

Combining two mammographic projections in a computer aided mass detection method

Saskia van Engeland and Nico Karssemeijer^{a)}

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

(Received 13 July 2006; revised 12 December 2006; accepted for publication 31 December 2006; published 9 February 2007)

A method is presented to improve computer aided detection (CAD) results for masses in mammograms by fusing information obtained from two views of the same breast. It is based on a previously developed approach to link potentially suspicious regions in mediolateral oblique (MLO) and cranio-caudal (CC) views. Using correspondence between regions, we extended our CAD scheme by building a cascaded multiple-classifier system, in which the last stage computes suspiciousness of an initially detected region conditional on the existence and similarity of a linked candidate region in the other view. We compared the two-view detection system with the single-view detection method using free-response receiver operating characteristic (FROC) analysis and cross validation. The dataset used in the evaluation consisted of 948 four-view mammograms, including 412 cancer cases with a mass, architectural distortion, or asymmetry. A statistically significant improvement was found in the lesion based detection performance. At a false positive (FP) rate of 0.1 FP/image, the lesion sensitivity improved from 56% to 61%. Case based sensitivity did not improve. © 2007 American Association of Physicists in Medicine. [DOI: 10.1118/1.2436974]

Key words: breast cancer, computer aided detection (CAD), masses, mammography

I. INTRODUCTION

In x-ray mammography it is standard practice to image the breast from two different angles. The most common projections are the mediolateral oblique (MLO) view and the cranio-caudal (CC) view. The MLO view is usually taken under a 45 deg angle. Most lesions are visible in both views. However, sometimes lesions are obscured by dense tissue in one of the views, or they may be projected outside the detector area. The MLO projection covers the largest tissue area and shows part of the pectoral muscle. When radiologists read mammograms they judge the different views in combination. In addition to comparison of MLO and CC projections they also make left/right comparisons and if previous mammograms are available they judge temporal changes. Computer aided detection methods, on the other hand, are mostly limited to analysis of one projection.

By processing MLO and CC views independently, CAD systems often mark abnormalities only in one view. Radiologists find this inconsistent, in particular if lesions appear rather similar in both views. This may reduce their confidence in the system and may affect their ability to use CAD in a beneficial way. This limitation of CAD has been identified as an important issue in recent experimental studies, which reported that it is more likely that radiologists ignore CAD results if these only mark a lesion in one view.^{1,2} The fact that CAD marks a lesion only in one view may be easily explained by differences in features computed for the regions in the two views. These will generally lead to differences in the levels of suspiciousness computed by the CAD system. During display, a threshold is applied to the level of suspiciousness to limit the number of regions rendered on the display. Due to this, a CAD marker may be visible in one

view while in the other it is just not displayed. To overcome this limitation CAD systems should be designed that make the display of marks dependent on both views.

It is interesting to note that computer aided detection methods in mammography are often evaluated using a case based sensitivity measure. This means that a true positive is counted if a lesion is marked in either one of the two views or in both. This should not be confused by a case based analysis of mammograms by the CAD system itself. In the light of the studies above more attention should be paid to lesion based analysis, or to measurement of the number of true positives marked in both views.

In the literature some approaches have been described to establish correspondence between MLO and CC views. Highnam *et al.*³ used a model-based method to find a curve in the MLO view that corresponds to the potential positions of a point in the CC view, while Good *et al.*⁴ reported a preliminary attempt of matching computer-detected objects in two views by exhaustive pairing of the detected objects and feature classification. However, remarkably few studies have been presented that use correspondence between CC and MLO views to improve detection results. Such a study was conducted by Paquerault *et al.*,⁵ who developed a two-view matching method in which a correspondence score is computed for each possible mass pair. By combining this correspondence score with their single-view detection score, mass detection results improved significantly.

In this study we focus on detection of masses, architectural distortion, and asymmetry, briefly referred to as masses in the rest of the paper. The work is based on a method to link potentially suspicious areas determined by our CAD scheme in MLO and CC views.⁶ In this paper we aim to use

this linking method to improve results of our single-view CAD scheme for mass detection. In particular, we investigate if a correspondence in CC and MLO views can be utilized to reduce false positives, based on the idea that false positives in different projections may be expected to be less correlated than true positives. By reclassification of CAD findings using two-view information we aim at decreasing the suspiciousness of false positives while maintaining the strength of the true positives. Moreover, by combining information from two views, the difference between the CAD output of true positive projections in two views may be reduced, which can improve consistency of the system.

