Computer-aided detection of masses in full-field digital mammography using screen-film mammograms for training

This content has been downloaded from IOPscience. Please scroll down to see the full text.

View the table of contents for this issue, or go to the journal homepage for more

Download details:

IP Address: 54.191.40.80
This content was downloaded on 29/08/2017 at 20:23

Please note that terms and conditions apply.

You may also be interested in:

The effect of feature selection methods on CAD of masses in mammograms
Rianne Hupse and Nico Karssemeijer

Comparison of computer-aided detection systems on FFDM and SFM
Jun Ge, Lubomir M Hadjiiski, Berkman Sahiner et al.

Use of border information in the classification of mammographic masses
C Varela, S Timp and N Karssemeijer

Computer-aided detection (CAD) of breast masses in mammography: combined detection and ensemble classification
Jae Young Choi, Dae Hoe Kim, Konstantinos N Plataniotis et al.

Automatic mass detection in mammography
Guido M te Brake, Nico Karssemeijer and Jan H C L Hendriks

Multiresolution local binary pattern texture analysis combined with variable selection for application to false-positive reduction in computer-aided detection of breast masses on mammograms
Jae Young Choi and Yong Man Ro

A novel featureless approach to mass detection based on SVM
Renato Campanini, Danilo Dongiovanni, Emiro Lampieri et al.
Computer-aided detection of masses in full-field digital mammography using screen-film mammograms for training

Michiel Kallenberg and Nico Karssemeijer

Department of Radiology, Radboud University Nijmegen Medical Centre, Geert Grooteplein Zuid 18, 6525 GA Nijmegen, The Netherlands

E-mail: m.kallenberg@rad.umcn.nl

Received 24 April 2008, in final form 17 October 2008
Published 12 November 2008
Online at stacks.iop.org/PMB/53/6879

Abstract

It would be of great value when available databases of screen-film mammography (SFM) images can be used to train full-field digital mammography (FFDM) computer-aided detection (CAD) systems, as compilation of new databases is costly. In this paper, we investigate this possibility. Firstly, we develop a method that converts an FFDM image into an SFM-like representation. In this conversion method, we establish a relation between exposure and optical density by simulation of an automatic exposure control unit. Secondly, we investigate the effects of using the SFM images as training samples compared to training with FFDM images. Our FFDM database consisted of 266 cases, of which 102 were biopsy-proven malignant masses and 164 normals. The images were acquired with systems of two different manufacturers. We found that, when we trained our FFDM CAD system with a small number of images, training with FFDM images, using a five-fold crossvalidation procedure, outperformed training with SFM images. However, when the full SFM database, consisting of 348 abnormal cases (including 204 priors) and 810 normal cases, was used for training, SFM training outperformed FFDM training. These results show that an existing CAD system for detection of masses in SFM can be used for FFDM images without retraining.

1. Introduction

Breast cancer is the most common cause of cancer death in women worldwide (Ferlay et al 2004). Nowadays it is generally accepted that early detection by screening increases the chance of survival. Mammography is one of the best available screening tools to detect breast cancer. However, studies indicate that many invasive breast cancers remain undetected at screening. A substantial fraction of these undetected cancers were visible in retrospect on
previous mammograms. Computer-aided diagnosis (CAD) systems have been developed to aid radiologists in the screening. These systems are able to detect and characterize breast masses and microcalcifications.

Most computer-aided detection (CAD) systems are currently developed for and based on screen-film mammography (SFM). In the last few years, however, the use of full-field digital mammography (FFDM) has increased rapidly in clinical practice because of the advantages of digital storage and because digital imaging can potentially improve breast cancer detection. FFDM has many advantages compared to SFM: in FFDM the stages of image formation can be optimized individually as image acquisition, processing and display are separated. FFDM provides higher signal-to-noise ratio, detective quantum efficiency and higher contrast sensitivity than SFM. In addition, FFDM enables soft-copy reading in which CAD is easily implemented (Pisano and Yaffe 2005).

Several clinical studies researched the performance of radiologists’ interpretation on FFDMs and SFMs. Pisano et al (2005) found that performance was comparable for both modalities when all women were taken into account. However, for women under the age of 50, women with radiographically dense breasts, or premenopausal or perimenopausal women FFDM was more accurate. Vigeland et al (2008) found that FFDM had a significantly higher detection rate for DCIS than SFM ($P < 0.001$). Skaane et al (2007) found that FFDM enabled a significantly higher cancer detection rate ($P = 0.02$). The interval cancer rate was lower for FFDM than for SFM, but the difference was not significant ($P = 0.35$).

