Temporal Change Analysis for Characterisation of Mass Lesions in Mammography

Sheila Timp, Celia Varela and Nico Karssemeijer

***Abstract*—In this paper we present a fully automated computer aided diagnosis (CAD) programme to detect temporal changes in mammographic masses between two consecutive screening rounds. The goal of this work was to improve the characterisation of mass lesions by adding information about the tumour behaviour over time. Towards this goal we previously developed a regional registration technique that finds for each mass lesion on the current view a location on the prior view where the mass was most likely to develop. For the task of interval change analysis we designed two kinds of temporal features: difference features and similarity features. Difference features indicate the (relative) change in feature values determined on prior and current views. These features may be especially useful for lesions that are visible on both views. Similarity features measure whether two regions are comparable in appearance and may be useful for lesions that are visible on the prior view as well as for newly developing lesions. We evaluated the classification performance with and without the use of temporal features on a dataset consisting of 465 temporal mammogram pairs, 238 benign and 227 malignant. We used cross validation to partition the dataset into a training set and a test set. The training set was used to train a Support Vector Machine classifier and the test set to evaluate the classifier. The average A_z value (area under the receiver operating characteristic curve) for classifying each lesion was 0.74 without temporal features and 0.77 with the use of temporal features. The improvement obtained by adding temporal features was statistically significant ($P = 0.005$). In particular similarity features contributed to this improvement. Furthermore we found that the improvement was comparable for masses that were visible and for masses that were not visible on the prior view. These results show that the use of temporal features is an effective approach to improve the characterisation of masses.**

***Index Terms*—Temporal change analysis, interval changes, mass characterisation, CAD, mammography**

I. Introduction

An important task of radiologists in mammography is to discriminate between benign and malignant lesions. In clinical practice a radiologist carefully analyses all detected lesions and classifies each lesion as benign, probably benign, suspicious, or highly suggestive of malignancy. The BI-RADS reporting system provides criteria on which radiologists should make this classification [1], [2]. The subsequent management of lesions mainly depends on this classification. For probably benign findings short-interval follow-up is suggested. For suspicious abnormalities biopsy should be considered. A good decision increases the number of correctly detected malignancies and reduces unnecessary additional examinations.

Mass lesions have some characteristics that can be used to discriminate between benign and malignant lesions [3], [4]. A very important one is the margin type of a lesion. Most benign masses possess well-defined sharp borders, while malignant tumours often have ill-defined, micro-lobulated or spiculated borders. Especially a spiculation pattern is strongly associated with the presence of a malignant lesion. There are only a few benign lesions that also show spiculation. These include a postoperative scar, a radial scar, fat necrosis or any other process resulting in marked fibrosis. Another characteristic that may be helpful in discriminating between benign and malignant lesions is the shape of a lesion. The shape of benign lesions is often round and oval, compared to a more irregular shape of most malignant lesions. The last important difference between benign and malignant lesions is the *tumour behaviour* over time. Benign masses tend to change slowly and have a more or less similar appearance on two consecutive screening mammograms. Malignant masses on the other hand may change considerably and become more suspicious during time. This paper focuses on the design of features that capture temporal changes to improve the characterisation of mass lesions.

Some studies have been done to evaluate the effect of using temporal information on either the detection [5]–[7] or characterisation [8], [9] of mass lesions. The last two studies are observer studies that evaluate the effect of prior views on the ability of radiologists to discriminate between malignant and benign lesions. In the study from Varela et al. six radiologists participated and the performance of each radiologist improved when using prior mammograms. Hadjiiski et al. [9] performed a study with eight radiologist and two breast imaging fellows and also found a significant improvement when the radiologists used prior views.

To our knowledge only one study compared the performance of a CAD system with and without using prior views [10]. The dataset for that study consisted of mammograms from two consecutive screening rounds with a visible mass lesion on the current and prior view. A radiologist first identified the mass lesion on current and prior mammograms after which a CAD programme calculated single view and temporal features. On a dataset consisting of 140 temporal image pairs the A_z value significantly increased from 0.82 to 0.88 when temporal features were added.

In this paper we develop a CAD programme for temporal change analysis to improve the characterisation of breast masses. This programme combines single view and temporal features to determine a likelihood of malignancy for each mass lesion. Our proposed method has some important advantages.

N. Karssemeijer is with the Department of Radiology, Radboud University Hospital Nijmegen, Geert Grooteplein Zuid 18, 6525 GA Nijmegen, The Netherlands (corresponding author; email: n.karssemeijer@rad.umcn.nl)

First, our method is almost completely automatic. It only requires manual identification of the mass on the *current* view, after which a regional registration programme is applied to identify a location on the prior view that best corresponds with the current mass lesion. Existing methods require manual identification of the mass on both *prior* and *current* views. Second, our method is not only suited for masses that are visible on the prior view but also for masses that are new. This corresponds with normal screening situations where only some lesions are visible on the prior view. Third, besides using difference features we include temporal features that measure changes in appearance between a mass region on the current view and a similar region on the prior view. These features discriminate between benign lesions that stay more or less constant and malignant lesions that change between two consecutive screening rounds.

Radiologists can use this programme as an aid to characterise mass lesions. When a radiologist uses this method he or she should provide the coordinates of the lesion on the current view. The programme then automatically finds a corresponding location on the prior view and determines single view and temporal features to estimate the likelihood that the lesion is malignant. Studies in the literature suggest that a radiologist can use a *likelihood of malignancy* determined by a CAD system to improve interpretation of lesions [9], [11], [12].

We evaluate the performance of our method on a dataset consisting of 238 benign and 227 malignant temporal mammogram pairs. Furthermore we split the dataset into two subsets. The first subset consists of masses that are visible on the prior view and the second subset of masses that are not visible on the prior view. We study which features are useful for each subset and determine the classification performance for each subset.

The remainder of this paper is organised as follows. Section II explains the proposed CAD method for characterisation of mass lesions. Section III describes the experiments, including the dataset in Section III-A and the classification results in Section III-B and III-C. The last section contains a discussion and conclusion.