II. METHODS

II.A. Initial region detection

The method for detection of suspicious regions in single mammographic views used in this work has been described in detail in previous publications. Just a brief description will be given here. An overview is shown in Fig. 1. Prior to the first detection stage, mammograms are segmented into three regions: breast tissue, pectoral muscle, and background. The image background is labeled by marking pixels with high exposure and low gradient values. This operation is followed by morphological transformations to remove labels and to fill small gaps. Subsequently, in the MLO views the pectoral muscle is segmented as a straight line using a method based on the Hough transform.⁷ After segmentation, locations in the tissue area are sampled at regular intervals, and at each location features are computed to determine the presence of a potentially suspicious pattern, such as densities or foci of radiating line patterns.^{8,9} A supervised neural network classifier, labeled as “A” in Fig. 1, is used to merge these features into a measure of suspiciousness. In this way a map of suspiciousness is obtained. We determine the maxima in this map to obtain a list of locations that may be of interest for further inspection. By applying a threshold to the suspiciousness at these locations, i.e., the height of the maxima, candidate locations are identified for further processing. As this threshold is fixed the number of initial locations varies from image to image. The detection sensitivity at this threshold cannot be improved by further processing. Therefore we use very low threshold, typically leading to around ten false positives per image on average.

In a second stage of processing the initial locations are inspected in more detail. The analysis starts with application of an algorithm for segmentation of locally dense regions,¹⁰ using the initial locations as seed points. This results in a set of candidate regions. Using the boundary of the candidate regions, a number of features are computed to represent relevant characteristics of the region surrounding the seed point. These features are used in a second supervised classifier, also implemented by a neural network. Region features that are used include region size, contrast, texture, compactness, acutance measures, and relative location in the breast.¹¹ The output of the single-view region classifier “B” is a measure of suspiciousness of the region.

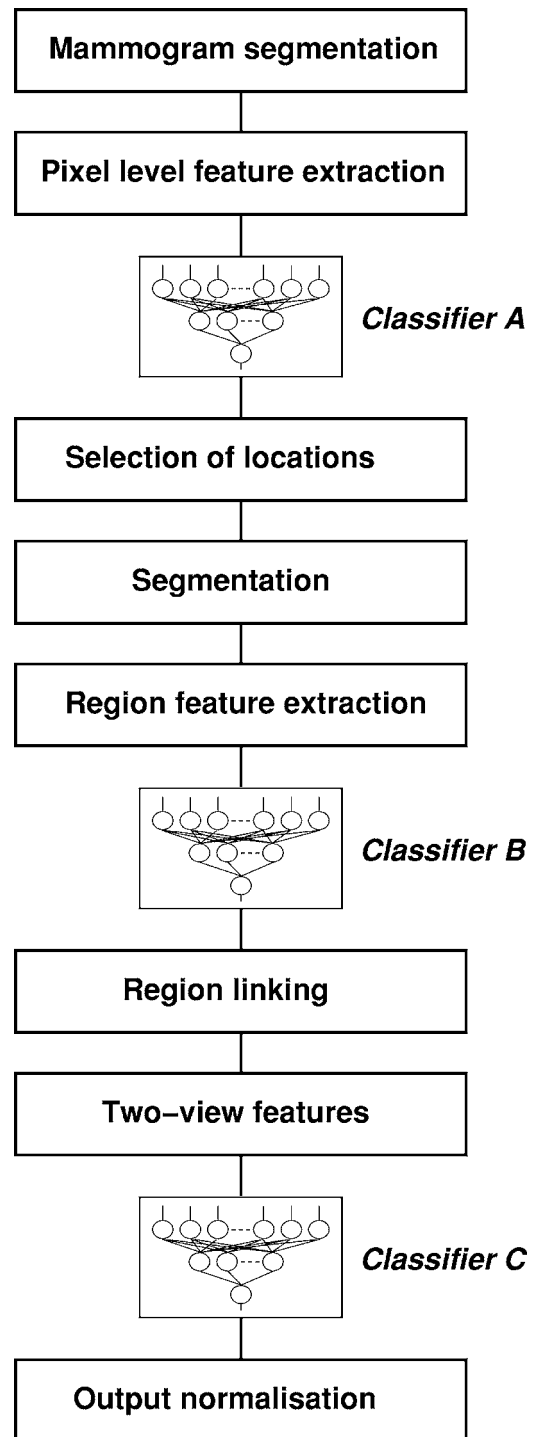


FIG. 1. Overview of the two-view detection method.

II.B. Classifier output conversion

The classifier used in the second processing stage is a multilayer perceptron with one hidden layer and one output node. The value of the output node is real number in the interval $[0,1]$. In the training phase the back-propagation rule is used to learn the network to map abnormal patterns to a value close to one and normal patterns to a value close to zero. After training the output can be used as a predictor of malignancy of new patterns. A higher output value means

that the input pattern is more suspicious. However, a certain output value cannot be translated to a well-defined performance level, which is sometimes inconvenient. For instance, if a CAD system is used in practice it is common to display only those regions for which the classifier output exceeds a certain threshold. For that setting one would like to know performance indicators of the system, in particular the average number of false positive marks per image. This is relatively easy to determine by applying the detection method to a series of representative normal cases. Of course, it is also important to know the sensitivity at the selected operating point. However, this cannot be measured easily in a standardized way, as enough representative samples should be available, along with reliable ground truth of lesions and detection criteria.