As FFDM is becoming more prevalent, the development of CAD systems that can deal with FFDM images is needed. So far, two research groups have reported on the development of a non-commercial FFDM CAD system (Li et al 2002, Wei et al 2005, 2007). Both systems are constructed by means of adaptations of a previously developed SFM CAD system. Next to these adaptations, which occur mainly in the preprocessing and prescreening stage, each system had been retrained with FFDM images to get the best result.

In order to develop a sound CAD system, a large database containing training samples is of paramount importance. Yet, as FFDM is not as widely used as SFM it is difficult to collect a sufficient amount of malignant abnormalities. Therefore, it would be of great value when the available databases of SFM images can be used to train FFDM CAD systems. One could for instance use a mix of SFM images and FFDM images as a training database. Retraining the CAD system with a small database of FFDM images would then not be necessary. In this paper, we examine the possibility of using SFM images as training images. We will focus on masses, architectural distortions and asymmetries. These will all be referred to as masses in the rest of the paper.

In order to optimize the performance of the CAD system, we have to ensure that the characteristics of the test images match the characteristics of the images with which the system is trained. As we train our system with SFM images we develop a method that converts the FFDM test images into SFM-like representations. The key point in this conversion method is the characteristic curve which describes the relationship between log relative exposure and optical density for an SFM image. In a raw FFDM image pixel values are linearly related to exposure values, regardless of the type of manufacturer. Therefore, apart from a constant scaling factor, the exposure values can be extracted from the raw FFDM images, which can then be converted into an SFM-like representation. Parameters of the conversion are determined for each individual image, using a method that is inspired by the automatic exposure procedure as implemented in conventional mammography (Karssemeijer et al 2005, Snoeren and Karssemeijer 2007, Kallenberg and Karssemeijer 2008b).

Secondly, we investigate the effect of using the SFM images as training samples as compared to training with FFDM images. In particular we are interested in CAD performance
operating at a high specificity, as optimizing CAD performance in this range might be most useful for clinical practice. We expect that with an identical amount of available images, FFDM training will be better. However as FFDM databases are normally small, we determine the number of FFDM images needed to train a CAD system with an acceptable performance.

2. Materials and methods

2.1. CAD system

Our CAD system consists of four stages: (1) preprocessing, (2) scanning of mass candidates, (3) region segmentation and (4) final classification. The details of the system can be found in previous publications (Karssemeijer and te Brake 1996, Karssemeijer 1998, te Brake and Karssemeijer 1999, Timp and Karssemeijer 2004). In the following section, we describe the essentials that are needed to understand the experiments we conducted.

In the first stage, the mammogram is segmented into the breast area and the background. In addition peripheral enhancement is applied, while in the MLO views the pectoralis is equalized (Karssemeijer 1998). In the second stage, four features are calculated on a regular grid with 1.6 mm of spacing. The features that are computed are related to the presence of spicules (Karssemeijer and te Brake 1996) and a central mass (te Brake and Karssemeijer 1999). The spiculation features are based on the orientation of lines in the neighborhood of the pixel of interest. Line-based orientation estimates are obtained from the output of directional, second-order Gaussian derivative operators at each point in the image. The orientation at which these operators have maximum response is selected. The pixel orientation map that results is used to construct two operators which are sensitive to radial patterns of straight lines. As orientations are insensitive to changes of image contrast, this procedure makes the detection robust for differences in pixel range and contrast among manufacturers and mammograms. The features for central mass detection are computed by a similar procedure based on first-order derivatives. The four features that result are subsequently combined with an ensemble of five neural networks (NN1). In this way a likelihood image is constructed. The local maxima in this image are taken as locations of interest.

In the third stage, the locations of interest are used as seed points for region segmentation. Region segmentation is based on dynamic programming (Timp and Karssemeijer 2004). In the fourth stage, features are calculated for each region found in stage three. In total seventeen features are selected: seven features that were computed at the initial stage, three location features and seven features derived from the segmented region (region size, three contrast measures, acutance, pectoral overlap and background density). An ensemble of five neural networks (NN2) is used to predict the malignancy of the regions.

