equal element values z. A discussion of the matrix augmentation is given below.

The training vectors for the normal tissue class dictionary are reshaped image blocks from normal tissue regions in the same four mammograms. The vectors are reshaped into a training matrix $\mathbf{Y}_2'$ which is augmented using precisely the same row vector as used for $\mathbf{Y}_1'$ yielding a second training matrix $\mathbf{Y}_2$:

$$\mathbf{Y}_2 = \begin{bmatrix} \mathbf{z} \\ \cdots \\ \mathbf{Y}_2' \end{bmatrix}.$$

ISSN: 1727-1983

K randomly picked vectors from the lesion class training matrix $\mathbf{Y}_1$ are now normalized and used as column vectors in the initial dictionary associated with this class. The initial dictionary associated with normal breast tissue is formed in the same manner, but now the K vectors are picked from the normal class training matrix $\mathbf{Y}_2$.

As already mentioned the dictionaries have normalized columns. This is why the vectors must be augmented prior to normalization. Consider the situation that one initial dictionary is created from a certain set of vectors and that a second initial dictionary is created from the same set of vectors multiplied by a constant. After normalization the two dictionaries will be exactly the same. Augmenting the vectors prior to normalization results in two differing dictionaries. Returning to the case of texture classification of mammograms, it is known that the average gray level value is higher for most lesions than for normal tissue. Augmenting the vectors ensures that the dictionaries respond differently to dark and bright regions. Some classification results with different values of the augmentation element z are given in the next section.

The initial dictionaries are now used as input to the MOD algorithm described previously. The resulting dictionaries are denoted $\mathbf{F}_1$ and $\mathbf{F}_2$, respectively.

The subsequent classification step is pixel-based, i.e. one pixel and its neighborhood is considered at a time. The pixels within each such image block are reshaped into a test vector $\mathbf{x}$ and augmented using the same element value z as for the training vectors. The test vector is first represented using the lesion class dictionary, yielding the representation error $r_1 = ||\mathbf{x} - \mathbf{F}_1\mathbf{w}_1||$. Then the normal class dictionary is used, resulting in the representation error $r_2 = ||\mathbf{x} - \mathbf{F}_2\mathbf{w}_2||$. The center pixel in the image block under consideration is assigned class 1 (lesion) if $r1 < r2$, otherwise it is assigned class 2 (normal tissue).

Results

The pixels of class 1 constitute noisy images of suspicious regions that must be enhanced, see Figure 1(b) for an example. The images are median filtered using a circular neighborhood of radius R, thus removing isolated points and resulting in connected regions. See Figures 1(c) and 1(d) for results obtained with $R = 2$ and $R = 6$. The parameters used were $z = 250$, block size $n = 9$, $K = 164$, and $S = 3$. We see that the number of regions is reduced as we increase the median filter radius R. Using a too high radius may result in loss of small true detections. For all images shown in this article, the true lesion is located within the green truth circle. For all images the pectoral muscle is classified as lesion tissue (see e.g. the lower left corner of the images in Figure 1). There is the possibility to remove such detections, but in rare cases there may exist lesions in the muscle region. It might be necessary to treat this region separately.

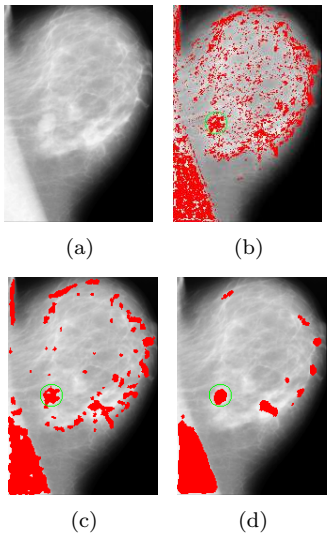


Figure 1: Classification results obtained using various values of the median filter radius R with $z = 250$, block size $n = 9$, total number of dictionary vectors $K = 164$ and $S = 3$ vectors in the linear combination. (a): Mammogram *mdb019ll*. (b): Result obtained without median filtering. (c): Result obtained using $R = 2$. (d): Result obtained using $R = 6$.

As stated in the introduction, one problem with CAD of mammograms in the relatively high number of False Positives (FP), i.e. the number of detections that are not lesions. The presented method is no exception; not all the regions classified are true lesions.