To relate the output of the classifier to performance of the system, we convert the neural network output to a measure L_S , which indicates innocence or normality of the region. This is done by applying the detection method to a series of normal mammograms and determining the average number of regions per image that have equal or higher scores than the region itself. In other words, the normality of a region is expressed by the average number of false positives per image at the least sensitive setting of the system at which the region is still marked. For the purpose of this study this conversion has the advantage that we can directly compare the output of different classifiers. For instance, we can study how the normality of normal and abnormal regions changes if we add stages to the CAD scheme, like the view correlation stage presented in this paper. It also has advantages when a detection scheme is evaluated using cross-validation, which involves training and testing of classifiers using different subsets. Normally the output of the neural networks trained with the various subsets should not be directly compared, as the output is influenced by the prevalence and difficulty of the normal and abnormal training patterns. To compute FROC performance for the whole dataset one can average FROC curves obtained for the subsets, but this is not ideal. When the classifier outputs are converted as we described, all data can be pooled for FROC analysis. It is noted that the conversion always is a monotone transform. As a consequence, FROC performance in a single dataset is not affected by the conversion.

II.C. Combined detection in MLO and CC views

In a previous study we investigated methods for determination of correspondence between regions in MLO and CC projections.⁶ Two of these methods were used in this study and are briefly discussed below. The common element of the two methods is a correspondence score, which is computed for each acceptable combination of a region with a region in the ipsilateral view. To determine if a region combination is acceptable, the distance to the nipple is used. An estimate of the location of the nipple is derived from the shape of the breast outline and the pectoral muscle. This procedure is fully automated. Experimentally it was found that in a large series of MLO-CC pairs, difference in distance to the nipple

was always smaller than 2.4 cm for corresponding regions. Therefore, we used this threshold value to determine if a region pair is acceptable. The correspondence score of a region pair is based on a set of features representing differences in appearance of the CC region and the MLO region, and on the nipple distance difference. Features include contrast difference and correlation. Using a database of regions with known correspondence, linear discriminant analysis is applied to learn how to discriminate between correct and incorrect region links in the feature space. This LDA scheme is subsequently used to predict if a new pair of regions corresponds. The output of the LDA scheme is the correspondence score.

The correspondence scores can be used to link regions in several ways. In this study we use two methods. The first is a method that links each region to the most likely region in the ipsilateral view, i.e., the one that has the highest correspondence score. The second is a one-to-one mapping obtained by sequentially assigning links between pairs with the highest correspondence score. In our previous work we found that the first method resulted in the highest fraction of correct true positive (TP) links. However, it is noted that by assigning links to the most likely candidate region a non-injective mapping is obtained, which is generally non-invertible. Region links computed for the MLO view may be different from links established with the CC view regions as starting point. As a consequence, this mapping does not lead to a unique definition of region pairs, which has implications for the two-view detection method.

After linking, several types of regions can be distinguished: regions that are part of a pair with one-to-one correspondence, regions that have several regions in the ipsilateral view mapping to them, and regions that are not mapped to any region. To avoid complications related to this, we did not design a method for classification of region pairs. Instead, we continue classification of individual regions, similar to the second stage of the detection method. By adding features extracted using information from the ipsilateral view we aim at improving the estimate of suspiciousness of candidate regions. Apart from being less complex, this approach has the advantage that the evaluation methodology used for single-view detection, lesion and case based FROC analysis, can still be used, which makes it easier to determine the benefit of two-view combination.

An outline of the two-view detection scheme is shown in Fig. 1. After application of the linking algorithm, most detections are linked to a region in the ipsilateral view. Not all detections are linked because it may happen that no acceptable region is available in the ipsi-latera view. To assess suspiciousness of a region in the presence of links a third stage is added to the detection scheme in which features from the two regions are combined in a supervised classifier "C." As input for the two-view classifier we investigated the use of single-view features, from the first and second stages of our scheme, and two-view features. These features are described in the following section.

TABLE I. Features investigated for the two-view classifier. The first column represents features used in the original single-view CAD scheme. In the second column the two-view features are shown. The correspondence score is the output of the LDA classifier that was used to discriminate between correct and incorrect region combinations.