In the CAD system two (ensembles of) neural networks are used: NN1 is used to scan pixels for mass candidates and NN2 is used to classify selected regions. Both neural networks are feedforward neural networks which consist of an input layer of \( n \) nodes representing the features, where \( n \) is the number of features, a hidden layer of eight nodes and an output layer of one neuron. The networks are trained by means of the backpropagation algorithm. Before training, features are normalized to zero mean and unit variance using all images in the training set. Each ensemble consists of five networks. The five networks differ by their initial weights and the order at which the patterns are presented to the training. Furthermore, the number of times a certain pattern is used for training differs among the networks as the samples are randomly drawn from the training database with repetitions.

The output of the networks NN1 and NN2 is a real number in the interval \([0–1]\) indicating the likelihood of malignancy. For each network this output is converted by a monotonously
decreasing function to obtain a normality score, which is defined as the number of false positives per image that would result when all mass candidates (NN1) or regions (NN2) with this score or lower would be marked. The look-up table that converts the network output into the normality score is calculated by applying the classifier to the normal images in the training set and determining the average number of mass candidates (NN1) or regions (NN2) per image marked by the system as a function of a threshold applied to the output level of the network (van Engeland and Karssemeijer 2007).

In pilot experiments we determined the optimal settings of the networks. For NN1 best results were obtained with a learning rate of 0.03, a proportion of abnormal to normal training patterns ($p_{\text{abn}}$) of 0.01 and a training length $L$ of $2 \times 10^6$ patterns. For NN2 we used a learning rate of 0.01, a $p_{\text{abn}}$ of 0.2 and an $L$ of $5 \times 10^5$ patterns. The output of NN1 was used to determine whether a mass candidate was further processed. The best results were obtained when the threshold for further analysis was set to a normality score of 3.5. Candidates with a higher normality score were classified as normal by the CAD system.

A free-response operator characteristic (FROC) curve was computed to measure performance. Because the output of the classifier was normalized the results of the subsets could be pooled to obtain one FROC curve representing independent test performance. Two types of FROC curves were computed: lesion-based and case-based curves. For case-based sensitivity it was regarded a true positive when the malignant mass was detected in at least one view, for lesion-based sensitivity views were treated independently. A detection was counted if the center of a mass of the region marked by CAD was inside a true mass region. When the center of the CAD region was outside a true mass region a false positive was counted. To determine the false positive rate only normal cases were considered. The mean sensitivity in the interval between 0.05 and 0.5 false positive (FP) marks/image served as a performance measure. To avoid domination of performance at high false positive rates the mean was calculated on a logarithmic scale using the following equation:

$$S = \frac{1}{\ln(0.5) - \ln(0.05)} \int_{0.05}^{0.5} \frac{s(f)}{f} \, df$$

where $s$ is the sensitivity and $f$ is the false positive rate.

2.2. Experiments

2.2.1. Datasets. The FFDM database used in this study consisted of 266 cases, of which 102 were biopsy-proven malignant masses and 164 normals. Each case consisted of four images: two craniocaudal views (CC) and two mediolateral oblique (MLO) views, making a total of 1064 images. The cases were collected from two medical centers in the Netherlands. Two hundred and six cases were collected from the Radboud University Nijmegen Medical Centre where they were acquired with a GE Senographe 2000D FFDM system. Sixty cases were collected from Preventicon, Utrecht, where they were acquired with a Hologic Selenia FFDM system. The GE system had a CsI phosphor/a:Si active matrix flat panel digital detector with a pixel size of 100 $\mu$m $\times$ 100 $\mu$m and 14 bits per pixel. The detector of Hologic Selenia FFDM system consisted of a photoconductor of amorphous selenium and a TFT array with a pixel size of 70 $\mu$m and 14 bits per pixel. All FFDM images were downsampled to a resolution of 200 $\mu$m by means of bilinear interpolation.

The SFM database comprised 1362 cases, containing 552 biopsy-proven malignant masses (including 204 priors) and 810 normals. The cases were obtained from multiple screening institutions in the Netherlands and they were acquired with a variety of mammography systems. Not all cases had four views because usually MLO views are only made at subsequent
screenings in the Netherlands. Six hundred and eighty-eight cases were digitized with a Canon CFS300 scanner; 674 cases with a Lumisys 85 scanner. Both scanners were operated with a pixel resolution of 50 μm and 12 bits per pixel. All images were downsampled to a resolution of 200 μm.