Results obtained with various values of z are shown in Figure 2. The parameters used were $R = 4$, $n = 9$, $K = 164$, and $S = 3$. We see that using $z = 1$ several regions that are smooth are classified as lesions even though they have a low average gray level. Increasing z to 50 and 250 gradually suppresses these regions, instead emphasizing regions of higher average gray level.

The effect of changing the size n of the image blocks is illustrated in Figure 3. We used $z = 250$, $R = 4$, $K = 164$, and $S = 3$. It is seen that using $n = 9$ gives a smaller number of regions than $n = 7$. In addition, the region follows the true outline of the lesion better.

The effect of varying the total number of dictionary vectors K is shown in Figure 4. $z = 250$, $R = 4$, and $S = 3$ was used. The results are in general different for the two K values. The small red region to the right of the green truth circle in Figure 4(d) is actually a lesion detected with $K = 100$. Figure 4(f) shows that the lesion is missed when $K = 40$ is used. This indicates that it probably is a good idea to use overcomplete dictionaries, i.e. the number of dictionary vectors K is larger than the dimension of the vectors N.

IFMBE Proc. 2005 11(1)
ISSN: 1727-1983 © 2005 IFMBE

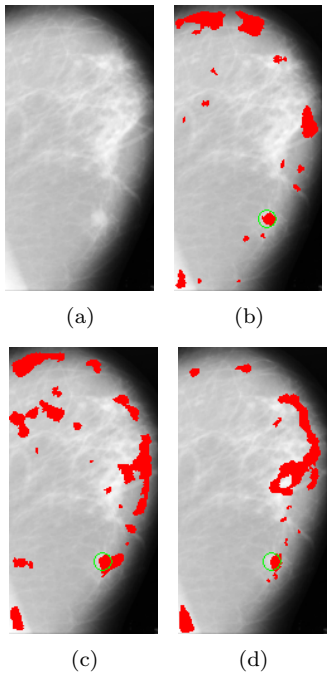


Figure 2: Classification results obtained using various values of z with median filter radius $R = 4$, block size $n = 9$, total number of dictionary vectors $K = 164$ and $S = 3$ vectors in the linear combination. (a): Mammogram *mdb023ll*. (b): Result obtained using $z = 1$. (c): Result obtained using $z = 50$. (d): Result obtained using $z = 250$.

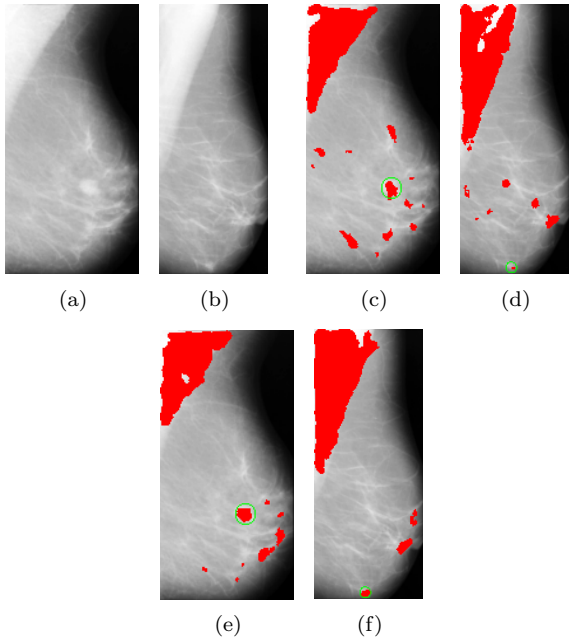


Figure 3: Classification results obtained using $n = 7$ and $n = 9$ with $z = 250$ with median filter radius $R = 4$, total number of dictionary vectors $K = 164$ and $S = 3$ vectors in the linear combination. (a): Mammogram *mdb005ll*. (b): Mammogram *mdb080rm*. (c): Result for *mdb005ll* obtained using $n = 7$. (d): Result for *mdb080rm* obtained using $n = 7$. (e): Result for *mdb005ll* obtained using $n = 9$. (f): Result for *mdb080rm* obtained using $n = 9$.

13 of the 14 test mammograms contain one lesion each while one mammogram contains two lesions. 14 of the 15 lesions are detected using $n = 9$, $z = 250$, $S = 3$, $K = 164$, and $R = 4$. A relatively good segmentation quality was achieved. The lesion not correctly detected had low contrast and was located in very dense tissue, see Figure 5.