Single-view features	Two-view features
spiculation features (f_1, f_2)	correspondence
focal mass features (g_1, g_2)	difference in distance to nipple (distance)
mass likelihood (L_A)	gray_scale_correlation
contrast measures ($\text{contrast}_1, \text{contrast}_2$)	polar_correlation
region_size	histogram_correlation
normality single-view (L_S)	normality of linked region ($L_{S_other_view}$)

II.D. Features for two-view classifier

As input to the two-view classifier a pool of 15 features was investigated, which are listed in Table I. In the following paragraphs we will give a short description of the used single-view features, which originate from the first and second steps of our CAD scheme. A detailed description of the two-view features can be found in a previous publication.⁶

- (i) **Spiculation features:** With respect to the detection of mass lesions in mammograms there are two important lesion characteristics, one is the presence of spicules and the other is the presence of a central mass. Our spiculation features are based on the idea that stellate lesions show a pattern of lines directed towards the center of a lesion. In the first step of our CAD scheme we use two spiculation features, which are both based on statistical analysis of local orientation patterns. Orientations are derived by second order Gaussian derivative filters. The first feature is a normalized measure of line concentration. The second feature calculates to what extent the locations with a line orientation towards the center are equally distributed in all directions. We will refer to these features as f_1 , respectively f_2 . Details can be found in Ref. 8.
- (ii) **Focal mass features:** For determining presence of a focal mass we use a similar approach as for the detection of spicules. The main difference is that instead of using line orientations we now calculate gradient orientations, using first order Gaussian derivatives. If a mass is present, the majority of image locations inside the mass will have gradient orientations directed towards the center of the mass. We derive two features from the calculated gradient orientations. The first is a normalized measure of gradient concentration, denoted by g_1 . The second feature g_2 represents whether or not the pattern is isotropic. Details can be found in Ref. 9.
- (iii) **Mass likelihood:** In the first step of our CAD scheme, pixel level features including the spiculation and focal mass features mentioned above are used as input to neural network classifier (labeled “A” in Fig. 1). The

output of this classifier is a continuous measure L_A representing the likelihood that a mass is present.

- (iv) **Region size:** In the second step of our CAD scheme regions are segmented at locations suspicious for presence of a mass. The size of a region, measured by the number of pixels of its segmented area, was used as feature (region_size).
- (v) **Contrast measures:** For the two-view classifier two contrast measures that performed best in the single-view analysis were included. The first is the difference in mean pixel value (which is linear with the optical density) between pixels inside [$E(I)$] and outside [$E(O)$] the segmentation,

$$\text{contrast}_1 = E(I) - E(O). \quad (1)$$

For the area outside the segmentation we took all pixels with a distance less than $0.6R$ from the contour, where R is the effective radius ($R = \sqrt{\text{area of the inside region}/\pi}$) of the region. The second contrast feature is the square of the difference between the mean pixel values inside and outside the contour, divided by the sum of the standard deviations of both areas,

$$\text{contrast}_2 = \frac{[E(I) - E(O)]^2}{\sigma(I) + \sigma(O)}. \quad (2)$$

- (vi) **Single-view likelihood of malignancy:** Next to the classifier output of the first detection stage, L_A , also the final output of our single-view CAD scheme, L_S , was investigated as input to the two-view classifier.

II.E. Two-view classifier

Based on correspondence scores, regions in the MLO and CC views were linked and two-view feature vectors for the linked region were composed. To select features for the two-view classifier, a forward selection procedure was used in combination with receiver operating characteristic (ROC) analysis. For this purpose, a LDA classifier was used. The optimal set of features was determined by adding features until the area under the ROC curve (A_z) did not further increase. To avoid bias in the computation of the A_z values a 50% cross-validation scheme was used at every feature selection step. The database used for feature selection included 948 cases and is described in detail the following section. The same database was used for evaluation of the two-view detection scheme. Because of the small number of features we used and the large number of regions (more than 18 000) we assumed that the risk of inducing a positive bias by this feature selection process was minimal.

For the two-view classifier we used a three-layer neural network with three hidden nodes. The network was trained with back-propagation. The input of this classifier was composed of the features selected by the ROC analysis. Just as for the single-view scheme, the output of this classifier is a continuous measure of the likelihood of malignancy of a region. Because the output has a variable scale, which de-

depends on the training schedule and the distribution of cases in the training sample, the output was converted to a normality level L_T in a similar way as the single-view classifier output conversion explained above.

It sometimes occurred that no link could be established for a region, because of limited availability of candidate regions in the other view. In principle, we could use the single-view output of classifier "B" as the final output of the detection scheme for these regions. However, this would not take into account that the absence of an existing link in the ipsilateral view is relevant information, which might be useful in the final estimate of suspiciousness of a lesion. Therefore, we decided to use another approach: When no link could be established the correspondence score and the other two-view features of a region were set to zero. This made it possible to process these regions also by the two-view classifier.