II. Single View and Temporal CAD Programme

Our CAD programme processes mammograms from consecutive screening rounds in which the most recent mammogram contains a visible lesion. Fig. 1 shows an example of a case that consists of three consecutive mammograms. This case contains two temporal mammogram pairs. In this example we see that priors are not always available for CC views. Our CAD programme therefore consists of two different parts: a single view part and a temporal part. The single view CAD programme is applied to all current images. In this programme several features are calculated to discriminate between benign and malignant lesions. Subsection II-B describes the single view part in detail. When prior views are available we also use the temporal CAD programme, see Subsection II-C. Before applying the CAD programme we first preprocess all prior and current images. This is described in the following section.

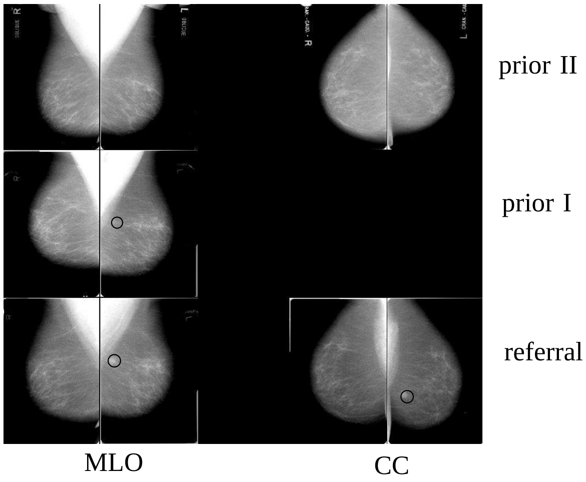


Fig. 1. Example of three consecutive mammograms of the same woman. Mammograms are displayed in chronological order. The bottom row represents the referral mammogram, the middle row the most recent prior mammogram (prior I), and the top row the second most recent prior mammogram (prior II). A malignant lesion is present in the left-MLO image of the referral. This lesion is also visible on the prior I left-MLO. This case provides two temporal mammogram pairs. The first temporal mammogram pair consists of a temporal image pair (left MLO referral–prior I) and a single image (left-CC referral). The second temporal mammogram pair consists of one temporal image pair: left MLO prior I–prior II.

A. Initial CAD programme

We apply the initial CAD programme to all prior and current images. We start with pre-processing each image: segmentation of the image into breast tissue, background tissue and pectoral muscle [13], peripheral enhancement to correct for differences in tissue thickness, and removal of the sharp transition in grey level from the breast area to the pectoral region [14]. We then apply a pixel level mass detection algorithm that estimates the potential presence of a tumour at each location inside the breast area. For this purpose we calculate at each location two features for the detection of a spiculation pattern or architectural distortion and two features for the detection of a focal mass. A neural network classifier combines these features into a single score that represents the likelihood that a mass is present at that location. We call this classifier output score the *mass likelihood*.

B. Single View CAD

After pre-processing the single view CAD programme processes all current images. First the mathematical centre of mass of the radiologists' annotation is determined for each mass lesion. A segmentation algorithm developed previously in our group uses this location as starting point to determine a contour for each lesion [15]. For each segmented lesion several single view features are determined that are useful for characterisation of mass lesions. A Support Vector Machine classifier combines these features into a single score that represents the probability that the lesion is malignant. The single view features we calculated are the following:

a) Spiculation Features: To determine the degree of spiculation we use an algorithm developed by Karssemeijer and te Brake [16]. This algorithms determines at each location two spiculation measures $f1$ and $f2$. To increase robustness we also add two features that calculate the average value of $f1$ and $f2$ inside the central part of the segmented lesion.

b) Border Features: We extensively studied the edges of circumscribed and ill-defined lesions and found that the perception of a well-defined margin very often coincides with the presence of a continuous boundary structure. On the contrary, ill-defined margins are often characterised by local randomness of the contour. Therefore we include a feature that represents the extent to which the margin is continuous or smooth. For details see [17].

c) Location Features: Malignant lesions have a preference to develop in the upper outer quadrant of the breast [18]. Inclusion of location features may thus be beneficial for the characterisation of masses. We determine for each lesion the location relative to the pectoral edge. For this purpose we define a new coordinate system which differs for medio lateral oblique (MLO) and cranio caudal (CC) views [19]. In MLO views the fitted pectoral edge serves as the y-axis, in CC views the chest wall boundary of the image. For both views the x-axis is the line perpendicular to the y-axis that has the longest distance from the y-axis to the breast boundary. The new coordinate system is used to define the relative x- and y-location of the centre of each segmented region. To compensate for different breast sizes these coordinates are normalised with the effective radius of the breast $r = \sqrt{A/\pi}$, where A is the size of the segmented breast region.

d) Morphological Features: We include two morphological features. The first one—size—is calculated as the area of the segmented region. Studies show that malignant masses on average are larger than benign ones [20]. The second morphological feature measures to what extent the segmented region is circularly shaped. We include this feature because benign masses often have a round or oval shape compared to a more irregular shape of malignant masses. We define circularity as

$$c = p^2/A,$$

where p is the perimeter and A the size of the region.

e) Detection of Micro Calcifications.: The presence of micro-calcifications at the location of a mass lesion is a sign of malignancy. Therefore we use a programme for the detection of micro-calcifications (ImageChecker, R2 Technology, Sunnyvale (CA)). As feature we use the number of calcifications found inside the segmented region.

C. Temporal CAD

The temporal CAD part consists of two steps. In the first step a regional registration technique determines for each lesion on the current view a location on the prior view where the mass most likely developed. In the second step the prior region is segmented and temporal features are calculated by combining information from prior and current regions.