Conclusions

A novel method of classifying mammograms into suspicious and non-suspicious regions has been presented. The results presented are promising. 14 of 15 lesions were detected with relatively high segmentation quality. There is, however, a number of false positives (FP) present in all the images. An FP reduction technique should be developed, perhaps one based on another set of learned dictionaries. For mammograms of dense breasts there is at present a potential for improvement. The method has only been tested on a limited number of mammograms, and needs testing on a larger database to provide more reliable sensitivity/specificity data. Given the promising results obtained for circumscribed lesions the method should be thoroughly tested on more subtle types of lesions, e.g. stellate lesions. Interesting problems involving more than two texture classes should be formulated and tested.

References

[1] *www.kreftregisteret.no*

[2] BIRD R. E., WALLACE T. W., and YANKASKAS B. C. (1992): 'Analysis of cancers missed at screening mammography', Radiology, 184(3), pp. 613–617

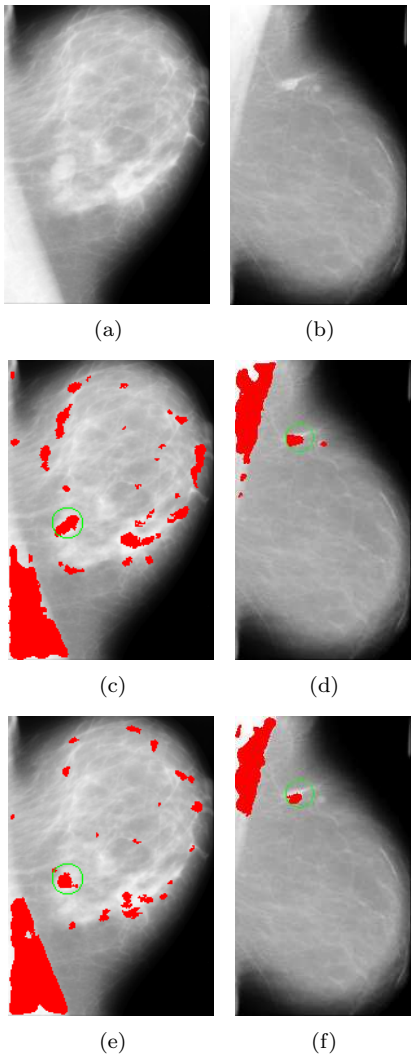
[3] *www.wiau.man.ac.uk/services/MIAS/MIAScom.html*

[4] ENGAN K., AASE S. O., and HUSØY J. H. (2000): 'Multi-frame compression: Theory and design', Signal Processing, (80), pp. 2121–2140

[5] SKRETTING K. and HUSØY J. H. (2005): 'Texture classification using sparse frame based representations', Eurasip JASP, Submitted for publication

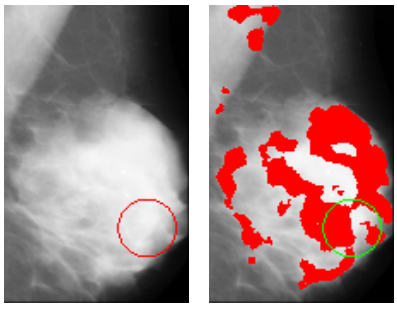
[6] HERREDSVELA J., GULSRUD T. O., and ENGAN K. (2005): 'Detection of circumscribed masses in mammograms using morphological segmentation', Proc. SPIE, volume 5747, pp. 902–913

ISSN: 1727-1983



(a) (b) (c) (d) (e) (f)

Figure 4: Classification results obtained using $K = 40$ and $K = 100$, with $z = 250$, median filter radius $R = 4$, and $S = 3$ vectors in the linear combination. (a): Mammogram *mdb019ll*. (b): Mammogram *mdb132rx*. (c): Result for *mdb019ll* obtained using $K = 100$. (d): Result for *mdb132rx* obtained using $K = 100$. (e): Result for *mdb019ll* obtained using $K = 40$. (f): Result for *mdb132rx* obtained using $K = 40$.



(a) (b)

Figure 5: The mammogram *mdb010rm* for which the lesion is not detected. (a): Mammogram and truth circle. (b): Mammogram and detected regions. $K = 164$, $n = 9$, $R = 4$, $S = 3$, and $z = 250$.

ISSN: 1727-1983