In our experiments we compared the performance of FFDM training with SFM training. For a proper comparison between both training procedures, we selected a subset of 1064 SFM images from the SFM database. In figure 1, the composition of the datasets is shown with regard to the types of masses. The sets are almost similar in composition: in both sets most of the malignant masses are spiculated (±40%) or ill-defined (±40%). Both sets contain solely images with two views and do not contain priors.

2.2.2. FFDM conversion. Our first experiment dealt with the conversion between an FFDM image and an SFM-like representation. Crucial in this conversion method is the simulation of the acquisition procedure of an SFM image.

In screen-film mammography, x-rays are absorbed by a scintillator which sends the signal as visible light to a film. After development of the film, exposure values are represented by optical density. The (nonlinear) relation between optical density and x-ray exposure is described by the characteristic curve (Hurter and Driffield 1890). In full-field digital mammography exposure values are collected with digital detectors. In these systems, the relation between x-ray exposure and pixel value is linear for raw images. Because of this linear relation the conversion between an FFDM image and an SFM-like representation can be accomplished by applying the characteristic curve to the raw image.

In this paper, we modeled the characteristic curve by

\[
\text{od}(E) = \text{od}_{\text{min}} + \frac{\text{od}_{\text{max}}}{1 + \left( \frac{E}{E_s} \right)^g}
\]  

(2)
where $E$ denotes the x-ray exposure and $\text{od}(E)$ the optical density. The parameters $s$ and $g$ represent the speed and gradient of the film, respectively.

For the parameter $g$ we selected values that match the characteristic curve of the Kodak MinR-2000 and Agfa HDR system. Effects of variation of $g$ were studied experimentally. For $\text{od}_{\text{min}}$ and $\text{od}_{\text{max}}$ we took 0.18 and 3.8, respectively.

To find an acceptable value for parameter $s$, we developed a method inspired by the acquisition procedure of an SFM image. In SFM the optical density of the (most dense part of) the breast is kept constant over mammograms. This is regulated by the automatic exposure control (AEC). The AEC is a device that monitors the exposure at the image receptor level, situated under the exposed breast. If the exposure exceeds a certain threshold the AEC sends a stop signal to the x-ray tube, thereby ensuring that the film is not under or overexposed. With this procedure the incident beam is adjusted to the density and thickness of each individual breast, which ensures that the exposure values of dense tissue are kept constant (Pisano and Yaffe 2005, Meeson et al 2001, Barnes 1993).

Obviously, in conventional systems the speed $s$ of the film/screen system is fixed and the exposure time is varied. For converting an FFDM image into an SFM-like representation we have to deal with a fixed exposure time, but we can vary the parameter $s$ of the characteristic curve. As only the fraction $E/s$ plays a role in the conversion, variations of speed and exposure time have the same end result.

In our method, we mimicked the AEC by scanning the FFDM image for the most dense part. A moving window of size $A_w \times A_h$ cm is placed on each mammogram to find the area with the smallest mean pixel value, $y_{\text{AEC}}$. For $A_w$ and $A_h$ we took 2.0 and 4.0 cm, respectively. In figure 2, the AEC simulation is illustrated. In this figure, the AEC window is represented
Utilizing SFM images in an FFDM CAD system

Figure 3. Overview of the conversion between a raw image and an SFM-like representation. In the left panel raw images are shown for a fat and dense mammogram. For each mammogram the parameter $s$ of the filmcurve is calculated by means of the AEC simulation (see figure 2), $y_{AEC}$, and thereby $s$ is relatively small for the dense mammogram. As a result the filmcurve, as shown in the middle panel, is shifted to the left. The shifting ensures that the dense parts of the mammograms are depicted at a constant optical density.

by the checked square. One can observe that the moving of the AEC window is restricted to a bounding box. The position and size of the bounding box are linked to the size of the image, ensuring that the box is placed on the breast, while it does not touch the pectoralis.

The minimal mean pixel value, determined by the AEC simulation, $y_{AEC}$, is used to fit the characteristic curve to the raw image. By applying the equation

$$s = y_{AEC} \left(1.1 \left( \frac{od_{max} - (od_{des} - od_{min})}{(od_{des} - od_{min})} \right) \right)^{\frac{1}{7}},$$

the curve is shifted and thereby adjusted to the mammogram, ensuring that the dense parts of the mammograms are represented with a constant optical density $od_{des}$. In our experiment we used an optical density $od_{des}$ of 1.2, which corresponds to the mean optical density value in dense areas in our training database.