II.F. Evaluation

The detection schemes were evaluated using a data set containing 412 abnormal mammograms and 536 normal mammograms. All mammograms had four views: the CC and MLO projections of the left and right breasts. For some women several mammograms were included taken from different screenings or clinical exams. For the purpose of this study these were treated as separate cases. Of the abnormal mammograms 164 were prior screening mammograms of screen-detected or interval cancers. All abnormal cases had a visible mass, architectural distortion, or asymmetry in at least one view, which was verified by pathology reports to be malignant. Lesions contours were marked by, or under supervision of, an experienced screening radiologist. Normal cases were verified to be normal by two year follow-up. Cases with benign abnormalities were excluded. In total there were 824 annotated regions in the 412 positive cases in the dataset. In 388 cases a mass region was visible in both views and in 10 of these cases multiple mass regions were marked, indicating a multifocal or multicentric cancer. All mammograms were recorded with film-screen systems. For digitization two scanners were used, a Lumisys 85 (248 cases) and a Canon CFS300 (700 cases). Both scanners were operated at a pixel resolution of 50 μm and 12 bits/pixel. For processing, the images were averaged down to a resolution of 200 μm per pixel, maintaining the original gray value depth.

Detection performance was tested using FROC analysis and 20-fold cross-validation. Both a lesion and a case based evaluation were carried out. In the lesion based analysis sensitivity was computed as the number of lesions detected divided by the total number of lesions. In the case based evaluation, a case is by definition regarded as a true positive (TP) if a true positive detection occurs in at least one of the two views. A lesion was counted as detected if the center of mass of a region marked by CAD was located inside the annotated cancer outline. All regions that did not meet this criterium were counted as false positives. It is noted that conversion of the classifier output to a standardized level of normality was included in the cross-validation scheme. By doing this we could pool the results of the classifiers in the different sub-

sets in one analysis to obtain an overall FROC curve. For statistical analysis we determined the FROC performance in each of the cross-validation subsets. The area under the FROC in the range of false positives per image less than 1.0 was used as performance measure. Statistical significance of differences between FROC curves was determined by application of a paired Wilcoxon test over the subset results. It is noted that the performance measured in the subsets was not affected by the conversion of the classifier output.

In this study we made several comparisons. First, the performance of the two-view classifier was compared with the single-view results. Second, we investigated the effect of using the two region linking methods on detection performance. Finally, we repeated the comparison of single- and two-view detection on a subset of cases, in which all positive cases with an incorrect link of a true mass to a false positive were removed. This was done to gain insight in the effect incorrect links. To get a more detailed view of differences of the single- and two-view classifier we also performed a pairwise comparison of the results for individual regions, by plotting histograms of $L_T - L_S$ for true positives and false positives. It is noted that in the linking process not all regions initially detected by the single view CAD program were included, to reduce computation. In each view we only took the five most important regions, i.e., those with the lowest single-view normality scores.

III. RESULTS

The number of detected regions analyzed by the two-view classifier was 18 725, of which 885 were true positives. There were 79 annotated regions that were not detected by CAD after the five regions per view threshold was applied. Many true mass regions were detected more than once by CAD: 74 regions were hit twice and 27 regions more than twice. Linking with the most likely candidate region in the ipsilateral view resulted in a correct link for 79% of the true positives, while 18% of the true positives were linked to a false positive and 3% of the true positives was not linked. For the one-to-one linking scheme, the percentage of correct true positive links was 64%, while 33% of the true positives were linked to a false positive.

Using forward feature selection, we determined an optimal set of features for our two-view classifier. To give an impression of the importance of the features used in the two-view classifier, Table II presents the individual feature performances. The output of our single-view CAD scheme L_S was the first feature that was selected by the feature selection algorithm. The features that were selected next were all two-view features, namely correspondence, polar_correlation, L_S _other_view, distance, and histogram_correlation.

We compared the performance of the two-view classifier with our original CAD performance using FROC analysis. The result of the image based evaluation is presented in Fig. 2, and shows an improvement when using the two-view classifier for both linking methods. Best results were obtained with the maximum correspondence method. For instance, at a false positive rate of 0.1 FP/image, the lesion based sensi-

TABLE II. A_z values for the LDA classifier for the single- and two-view features. Using forward feature selection the following features were selected: L_S , correspondence, polar_correlation, L_S _other_view, distance, and histogram_correlation.