1) Regional Registration: Our temporal CAD programme starts with a regional registration method that aims at finding for each lesion on the current view a corresponding location on the prior view. Below we shortly describe this method. For details see [14] and [20]. First we globally register prior and current views by applying a centre of mass alignment [21]. After alignment we use the centre coordinates of the current lesion (c_x, c_y) as midpoint of a circular search area on the prior view with radius 2 cm. Inside this search area we select the location where the mass most likely developed. To this end we calculate at each location (m, n) inside the search area three registration measures: mass likelihood, distance and grey scale correlation. The mass likelihood indicates the potential presence of a mass at each location, see Section II-A. The distance measure indicates the distance from (m, n) to the centre of the search area (c_x, c_y). The last measure is Pearson's correlation between the current region and a similar region on the prior view centred at (m, n). A linear discriminant analysis (LDA) classifier combines these measures—mass likelihood, distance and correlation—into a single score: the *registration score*. We then select the location with the highest registration score as match for the current mass lesion. When the selected location is inside the manual outline of the lesion we call the link correct. Fig. 2 shows some examples of temporal image pairs and the location selected by the regional registration programme.

2) Temporal Features: We use the location (m_s, n_s) on the prior view with the highest registration score as starting point for a segmentation algorithm [15]. This algorithm determines a contour of the region on the prior view, independent of whether the lesion is visible or not. We determine single view features for the segmented region on the prior view and then calculate two kinds of features that measure temporal changes: difference features and similarity features.

a) Difference Features: Difference features measure changes in feature values between the prior and the current region. In our experiments we determine difference features for all single view features except for the location features. For the feature 'size' we use the relative change between the feature value of the current region and the feature value of the prior region. For the other features we use the absolute change as these features are already normalised measures. Difference features may be especially helpful when the tumour is already visible on the prior view. When the lesion is not yet visible on the prior view the contour defined by our segmentation programme is not meaningful. Features that depend on the contour such as the size of a region will not be useful in that case.

b) Similarity Features: The second group of temporal features measure the similarity between the current region and a similar region on the prior view.

- Regional Registration Score. The first similarity feature is the output from the regional registration programme. This feature corresponds with the likelihood that a correct link has been established. A very low registration score may indicate that the lesion is not visible on the prior view. The classifier might use this information to determine the relative usefulness of temporal difference features. The registration score on its own may also help to characterise mass lesions. A very high registration score for example may indicate the presence of a benign mass when the mass is obvious on the prior view—resulting in a high mass likelihood—and similar on prior and current

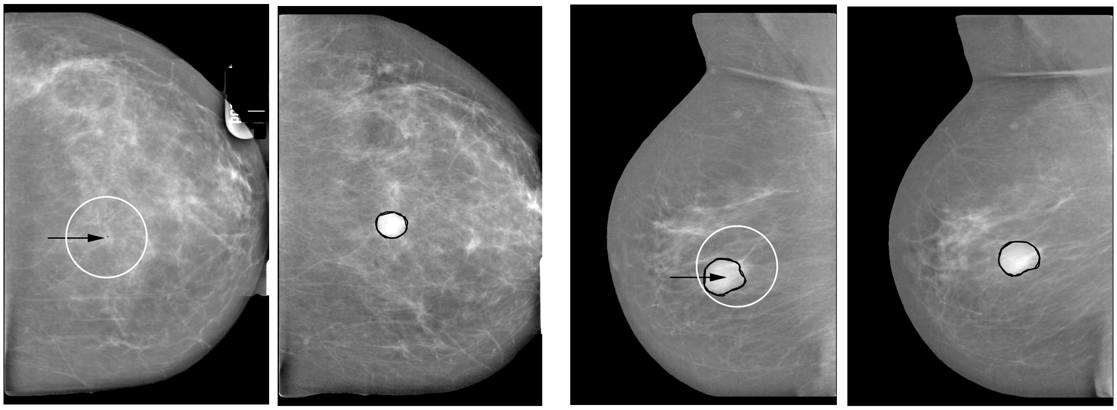


Fig. 2. Pairs of temporal images. Right and left images correspond to current and prior views. In each prior view the arrow indicates the location selected by the regional registration programme. Fig. 2(a) shows a malignant mass. The mass is not yet visible on the prior view. The registration programme selects the most probable location on the prior view. Fig. 2(b) shows a benign mass that is similar on prior and current views.

views—resulting in a high correlation measure. A low registration score on the other hand may suggest the presence of a malignant lesion as malignant lesions often change more between two consecutive screening rounds.

- Relative Grey Level Change. The second similarity feature calculates the relative difference in grey level between the current and prior region. For this purpose we transform the current image such that its grey level histogram matches that of the prior image. We first calculate for the prior and the current image the cumulative histograms of the grey values inside the breast area. For each grey level y the cumulative histograms are

$$f_C(y) = \sum_{i=0}^{y} H_C(i) \qquad f_P(y) = \sum_{i=0}^{y} H_P(i),$$

where H_C and H_P are the grey value histograms calculated inside the breast area on the current and the prior image respectively. We then transform each grey level y of the current image

$$\tilde{y} = f_P^{-1}(f_C(y)).$$

After histogram matching we determine the relative grey level change between a similar region on the prior and the current view. We use the segmented region on the current view as a template and put this template over the selected location (m_s, n_s) on the prior image. The relative grey level change between both regions is

$$RGLC = \frac{1}{N} \sum_{(m,n) \in C} (\tilde{y}_c(m,n) - y_p(m', n')),$$

where the summation is performed over all locations (m, n) inside the current region C. N denotes the number of pixels inside C, $\tilde{y}_c(m, n)$ denotes the transformed grey level at location (m, n) in C and $y_p(m', n')$ the grey level at the same relative location in the prior region with centre (m_s, n_s).

D. Case Based Classification

As classifier we use a Support Vector Machine [22] where we use the implementation provided in the freely available package from CRAN [23]. We use the radial basis kernel for training and testing. For testing we use the probability model for classification assuming equal priors. The probability model for classification fits a logistic distribution using maximum likelihood to the classifier outputs. The probabilistic regression model assumes (zero-mean) Laplace-distributed errors for the predictions, and estimates the scale parameter using maximum likelihood [23].