The optical density value is subsequently converted to pixel value $y$ by applying the digitization equation of the calibrated digitizers used in the study:

$$y(od) = 4096 \left( \frac{4.1 - od}{4.1} \right).$$

Figure 3 shows an overview of the conversion between a raw image and an SFM-like representation. In the left panel, the raw image is depicted of a fatty and a dense mammogram. For each mammogram the parameter $s$ is calculated by means of the AEC simulation. In the dense mammogram the mean pixel value of the AEC region (filmcurve parameter $y_{AEC}$
(and thereby $s$) is relatively small. As a result the conversion curve is shifted to the left, which is shown in the middle panel. The resulting images are depicted in the right panel.

The effect of the parameter $g$ on the performance of mass detection was investigated experimentally. In this experiment, we used our CAD system trained on all images in the SFM database. As a test set we used all images in our FFDM database. To investigate the effect of the parameter $g$ each FFDM image was converted with the characteristic curve with gradient $g$. The converted image was then subjected to the CAD algorithm, after which FROC curves were computed. As a performance measure the mean sensitivity was determined in the interval between 0.05 and 0.5 FP marks/image, by applying equation (1). As we did this procedure for each gradient $g$ we were able to determine the effect of the parameter $g$ on the performance of mass detection.

2.2.3. Training. In a second experiment, we investigated the effect of using SFM images as training samples compared to training with FFDM images. To assess the effect of the training procedure we classified each mammogram two times. In the first case we used neural networks trained with SFM cases; in the second case the classifiers were trained with the FFDM images using a crossvalidation procedure. For this purpose the FFDM dataset was randomly split for each complete evaluation cycle in five subsets, making sure that different views of the same exam ended up in the same dataset. The FFDM images were converted into an SFM-like representation using the method described above. In this method the gradient parameter $g$ was set to 2.0.

In order to compare SFM training to FFDM training, one has to take into account the number of training samples the classifiers are trained with. We expected that the number of mass training samples would have a large effect on detection performance, so we varied the number of masses that were used in the training. This was done by randomly selecting masses from the four subsets used for training in the crossvalidation procedure, while keeping all normal images in the training set. This was done for each of the five runs in the crossvalidation process. Note that the composition of the training set did not affect the proportion of abnormal to normal training patterns at which the samples were presented to the neural networks. In the neural network training procedure, this proportion, defined as $p_{abn}$ in section 2.1, was kept constant. The whole cycle of training and testing was repeated eight times, in order to statistically validate the results of the experiments. In these repetitions we varied the distribution of the images over the partitions and initialized the neural networks with different weights. Subsequently, differences in $S$ were statistically tested using an unpaired Student $t$-test with the assumption of equal variance.

3. Results

3.1. FFDM conversion

Our first experiment dealt with the conversion between an FFDM image and an SFM-like representation using a fitted characteristic curve. To evaluate the effect of the parameters of the curve on the performance of mass detection, the parameter $g$ was varied. For each value of $g$ CAD performance was assessed using FROC curves. In the experiment, the CAD system was trained with SFM images.

Figure 4 shows the FROC curves for a range of values for the gradient ($g$) averaged over eight runs. In the figure, one can see that the CAD system is robust for changes of the parameter within the range of 1.5–3.0. The performance, measured as mean case-based sensitivity in the interval from 0.05 to 0.5 FP marks/image, ranges from 77.9% ($\pm$ 0.2%) (standard error of...
the mean (SEM)) for a g of 3.0 to 79.0% (± 0.3% (SEM)) for a g of 1.5. For a gradient of 5.0 the performance is worse. The mean sensitivity in the chosen interval drops to 69.2% (± 0.5% (SEM)). Lesion-based sensitivity shows the same pattern. Again, a gradient of 5.0 yields the worst results with a mean sensitivity of 53.0% (± 0.4% (SEM)), whereas at gradients between 1.5 and 3.0 the performance is best. Mean sensitivity ranges from 60.3% (± 0.8% (SEM)) for 1.5 to 61.8% (± 0.6% (SEM)) for 2.0. In the interval between 0.01 and 0.05 a gradient of 3.0 performs best.

To address the robustness of the CAD system for changes in the gradient of the filmcurve, we inspected to what extent the feature values varied with gradient changes. It was found that only one of the seventeen features in our CAD system, namely one of the region contrast measures, was affected by the gradient in a substantial manner. For this feature, a gradient between 2.0 and 3.0 matched the SFM images best. Other features showed only minor changes when the gradient was varied.