Features	A_z
Single-view features	
f_1	0.547
f_2	0.594
g_1	0.753
g_2	0.774
L_A	0.837
region_size	0.737
contrast ₁	0.795
contrast ₂	0.750
L_S	0.899
Two-view features	
L_S _other_view	0.765
distance	0.619
gray_scale_correlation	0.676
histogram_correlation	0.626
polar_correlation	0.742
correspondence	0.826

tivity increases from 52% to 61%. This improvement was statistically significant ($p=0.026$). Results obtained with the one-to-one linking method were worse, but the difference between the FROC curves of the two linking methods was not statistically significant. The differences are consistent with the fraction of correct TP-TP links obtained by the two methods, which were 79% and 64% for the maximum correspondence and one-to-one linking scheme, respectively. Case based FROC results were also computed for both linking methods. These results were not different from the single-view results. Lesion based performance improves because more lesions are detected in two views. This can be seen in Fig. 3 where the fraction of lesions detected in two views is shown as a function of the number of false positives per image.

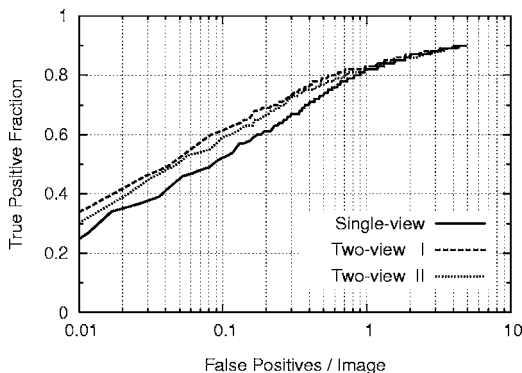


FIG. 2. Lesion based FROC results for the single- and two-view mass detection schemes. Two region linking methods were applied, based on maximum correspondence (I) and one-to-one linking (II). The following features were used as input for the two-view classifier: L_S , correspondence, polar_correlation, L_S _other_view, distance, and histogram_correlation.

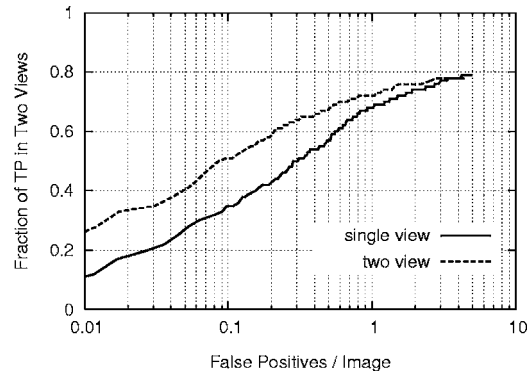


FIG. 3. Two-view detection FROC for linking method based on maximum correspondence. True positives are only counted if a lesion is detected in both views.

To have a closer look at the effect of the two-view classifier we made a pairwise comparison of the normality levels of the regions determined by the single- and two-view classifiers. Histograms of $L_T - L_S$ were computed for false positives and for the true positive CAD regions. Figure 4 shows the effect of the two-view classifier on the FP regions, where the FP regions are sorted into four groups based on the original (single-view) normality score. For the relatively suspect false positives (normality score < 1.0) the histograms show a shift to the right, which means that on average these false positives are rated more normal. This is consistent with an improvement of the FROC curve. However, it can also be seen that some of the less suspicious false positive regions (normality score > 1.0) are rated more suspicious by the two-view classifier. In the histogram for the TP regions (Fig. 5), a

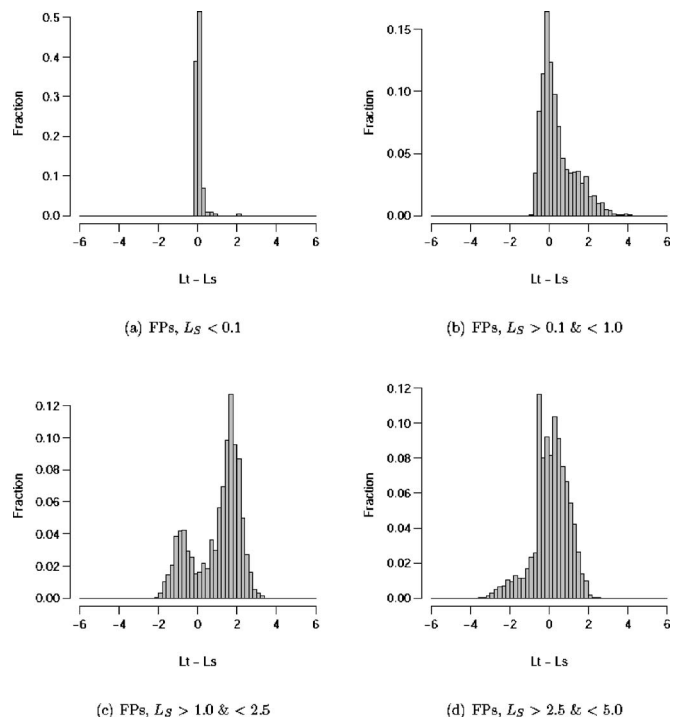


FIG. 4. Histograms of normality score changes $L_T - L_S$ observed after application of the two-view classifier, for different groups of false positives.