As not all images in our dataset have prior views (see Fig. 1) we train two different classifiers: a single view classifier and a temporal classifier. For both classifiers we apply a 20-fold cross validation scheme to partition the dataset into a training set and a test set. The *single view classifier* estimates for each image the posterior probability $p(m|\mathbf{x_{sv}})$ that the image contains a malignant lesion given the feature vector $\mathbf{x_{sv}}$ containing single view features extracted from the current region . The case based malignancy score $\zeta(l)$ combines the posterior probabilities from available MLO and CC views to estimate the likelihood that a lesion is malignant. When both MLO and CC views are present we use the sum rule to determine the case based malignancy score [24]:

$$\zeta(l) = \frac{1}{2}(p_{mlo}(m|\mathbf{x_{sv}}) + p_{cc}(m|\mathbf{x_{sv}})).$$

When only the MLO image is available the case based malignancy score is equal to the posterior probability from the MLO view.

$$\zeta(l) = p_{mlo}(m|\mathbf{x_{sv}}).$$

To include temporal information we train a second (temporal) classifier that determines the posterior probability $p(m|\mathbf{x_t})$ that a region is malignant given a temporal feature vector $\mathbf{x_t}$ containing single view, difference and/or similarity features. The case based malignancy score $\zeta(l)$ indicates the likelihood that the lesion is malignant and depends on the available views of the current and the prior mammogram. We distinguish the following situations.

- The temporal mammogram pair only contains a temporal MLO image pair. For the current mammogram no CC views are available. This situation corresponds with the second mammogram pair in Fig. 1. We use the posterior probability from the MLO view as the case based malignancy score:

$$\zeta(l) = p_{mlo}(m|\mathbf{x_t}).$$

- The temporal mammogram pair consists of a temporal MLO image pair and single view CC images. Prior CC views are not available. See first mammogram pair in Fig. 1. To determine the case based malignancy score we use the sum rule to combine the posterior probabilities from the temporal MLO classifier and the single view CC view classifier:

$$\zeta(l) = \frac{1}{2}(p_{mlo}(m|\mathbf{x_t}) + p_{cc}(m|\mathbf{x_{sv}})).$$

- The temporal mammogram pair consists of a temporal CC image pair and a temporal MLO image pair. We use the sum rule to combine the posterior probabilities from the temporal classifier for both the CC and the MLO view:

$$\zeta(l) = \frac{1}{2}(p_{mlo}(m|\mathbf{x_t}) + p_{cc}(m|\mathbf{x_t})).$$

For the evaluation of the single view and the temporal CAD system we use Receiver Operating Characteristic (ROC) methodology [25], [26]. We quantify the classification accuracy as the area under the case based ROC curve (A_z value). To test whether temporal features improve the performance we perform a paired comparison of both conditions—CAD with and without the use of temporal features—with regard to differences in the area under the two estimated ROC curves. For this purpose we use the freely available CLABROC software [27].

III. Experiments

A. Dataset

The mammograms used in this study all come from the Dutch Breast Cancer Screening Programme. All women aged 50-75 are invited bi-annually to participate in this programme. Two mammographic views—medio lateral oblique (MLO) and cranio caudal (CC)—are obtained at the initial screening. At subsequent screenings only medio lateral views are obtained, unless there is an indication that additional cranio caudal views would be beneficial. All images were digitised with a Canon CFS300 laser scanner at a pixel resolution of 50 μm and averaged to a resolution of 200 μm maintaining the original grey value resolution of 12 bits. All visible masses were annotated by or under supervision of an expert radiologist.

For the experiments we used consecutive mammograms from a collection of cases that were referred between 1996 and 2000. These cases consist of mammograms at referral and mammograms from up to two previous screening rounds. All images from two consecutive screening rounds form a temporal mammogram pair. In a temporal mammogram pair we call the most recent mammogram the current mammogram and the mammogram from one screening round earlier the previous or prior mammogram. Fig. 1 shows an example of a case that contains two temporal mammogram pairs. The first mammogram pair consists of the referral mammogram and the mammogram from the screening round prior to referral. In this temporal pair we call the referral mammogram the current mammogram. This case contains a second mammogram pair because the mass lesion is also visible on the mammogram prior to referral. In this second mammogram pair the mammogram prior to referral represents the current mammogram and the mammogram obtained one screening round earlier the prior mammogram. This means that at the time the current mammogram was taken the woman was not referred for further examination. These mammogram pairs make up 35% of the total dataset and often contain subtle lesions that are difficult to characterise.

TABLE I
Information about the subsets. For each subset the number of mammogram (image) pairs is given.

Set Name	Number of Pairs	Benign Pairs	Malignant Pairs
total dataset	465 (720)	238 (356)	227 (364)
temporal dataset	465 (542)	238 (279)	227 (263)
single view dataset	178 (178)	77 (77)	101 (101)
subset with visible priors	202 (246)	108 (133)	94 (113)
subset with normal priors	263 (296)	130 (146)	133 (150)

We constructed the dataset for the experiments by collecting all temporal mammogram pairs in which the current MLO view contained exactly one visible mass lesion. This resulted in 465 temporal mammogram pairs, 238 benign and 227 malignant. These mammogram pairs contain 720 image pairs: 465 MLO image pairs, 77 CC image pairs, and 178 single CC view images.

The total dataset can be divided into a *temporal dataset* and a *single view dataset*. The *temporal dataset* contains 542 temporal image pairs, 465 MLO and 77 CC. The *single view dataset* consists of 178 single view CC images. We constructed two different subsets of the temporal dataset. The first subset consists of masses that are also visible on the prior view and is called the *subset with visible priors*. This set contains 202 mammogram pairs. The second subset consists of masses that were not visible on the prior view and is therefore called the *subset with normal priors*. This set contains 263 mammogram pairs. Table I summarises information about the subsets. We evaluated the benefit of using temporal features on the total dataset as well as on different subsets.