3.2. Training

Our second experiment dealt with the effects of using the SFM images as training samples compared to training with FFDM images. It was found that SFM training performed better, as shown in figure 5. When we trained the CAD system with SFM images the mean case sensitivity in the interval from 0.05 to 0.5 FP marks/image was 78.7% (±0.2% (SEM)), while training with FFDM images yielded 76.8% (±0.3% (SEM)). Lesion sensitivity was 61.9% (±0.2% (SEM)) for SFM training and 59.2% (±0.3% (SEM)) for FFDM training. Both differences were statistically significant ($p = 0.0003$ and $p < 0.0001$, respectively) as determined by an unpaired Student $t$-test, with equal variance. It should be remarked however
Case and Lesion Sensitivity

Figure 5. FROC curves showing the performance of FFDM training versus SFM training. The upper pair shows case sensitivity, while the lower pair shows lesion sensitivity. The curves are averaged over eight runs. The solid vertical lines represent the borders between which the mean sensitivity is calculated as a performance measure. As one can observe the performance of SFM training is slightly better.

Mean Sensitivity between 0.05 and 0.5 FP/image

Figure 6. The effect of the number of masses in the training database on the performance of mass detection and classification. The upper lines represent the case-based sensitivity, while the lower lines represent the lesion-based sensitivity. The error bars represent the standard error of the mean. It can be seen that the performance increases steadily while masses are being added to the training. Furthermore, it can be seen that FFDM training outperforms SFM training when the system is trained with the same number of masses. However when the full SFM database is used, SFM training outperforms FFDM training.
that with the same training procedures the SFM training needed more training images (552 abnormal (including 204 priors) and 810 normal cases) to achieve this result.

In order to investigate the effect of the number of training samples further, we evaluated the effect of adding masses to the training database. The performance of the CAD system is measured as mean sensitivity in the interval between 0.05 and 0.5 FP marks/image. In figure 6, the performance of the CAD system is depicted as a function of the number of masses in the training database. We found that performance steadily increases with the number of masses in the training set, although it may be noted that for SFM training the performance increase between 64 and 80 mass cases is less than expected. Interestingly, one can observe that FFDM training outperforms SFM training when the system is trained with the same number of malignant masses. For case-based sensitivity this effect was significant for 8, 16, 64 and 80 cases ($p = 0.02$, $p = 0.003$, $p = 0.0006$, $p < 0.0001$ (two tailed), respectively). For image-based sensitivity this effect was significant for 8, 16, 32, 64 and 80 cases ($p = 0.009$, $p = 0.001$, $p = 0.002$, $p = 0.0004$ and $p < 0.0001$ (two tailed), respectively). However, when the large SFM database, consisting of 552 malignant masses (including 204 priors) and 810 normals, is used to train the CAD system the advantage of a large database becomes clear. For both case- and lesion-based sensitivity SFM training outperforms FFDM training ($p = 0.0003$ and $p < 0.0001$ (two tailed), respectively), as shown in figures 5 and 6.

4. Discussion and conclusion

The CAD system was shown to be robust for changes of the gradient parameter within the range of 1.5–3.0, in the interval between 0.01 and 0.05 FP marks/image. In SFM the average image gradient (AG) is measured as the slope of the characteristic curve for ODs from 0.25 and 2.0 above base plus fog. Typical AG values lie between 3.0 and 4.0, which corresponds to gradients between 2.0 and 2.7 in our method (Tsalafoutas et al 2004). The sensitivity of the CAD system in the interval between 0.01 and 0.05 FP marks/image was highest when values of the gradient are in this range. It should be noted, though, that error bars in this area of the FROC curve are high, due to a low number of false positives that remain.

In previous research it was found that the proportion of malignant versus normal cases in the training set influences the classifier output. To overcome undesirable effects of unbalanced training sets we normalized the classifier output by a monotonously decreasing function. The look-up table that converts the network output into the normality score is calculated by applying the classifier to the normal images in the training set and determining the average number of mass candidates (NN1) or regions (NN2) per image marked by the system as a function of a threshold applied to the output level of the network. Note that this is also a common procedure in clinical CAD systems used in practice. It standardizes the CAD output, allowing adjustment of its specificity in a PACS environment. The normality scores are not used directly as false positive rate in the FROC curve. Instead FROC curve is computed by thresholding the normality scores and recomputing false positives and true positives at a range of threshold values. In pilot experiments we determined that 500 normal images are appropriate for the normalization procedure. In this paper, the number of images that was used to determine the normality score ranges from 540 to 684.