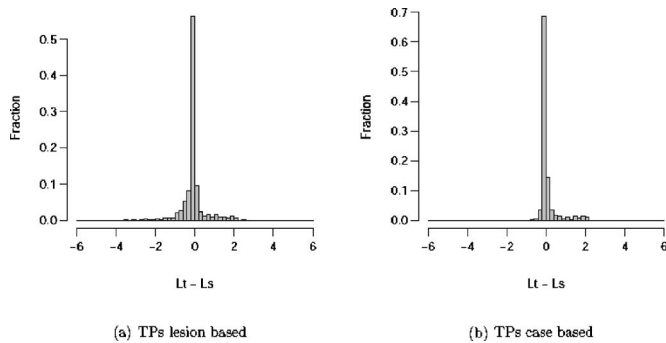


FIG. 5. Histograms of normality score differences $L_T - L_S$ before and after applying the two-view classifier for true positives. Differences computed by case are shown right.

slight shift of the normality scores to lower values can be seen, which means that the regions become somewhat more suspicious in the two-view detection scheme. However, if we look at the case based calculation, where for every case the minimum normality score of the TP regions in the MLO and CC view was used, we see the opposite effect.

To investigate the effect of incorrect links, we constructed a new case sample based on the linking results of the maximum correspondence method. All cases that had incorrect true positive links were removed. Figures 6 and 7 show FROC results for this set of cases. A strong increase of the number of lesions detected in two views can be observed, while also the case based detection performance improves slightly. However, case based improvement was not statistically significant.

IV. DISCUSSION AND CONCLUSIONS

Using lesion based FROC analysis we showed that detection results improve when using two-view information (see Fig. 2). However, in the case based evaluation we found no improvement. Obviously, case based performance is harder to improve, as sensitivity is completely determined by the views in which the masses have the highest malignancy rating (or the lowest normality score). It might be that most of the masses that were found more suspicious after application

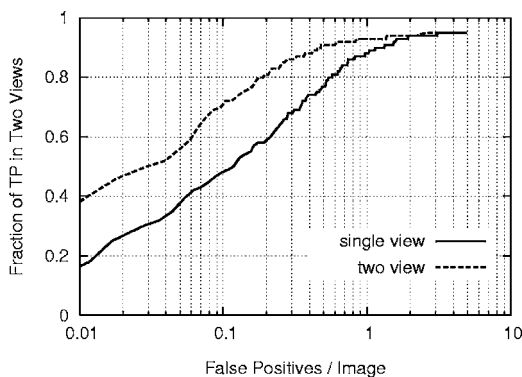


FIG. 6. Two-view detection FROC for a subset of the data in which positive cases with incorrect true positive links were removed. True positives are only counted if a lesion is detected in both views.

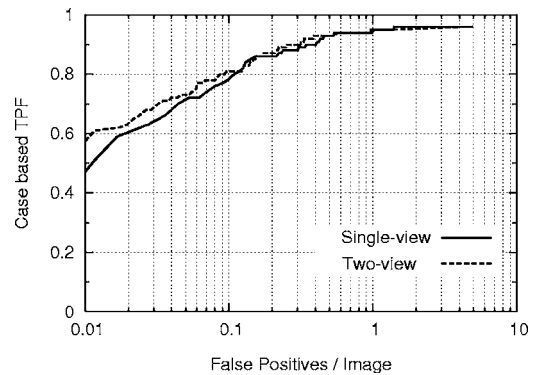


FIG. 7. Case based FROC results for a subset of the data in which positive cases with incorrect true positive links were removed.

of the two-view classifier were rated more suspicious in the other view. This might partly explain the fact that the case based performance did not change. As can be seen in Fig. 5, the positive effect of the two-view classifier on the malignancy score of the TP regions is not visible in the case based evaluation. On the contrary, we see a small negative effect. The reason that case based FROC performance did not get worse due to this effect lies in changes of the false positive ratings. Figure 4 shows that the effect of the two-view classifier on the FP regions can be quite strong. There is a net effect of rating them more normal. However, it seems that this benefit is canceled by the negative effect on the TP regions in the case based evaluation. This effect can be seen more clearly when calculating the sum of differences of the normality scores of the TP regions before and after application of the two-view classifier. In the image based evaluation this sum of differences is 38.4 against -38.9 in the case based evaluation.

For the two-view classifier it is important that the number of incorrect links between a TP region in one view and a FP region in the other view is as low as possible. Links between true and false positive regions may severely degrade detection performance, as this will generally lead to more suspicious ratings of the false positives and less suspicious ratings of the true positives. In Fig. 2 results of two different linking methods can be compared. As expected, best performance was achieved using the method with the highest percentage of correctly established TP-TP region links. The percentage of true positives with correct TP-TP links was 79% for this method. It seems that it is still worthwhile, though, to put more effort in improving the linking scheme, as the negative effect of incorrect links appears to be quite large. When cases with incorrect true positive links are removed the improvement obtained by two-view matching is larger (Fig. 7).