We determined the mass size on both current and prior views as the area inside the annotation. Figure 3 shows the distribution of the mass size for benign and malignant masses with visible priors. For benign masses the mean (median) mass size in cm^2 was 1.8 (1.0) on the prior and 2.0 (1.1) on the current view, for malignant masses the mean (median) size was 1.2 (0.9) on the prior and 2.4 (1.5) on the current view.

B. Single View Classification

Table II gives the performance of the individual features measured as the area under the ROC curve (A$_z$ value) for the total dataset consisting of 238 benign and 227 malignant mass lesions.

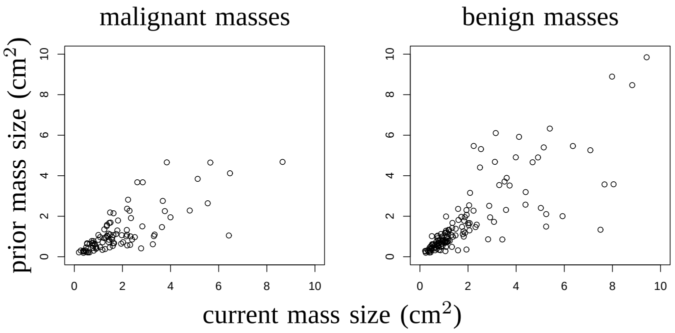


Fig. 3. Mass size on prior and current views.

TABLE II
SUMMARY OF THE SINGLE VIEW FEATURES. FOR EACH FEATURE WE CALCULATED THE INDIVIDUAL A_z VALUE FOR THE TOTAL DATASET CONSISTING OF 238 BENIGN AND 227 MALIGNANT MASS LESIONS.

Feature Name	Description	A_z
$f1$	spiculation	0.68
$f2$	spiculation	0.68
$\overline{f1}$	mean value of f1 inside region	0.68
$\overline{f2}$	mean value of f2 inside region	0.66
size	size of the segmented region	0.58
circularity	ratio between perimeter and size	0.52
calcification	number of calcifications	0.54
locx	relative x-location	0.56
locy	relative y-location	0.50
BC	continuity of the contour	0.62

For each region we constructed a single view feature vector that contained all single view features as described in Table II. Table III gives the performance obtained with this feature vector for the total dataset, the subset with visible priors and the subset with normal priors. This table shows that there is a large difference in classification performance between the subset with visible priors and the subset with normal priors.

TABLE III
A_z VALUE AND STANDARD DEVIATION FOR CASE BASED EVALUATION USING DIFFERENT FEATURE VECTORS. THE SINGLE VIEW FEATURE VECTOR CONSISTS OF FEATURES EXTRACTED FROM THE CURRENT MASS LESION. TEMPORAL FEATURES CONTAIN INFORMATION OF BOTH PRIOR AND CURRENT REGIONS.

Dataset	Single View FV	Temporal FV I	Temporal FV II	Temporal FV III
total dataset	0.74±0.02	0.78±0.02 †	0.74±0.02	0.77±0.02 †
visible priors	0.79±0.03	0.83±0.03 †	0.81±0.03	0.83±0.03
normal priors	0.72±0.03	0.75±0.03	0.70±0.03	0.73±0.03

† Statistically significant.

For the former the average A_z value is 0.79, for the latter 0.72. This may be explained by the observation that masses in the set with visible priors are often quite obvious on the current view. This often results in more distinct tumour characteristics making it easier to characterise these lesions. The set with normal priors on the other hand consists of masses that are only visible on the current view. This set therefore also contains very subtle masses which are harder to classify.

TABLE IV
SUMMARY OF THE TEMPORAL FEATURES. FOR EACH FEATURE WE CALCULATED THE INDIVIDUAL A_z VALUE.

Feature Name	Description	A_z
Similarity Features		
registration prob	probability that match is correct	0.60
RGLC	relative grey level change	0.63
Difference Features		
size_diff	relative difference in size	0.61
$f1$_diff	absolute difference in $f1$	0.62
$f2$_diff	absolute difference in $f2$	0.62
BC_diff	difference in continuity of the border	0.61

C. Temporal Classification

1) Registration: The regional registration programme used a search area with a radius of 2 cm. The programme correctly linked 79% of all visible masses to their corresponding priors. For correctly linked regions the mean Euclidean distance between the registered locations and the manually identified locations was 1.2 mm ± 0.96 mm , range [0-8.6mm]. In 21% of the cases no correct link could be established. We used all lesions (thus also misregistered lesions) for the temporal analysis. For all masses, also the misregistered ones, temporal features were based on the locations determined by the algorithm.

2) Individual Temporal Features: For all temporal features we determined the individual performance as the area under the ROC curve (A$_z$ value) for the temporal dataset consisting of 238 benign and 227 malignant mass lesions. To limit the number of features we selected the four best performing difference features for further use. The best features were relative difference in size, difference in border continuity, and two features that represent the difference in spiculation. Table IV summarises the individual performance of the selected difference features and both similarity features on the temporal dataset.

3) Temporal Feature Vectors: For each region we constructed three different temporal feature vectors, see Table V. The first temporal feature vector contained single view and similarity features. The second temporal feature vector contained single view and difference features. The last temporal feature vector contained single view, similarity, and difference features.

Table III gives the classifier performance obtained with the different feature vectors for each dataset. We compared the performance obtained with the *single view feature vector* with the performance obtained with the different temporal feature vector. The improvement obtained with *temporal feature vector I* was statistically significant for the total dataset ($P = 0.005$, two-tailed) and for the subset with visible priors ($P = 0.02$, two-tailed). The improvement for the subset with normal priors however was not significant ($P = 0.11$, two-tailed). Fig. 4 shows ROC curves obtained with the *single*

TABLE V
SUMMARY OF THE DIFFERENT FEATURE VECTORS. THE FIRST SET ONLY CONTAINS SINGLE VIEW FEATURES EXTRACTED FROM CURRENT LESIONS. THE TEMPORAL FEATURE VECTORS CONTAIN SINGLE VIEW AND TEMPORAL FEATURES.