In this study, images were used from two manufacturers. It was concluded that the conversion method works well for both manufacturers. However, a direct comparison between the CAD performance on images of both imaging systems was not possible as the images were taken from different patients. We expect, however, that differences are small as we designed our features to be insensitive for changes in gain, offset and pixel range. An alternative
approach, though, to overcome differences in imaging systems, would be to transform the images to the standard mammogram form as introduced by (Highnam and Brady 1999). However, for this method multiple acquisition parameters should be available, as is generally not the case in film images.

We investigated the effect of using the SFM images as training samples compared to training with FFDM images. We found that FFDM training outperforms SFM training when a small number of mass cases were used for training. A plausible explanation for this result is the possible mismatch between the characteristics of the FFDM database and the SFM database, which can result in problems such as overfitting.

Possible differences can be grouped in two categories. The first category is related to differences in lesion characteristics: it should be noted that the mammograms in our SFM set are solely cases of the screening, whereas the FFDM set contains diagnostic images as well. In addition, the images in our FFDM set are cases from the first screening round, whereas the SFM cases were collected from multiple screening rounds. As a result, differences can occur in breast density and subtlety of the lesions. A second reason for a mismatch that fits in the first category is the distribution of the lesion types in both datasets. However, we saw in figure 1 that both sets are quite comparable with respect to lesion types.

The second category of differences deals with differences in image characteristics: first, FFDM is known to have superior characteristics compared to SFM. Second, the SFM images in our database are collected from multiple screening institutions and are taken over a range of years, whereas the FFDM cases are from two institutions, where they were collected in a short period of time. This implies that for SFM a large variety of filmcurves are used, whereas the FFDM images are subjected to the same filmcurve (apart from an offset). As variation in acquisition procedure causes a variation that is not related to the distinction between mass and normals, one might argue that the SFM database contains more noise, which hampers the training of a classifier.

In this study, we also investigated the effect of the number of masses in the database on the performance of mass detection. We found that performance increases steadily while masses were added to the training set. This result is in accordance with previous (simulation) studies (Chan et al 1999, 2004, Sahiner et al 2008, Fukunaga and Hayes 1998, Kallenberg and Karssemeijer 2008a). It may be noted that in the interval between 64 and 80 mass cases, for SFM training the increase in performance was less than expected. Apparently, at this point the 16 extra masses do not help the CAD system further. Possibly the 16 extra masses do not differ sufficiently from the 64 masses in the feature space, so that the classifier is not provided with additional information. However, when an extra amount of 620 masses is added (i.e. when the full database of 684 malignant masses (including 256 priors) and 1013 normals was used) the advantage of a large training database becomes clear. When the full SFM database is used for training, SFM training outperformed FFDM training. In future experiments, we will investigate the possibility of combining SFM training with FFDM training to increase the database further.

In conclusion, this study shows that training an FFDM CAD system with SFM images is possible and is, as such, a good alternative or extension to retraining with FFDM images, as has been done in current systems described in the literature.

References

Barnes G 1993 Mammography equipment: compression, scatter control, and automatic exposure control Syllabus: A Categorical Course in Physics (Oak Brook: RSNA publications) pp 59–68
Utilizing SFM images in an FFDM CAD system


Highnam R and Brady J 1999 *Mammographic Image Analysis* (Dordrecht: Kluwer)

Hurter F and Driffield V 1890 Photo-chemical investigations and a new method of sensitiveness of photographic plates *J. Soc. Chem. Industr.* 455–69


Li L, Clark R and Thomas J 2002 Computer-aided diagnosis of masses with full-field digital mammography *Acad. Radiol.* 9 4–12


Sahiner B, Chan H P and Hadjiiski L 2008 Classifier performance estimation under the constraint of a finite sample size: resampling schemes applied to neural network classifiers *Neural Netw.* 21 873–93


Tsafaloutas I A, Dimakopoulou A D, Koulentianos E D, Serferoglou A N and Yakounakis E N 2004 Variation of the sensitometric characteristics of seven mammographic films with processing conditions *Br. J. Radiol.* 77 666–71