To our knowledge, the only study presented so far on the use of MLO and CC information to improve mass detection results is that of Paquerault *et al.*⁵ They developed a two-view matching method that results in a correspondence score for each possible mass pair. By combining this correspondence score with their single-view detection score, their classification results also improved significantly. Their lesion based detection sensitivity was found to improve from 62%

with a one-view detection scheme to 73% with their two-view scheme, at a false positive rate of 1 FP/image. The corresponding case based detection sensitivity improved from 77% to 91%. These results seem to be in contradiction with our results. Figure 2 shows that at the false positive level of 1 FP/image the sensitivity of our scheme is improved only very slightly by using two-view information (from 82% to 83%). This requires further investigation. In the paper by Paquerault *et al.* no information was given about the percentage of correctly established links.

Our detection results mainly improved for operating points with high specificity (low number of false positives per image) in the FROC curve. For instance, at a false positive rate of 0.1 FP/image, the lesion based sensitivity increased from 52% to 61%. While this reflects better performance of the system in distinguishing false positives from true masses, it may be less relevant when a CAD system is merely used to prompt regions at a higher false positive rate. However, we think that CAD systems in breast cancer screening will evolve to applications that display estimates of malignancy to the radiologists, in order to help them to make better decisions regarding referral. In those applications excellent performance of CAD is required at operating points lower than 0.1 FP/image.

In summary, we found that by establishing correspondence between regions detected in two views detection performance can be improved, but that improvements thus far are only seen in lesion based evaluation. This is important, though, as this means that results of the CAD system become more consistent: It happens less often that a lesion is only marked in one view. This may lead to increased confidence of radiologists in the system.

ACKNOWLEDGMENT

This work was funded by Grant No. NKB 2001-2380 of the Dutch Cancer Society.

⁴Electronic mail: n.karssemeijer@rad.umcn.nl

¹R. M. Nishikawa, A. Edwards, R. A. Schmidt, J. Papaioannou, and M. N. Linver, "Can radiologists recognize that a computer has identified cancers that they have overlooked?" *SPIE Medical Imaging*, Vol. 6146 (2006).

²B. Zheng, D. Chough, P. Ronald, C. Cohen, C. M. Hakim, G. Abrams, M. A. Ganott, L. Wallace, R. Shah, J. H. Sumkin, and D. Gur, "Actual versus intended use of CAD systems in the clinical environment," *SPIE Medical Imaging*, Vol. 6146 (2006).

³R. P. Highnam, Y. Kita, J. M. Brady, B. J. Shepstone, and R. E. English, "Determining correspondence between views," in *Digital Mammography*, edited by N. Karssemeijer, M. A. O. Thijssen, J. H. C. L. Hendriks, and L. J. T. O. van Erning (Kluwer, Dordrecht, 1998), pp. 111–118.

⁴W. F. Good, B. Zheng, Y.-H. Chang, X. H. Wang, G. Maitz, and D. Gur, "Multi-image cad employing features derived from ipsilateral mammographic views," *SPIE 1999 Image Processing*, Vol. 3661, pp. 474–485 (1999).

⁵S. Paquerault, N. Petrick, H. P. Chan, B. Sahiner, and M. A. Helvie, "Improvement of computerized mass detection on mammograms: fusion of two-view information," *Med. Phys.* **29**(2), 238–247 (2002).

⁶S. van Engeland and N. Karssemeijer, "Finding corresponding regions of interest in mediolateral oblique and craniocaudal mammographic views," *Med. Phys.* **33**(9), 3203–3212 (2006).

⁷N. Karssemeijer, "Automated classification of parenchymal patterns in mammograms," *Phys. Med. Biol.* **43**, 365–378 (1998).

⁸N. Karssemeijer and G. M. te Brake, "Detection of stellate distortions in mammograms," *IEEE Trans. Med. Imaging* **15**, 611–619 (1996).

⁹G. M. te Brake and N. Karssemeijer, "Single and multiscale detection of masses in digital mammograms," *IEEE Trans. Med. Imaging* **18**, 628–639 (1999).

¹⁰S. Timp and N. Karssemeijer, "A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography," *Med. Phys.* **31**(5), 958–971 (2004).

¹¹G. M. te Brake, N. Karssemeijer, and J. H. Hendriks, "An automatic method to discriminate malignant masses from normal tissue in digital mammograms," *Phys. Med. Biol.* **45**(10), 2843–2857 (2000).