Name	**Description**	**Nr**
single view feature vector	single view features	10
temporal feature vector I	single view and similarity features	12
temporal feature vector II	single view and difference features	14
temporal feature vector III	single view, difference, and similarity features	16

view feature vector and *temporal feature vector I*. For *temporal feature vector II*—containing single view and difference features—the classification performance only improved for the set with visible priors. This improvement was not significant ($P = 0.22$, two-tailed). For the set with normal priors the performance even decreased, indicating that difference features may not be useful for lesions that are not visible on the prior view. The last temporal feature vector—*temporal feature vector III*—contained single view, difference, and similarity features. The use of this feature vector improved the classifier performance for the total dataset ($P = 0.05$, two-tailed) and the subset with visible priors ($P = 0.06$, two-tailed).

To estimate the usefulness of difference features we compared the performance obtained with *temporal feature vector I* with the performance obtained with *temporal feature vector III*. For the subset with visible priors both feature vectors resulted in equal performances, indicating that difference features did not have an additional effect when similarity features were already used. For the subset with normal priors the addition of difference features even lead to a decrease in performance. These results suggest that similarity features are preferred over a combination of similarity and difference features.

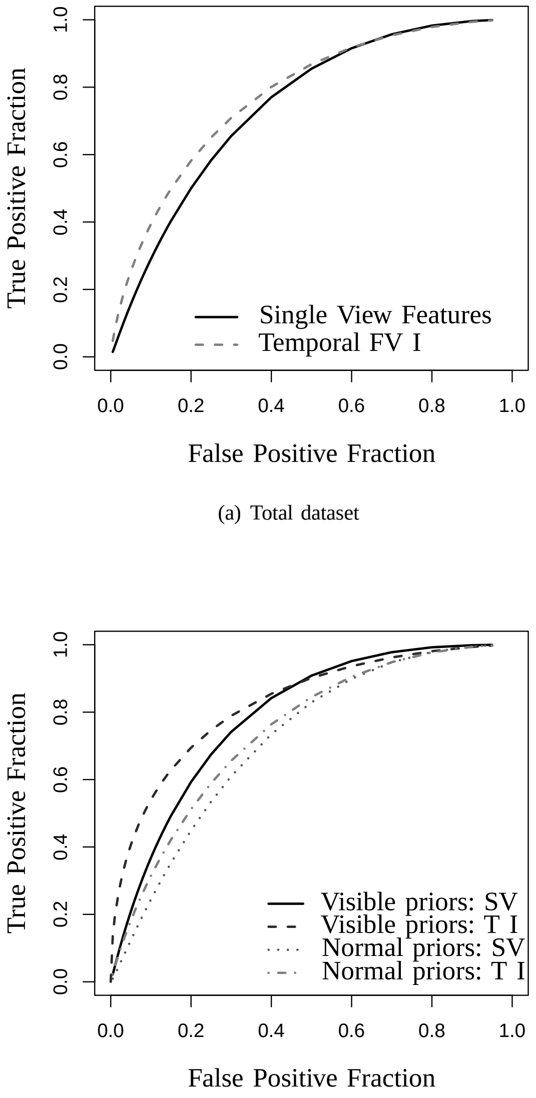
D. Comparison with Radiologists

In this section we will compare the performance of the CAD programme with the performance obtained by trained radiologists. For this purpose we use the results from an observer experiment Varela et al. [8] performed. In this experiment six radiologists rated a set of 99 benign and 99 malignant cases. These cases are a subset of the dataset used in the current study. On this subset the radiologists obtained an average A_z of .76 when prior mammograms were not used and an average A_z of .80 when priors were used. We compared this with the performance of our automated classification system on the same subset of cases. With the temporal CAD programme we obtained an A_z value of 0.81, which is slightly higher than the mean the A_z value of the radiologists. Also the improvement obtained by using temporal information on this subset was similar to the improvement the radiologists obtained with using priors: without temporal information the A_z value of CAD was 0.77.

IV. DISCUSSION

In this paper we present a CAD programme for characterisation of mass lesions. This programme uses single view and temporal information. The first step in the temporal CAD programme is regional registration to link each lesion on the current view to a corresponding location on the prior view. In total 79% of all lesions were linked correctly. This percentage is slightly lower than the percentage that we found in [20] (82%). This difference might be caused by the fact that the dataset we used for this study contains as much benign as malignant lesions. The dataset in the previous study contained proportionally more malignant lesions, which are more often correctly linked by the registration programme [20].

The performance obtained in a CAD study strongly depends on the used dataset. In our study we used a relatively difficult dataset that was sampled from cases referred in a screening programme with 40% positive predictive value (partly due to double reading). Obvious benign cases therefore were not referred. We believe that because of this low referral rate benign cases in our dataset are biased towards cases that show



(a) Total dataset

(b) Subsets visible and normal priors

Fig. 4. Case based ROC curves for the single view (SV) feature vector and the temporal feature vector I (T I).

temporal changes, as these cases look more suspicious and are thus more readily referred. Benign cases that remain constant between two screening rounds were often not referred. Therefore this dataset may have been more difficult to improve by adding features that capture temporal changes. In addition this dataset contained mammogram pairs in which the current mammogram was the mammogram prior to referral. These mammogram pairs make up 35% of the total dataset and often contain subtle lesions that are difficult to characterise. To demonstrate the difficulty of the dataset we compared the performance of the CAD system with the performance of radiologists on a subset of cases. On this subset radiologists obtained an average A_z of 0.76 without using priors and 0.80 when priors were used [8]. On the same subset our temporal CAD programme obtained a performance of 0.77 without temporal features and 0.81 with temporal features. The improvement obtained by the CAD programme is thus similar to the improvement obtained by radiologists.

Fig. 4 shows that the temporal CAD programme only improves the performance in the lower range of sensitivity (less than 0.8). At high sensitivity the ROC curves for the single view and temporal CAD programme overlap. Importance of the high sensitivity range of ROC curves is sometimes emphasized in classification studies, with reduction of unnecessary biopsies in mind as a potential clinical application. In this study however we are as well interested in improved sensitivity in screening using CAD. The study sample we used contained also lesions that were missed in screening by two readers, most likely because of misinterpretation. The obtained improvement at moderate sensitivity levels suggests that part of the cases missed during screening may be correctly referred when using the temporal CAD programme.

The temporal CAD programme determined two kinds of temporal features: difference and similarity features. The first similarity feature is the registration score. This feature combines three registration measures including the correlation between the current region and a similar region on the prior. Results show that a very low registration score is more often seen for malignant masses than for benign ones and vice versa. The second similarity feature is the relative grey level change (RGLC). An advantage of this feature is that it measures both changes in size and changes in contrast: the relative grey level of a lesion increases when a lesion becomes more dense and also when a lesion increases in size. Furthermore this feature works both for masses that are visible on the prior view and for masses that are new. Prior to calculating the RGLC we use histogram equalization to compensate for differences in exposure between two screening rounds. For this purpose we also tested two other grey scale registration methods in combination with our temporal features: 1) grey scale registration based on mutual information, and 2) a more sophisticated method developed by Snoeren [28] in which a physics based model is used to register the grey scale of temporal mammogram pairs. The use of these advanced grey scale registration methods led to an equal performance as the one obtained by using histogram equalization.

To detect temporal changes we also evaluated the performance of Pearson's correlation measure. In a study comparing twelve different similarity features Filev et al. found that this was the best performing one [29]. The individual A_z value for this feature however was rather low (0.54) compared to the A_z value of the RGLC (0.63). This difference is probably caused because the RGLC measures both changes in contrast and changes in size. The correlation feature on the other hand may fail to detect a change in density when the overall shape of a lesion remains the same.

The second group of temporal features we used were difference features. These features only improved the performance for lesions that were already visible on the prior view. On this subset the performance increased from 0.79 to 0.81. In a previous study Hadjiiski et al. [10] evaluated the effect of difference features on a set of masses with visible priors and found an improvement in A_z value from 0.82 to 0.88. The most obvious cause for the small improvement we obtained with difference features compared to [10] is the difficulty of our dataset. Other differences between both studies concern the used features and the registration method. Hadjiiski et al. used texture difference features, while in this study we used spiculation and morphological difference features. Texture features may be useful when a mass lesion is subtle on the prior view. The last difference concerns the registration method. In our experiments we used an automated registration programme while in [10] a radiologist indicated the location of the lesion on the prior view. To investigate whether this influenced the results we did an experiment in which we used the centre of the manual segmentation on the prior view instead of the location selected by the registration programme. In this experiment the classification performance increased to 0.82 for the set of visible masses. This result differs not much from the proposed CAD method indicating that we can use our automated registration programme instead of manual annotations.

It is remarkable that the classification performance for masses with normal priors decreased when using difference features. This decrease may be caused by the CAD programme not distinguishing between lesions with visible and lesions with normal priors. When the lesion is not visible it is not possible to determine an appropriate contour of the prior region and the segmentation programme will use accidental mammographic structures to determine a contour of the prior region. Consequently, features that depend on this contour will not be meaningful. The addition of these meaningless features may result in a lower classification performance.

Figs. 2 and 5 show some examples to illustrate potential benefits and drawbacks of the temporal CAD programme. For these examples we compared the malignancy score from the single view classifier with the score from the best temporal classifier using *temporal feature vector I*. Figs. 2(a) and 2(b) show two examples where the temporal classifier performed better than the single view classifier. Fig. 2(a) shows a malignant mass that is not visible on the prior view. The whole area on the prior view is 'empty' resulting in a low registration score and a RGLC. Fig. 2(b) concerns a benign mass that is almost identical in appearance on the prior and the current view. The use of temporal features therefore resulted in a better characterisation of the lesion. Fig. 5 shows an example where the single view classifier performed better than the temporal classifier. The malignant lesion is similar in appearance on both prior and current views. Consequently both temporal features were suggestive for a benign lesion, resulting in a lower malignancy score. The temporal classifier also performs worse than the single view classifier when a benign mass changes considerably in appearance. Temporal features are thus especially useful when they are combined with single view features. Temporal features can also be used in a detection programme; similarity features can then help to indicate the presence of a new lesion.

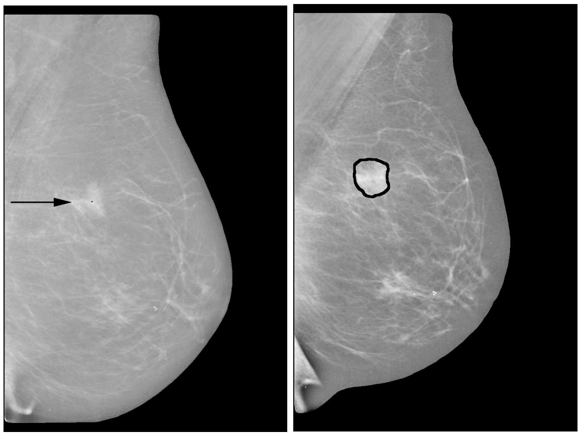


Fig. 5. Malignant mass that is similar on prior and current views.

V. Acknowledgement

This project was funded by a grant from the Dutch Cancer Society (KUN 2001-2380) and sponsored by fundraising activities of cycling club ”Bergh in het zadel”.

References

[1] C. D’Orsi and D. Kopans, “Mammography interpretation: the BI-RADS method,” *American Family Physician*, vol. 55, no. 5, pp. 1548–1552, 1997.

[2] S. Orel, N. Kay, C. Reynolds, and D. Sullivan, “BI-RADS categorization as a predictor of malignancy,” *Radiology*, vol. 211, no. 3, pp. 845–850, 1999.

[3] M. Friedrich and E. Sickles, Eds., *Radiological diagnosis of breast diseases*. Springer, 2000.

[4] M. Homer, *Mammographic interpretation*. The McGraw-Hill Companies, Inc., 1997.

[5] L. Bassett, B. Shayestehfar, and I. Hirbawi, “Obtaining previous mammograms for comparison: usefulness and costs,” *AJR*, vol. 163, no. 5, pp. 1083–1086, 1994.

[6] M. Thurfjell, B. Vitak, E. Azavedo, G. Svane, and E. Thurfjell, “Effect on sensitivity and specificity of mammography screening with or without comparison of old mammograms,” *Acta Radiol*, vol. 41, no. 1, pp. 52–56, 2000.

[7] M. Callaway, C. Boggis, S. Astley, and I. Hutt, “The influence of previous films on screening mammographic interpretation and detection of breast carcinoma,” *Clin Radiol*, vol. 52, no. 7, pp. 527–529, 1997.

[8] C. Varela, N. Karssemeijer, J. Hendriks, and R. Holland, “Use of prior mammograms in the classification of benign and malignant masses,” *Eur J Radiol*, vol. 56, no. 2, pp. 248–255, 2005.

[9] L. Hadjiiski, H.-P. Chan, B. Sahiner, M. Helvie, M. Roubidoux, C. Blane, C. Paramagul, N. Petrick, J. Bailey, K. Klein, M. Foster, S. Patterson, D. Adler, A. Nees, and J. Shen, “Improvement in radiologists’ characterization of malignant and benign breast masses on serial mammograms with computer-aided diagnosis: An ROC study,” *Radiology*, vol. 233, no. 1, pp. 255–265, 2004.

[10] L. Hadjiiski, B. Sahiner, H.-P. Chan, N. Petrick, M. Helvie, and M. Gurcan, “Analysis of temporal changes of mammographic features: computer-aided classification of malignant and benign masses,” *Med Phys*, vol. 28, no. 11, pp. 2309–2317, 2001.

[11] Z. Huo, M. Giger, C. Vyborny, and C. Metz, “Breast cancer: effectiveness of computer-aided diagnosis observer study with independent database of mammograms,” *Radiology*, vol. 224, no. 2, pp. 560–568, 2002.

[12] H.-P. Chan, B. Sahiner, M. Helvie, N. Petrick, M. Roubidoux, T. Wilson, D. Adler, C. Paramagul, J. Newman, and S. Sanjay-Gopal, “Improvement of radiologists’s characterization of mammographic masses by using computer-aided diagnosis: An ROC study,” *Radiology*, vol. 212, no. 3, pp. 817–827, 1999.

[13] N. Karssemeijer, “Automated classification of parenchymal patterns in mammograms,” *Phys Med Biol*, vol. 43, no. 2, pp. 365–378, 1998.

[14] S. Timp and N. Karssemeijer, “Interval change analysis to improve computer aided detection in mammography,” *Med Image Anal*, vol. 10, no. 1, pp. 82–95, 2006.

[15] ——, “A new 2D segmentation method based on dynamic programming applied to computer aided detection in mammography,” *Med Phys*, vol. 31, no. 5, pp. 958–971, 2004.

[16] N. Karssemeijer and G. te Brake, “Detection of stellate distortions in mammograms,” *IEEE Trans Med Imaging*, vol. 15, pp. 611–619, 10 1996.

[17] C. Varela, S. Timp, and N. Karssemeijer, “Use of border information in the classification of mammographic masses,” *Phys Med Biol*, vol. 51, no. 2, pp. 425–441, 2006.

[18] S. Caulkin, S. Astley, J. Asquith, and C. Boggis, “Sites of occurrence of malignancies in mammograms,” in *4th International Workshop on Digital Mammography, Nijmegen, the Netherlands*, N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, Eds. Kluwer, Dordrecht, 1998, pp. 279–282.

[19] F. Georgsson, “Differential analysis of bilateral mammograms,” *Int. Journal of Pattern Recognition and AI*, vol. 17, no. 7, pp. 1207–1226, 2003.

[20] S. Timp, S. van Engeland, and N. Karssemeijer, “A regional registration method to find corresponding mass lesions in temporal mammogram pairs.” *Med Phys*, vol. 32, no. 8, pp. 2629–2638, 2005.

[21] S. van Engeland, P. Snoeren, N. Karssemeijer, and J. Hendriks, “A comparison of methods for mammogram registration.” *IEEE Trans Med Imaging*, vol. 22, no. 11, pp. 1436–1444, 2003.

[22] N. Cristianini and J. Shawe-Taylor, *An introduction to Support Vector Machines and other kernel based learning methods*. Cambridge University Press, 2000.

[23] Hornik, “The R FAQ,” ISBN: 3-900051-08-9, 2005, available online: http://CRAN.R-project.org/doc/FAQ.

[24] J. Kittler, M. Hatef, R. Duin, and J. Matas, “On combining classifiers,” *Trans Pattern Anal Machine Intell*, vol. 20, no. 3, pp. 226–239, 1998.

[25] C. Metz, “ROC methodology in radiographic imaging,” *Invest Radiol*, vol. 21, no. 9, pp. 720–733, 1986.

[26] C. Metz, B. Herman, and J. Shen, “Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously-distributed data,” *Stat Med*, vol. 17, no. 9, pp. 1033–1053, 1998, available online: http://www-radiology.uchicago.edu/krl/rocstudy.htm.

[27] C. Metz, B. Herman, and C. Roe, “Statistical comparison of two ROC-curve estimates obtained from partially-paired datasets,” *Med Decis Making*, vol. 18, no. 1, pp. 110–121, Jan-Mar 1998, available online: http://www-radiology.uchicago.edu/krl/rocstudy.htm.

[28] P. R. Snoeren and N. Karssemeijer, “Gray scale registration of mammograms using a model of image acquisition,” *Inf Process Med Imaging*, vol. 18, pp. 401–412, Jul 2003.

[29] P. Filev, L. Hadjiiski, B. Sahiner, H.-P. Chan, and M. Helvie, “Comparison of similarity measures for the task of template matching of masses on serial mammograms,” *Med Phys*, vol. 32, no. 2, pp. 515–529, 2005.