# Comparison of similarity measures for the task of template matching of masses on serial mammograms

Peter Filev, Lubomir Hadjiiski,[a] Berkman Sahiner, Heang-Ping Chan, and Mark A. Helvie
*Department of Radiology, The University of Michigan, Ann Arbor, Michigan 48109-0904*

(Received 17 May 2004; revised 16 September 2004; accepted for publication 23 November 2004; published 27 January 2005)

We conducted a study to evaluate the effectiveness of twelve different similarity measures in matching the corresponding masses on temporal pairs of current and prior mammograms. To perform this comparison we implemented each of the twelve similarity measures in the final stage of our multistage registration technique for automated registration of breast lesions in serial mammograms. The multistage technique consists of three stages. In the first stage an initial fan-shape search region was estimated on the prior mammogram based on the geometrical position of the mass on the current mammogram. In the second stage, the location of the fan-shape region was refined by warping, based on an affine transformation and simplex optimization. A new refined search region was defined on the prior mammogram. In the third stage, a search for the best match between the lesion template from the current mammogram and a structure on the prior mammogram was carried out within the search region. Our data set consisted of 318 temporal pairs. We performed three experiments, using a different subset of the 318 temporal pairs for each experiment. In each experiment we further tested how the performance of the similarity measures varied as the size of the search region increased or decreased. We evaluated the twelve similarity measures based on four criteria. The first criterion was the mean Euclidean distance, which was the average distance of the true location of the mass to the location detected by the similarity measure. The second criterion was the percentage of temporal pairs that were aligned so that 50% or more of the lesion area overlapped. The third criterion was the percentage of pairs that were aligned so that 75% or more of the lesion area overlapped. The fourth and final criterion was the robustness of the similarity measure. Our results showed that three of the similarity measures, Pearson's correlation, the cosine coefficient, and Goodman and Kruskal's Gamma coefficient, provide significantly higher accuracy ($p<0.05$) in the task of matching the corresponding masses on serial mammograms than the other nine similarity measures. 
[DOI: 10.1118/1.1851892]

## INTRODUCTION

Mammography is currently the most effective method for detection of breast cancer. One of the important methods used by radiologists to detect developing malignancy in mammographic interpretation is the analysis of interval changes between serial mammograms. A variety of computer-aided diagnosis (CAD) techniques have been developed to detect mammographic abnormalities and to distinguish between malignant and benign lesions. We are studying the use of CAD techniques to assist radiologists in interval change analysis.

A few approaches to lesion registration between current and prior mammograms have been studied by investigators.[1–9] Sallam *et al.*[1] have proposed a warping technique for mammogram registration based on manually identified control points. A mapping function was calculated for mapping each point on the current mammogram to a point on the prior mammogram. Vujovic *et al.*[2] have proposed a multiple-control-point technique for mammogram registration. They first determined several control points independently on the current and prior mammograms based on the intersection points of prominent anatomical structures in the breast. A correspondence between these control points was established based on a search in a local neighborhood around the control point of interest.

The previous techniques depend on the identification of control points. However, because the breast is mainly composed of soft tissue that can change over time, there are no obvious invariant landmarks on mammograms. Furthermore, because of the elasticity of the breast tissue, there is large variability in the positioning and compression used in mammographic examination. As a result, the relative positions of the breast tissues projected onto a mammogram vary from one examination to the other. Techniques that depend on identification of control points will not be generally applicable to registration of breast images.

Gopal *et al.*[3] and Hadjiiski *et al.*[4,6] have developed a multistage technique that defines a transformation to locally map the position of the mass on a current mammogram to that on the prior mammogram. A local search for the mass is then performed on the prior mammogram. Good *et al.*[5] have also developed a technique that defines a transformation to map all points from the current mammogram onto a prior mammogram. The current mammogram is then subtracted from the prior mammogram. S. Van Engeland *et al.*[8] and Hadjiiski

*et al.*[9] used warping methods to align the current and prior mammograms. More detailed overview of the above methods can be found in the literature.[6,8]

In this study we focused on the multistage technique for automated registration of breast lesions in temporal pairs, developed by Hadjiiski *et al.*[6] In this method, initially, an automated procedure is used to detect the breast boundary on the current and prior mammograms. In the first stage of the process the location of the mass on the current mammogram is determined in a polar coordinate system with the nipple as the origin. By using the radial distance $R_{\text{curr}}$ between the nipple and mass centroid an arc is drawn which intersects the breast boundary. Angles are estimated at the radial distance $R_{\text{curr}}$ between the (nipple, mass centroid) and (nipple, intersections with the breast boundary) axis. The location of the current mass is determined by $R_{\text{curr}}$ and the obtained angles. Using the radial distance $R_{\text{curr}}$ to draw an arc centered at the nipple centroid on the prior mammogram, the two intersect points with the breast boundary on the prior mammogram are determined. Based on the angles obtained on the current mammogram and radial distance $R_{\text{curr}}$, the initial position of the lesion on the prior mammogram is estimated. An initial fan-shaped search region is then defined on the prior mammogram centered at the predicted location of the mass centroid. A fan-shaped template centered at the mass is also defined on the current mammogram. This fan-shaped region is then refined in the second stage by warping. The affine transformation in combination with simplex optimization was iteratively used to warp the fan-shaped template and further maximize the correlation measure with the breast structures on the prior mammogram. In the third stage the mass template from the current mammogram is matched to the corresponding lesion on the prior mammogram. The mass location on the prior mammogram is determined by maximizing the correlation similarity measure between the template and the structures within the search region.

In the current study we compared the effectiveness of correlation, as it is used in this technique as a similarity measure, to eleven other similarity measures. Our goal is to select the most effective similarity measures for locating the corresponding mass in the third stage of the automated registration technique.[6] Twelve similarity measures were compared in this study. The similarity measures included: Correlation, mutual information (scaled version), mutual information (unscaled version), increment sign correlation, gradient difference, pattern intensity, ordinal correlation, rank transform, cosine coefficient, Gamma coefficient, correlation standardized by the median, and the extended Jaccard measure. In addition to the accuracy of matching the masses using these similarity measures, we further tested their robustness by evaluating the dependence of the accuracy of matching on the size of the search regions.

## SIMILARITY MEASURES

In this section we will describe briefly the twelve similarity measures that were compared in this study. Figure 1 presents an example of a search region containing the mass in the prior year mammogram [Figs. 1(a) and 1(c)] and the current mass template [Figs. 1(b) and 1(d)] for two patients. In the following discussion, $I_{\text{current}}(i,j)$ represents the array containing the pixel values of the lesion template from the current mammogram and $I_{\text{prior}}(i,j)$ represents an array containing the pixel values of a sub-region within the search region on the prior mammogram, having the same dimensions as $I_{\text{current}}(i,j)$. The location of the sub-region is moved one pixel at a time over the entire search region on the prior mammogram, and at each location the $I_{\text{prior}}(i,j)$ array takes the pixel values of the current sub-region. The similarity between $I_{\text{current}}(i,j)$ and $I_{\text{prior}}(i,j)$ is calculated using one of the twelve similarity measures. This is basically a template matching operation in which the matching index is calculated using one of the twelve similarity measures. For a given similarity measure, the best match between the current mass template and a structure within the search region on the prior mammogram is found when the value of the similarity measure between $I_{\text{current}}(i,j)$ and $I_{\text{prior}}(i,j)$ is at a maximum. The structure is then considered to be the mass on the prior mammogram that corresponds to the mass of interest in the current mammogram. The mass found by a given similarity measure is compared to the ground truth, which was identified by an experienced radiologist based on available diagnostic and biopsy information, using the accuracy measures

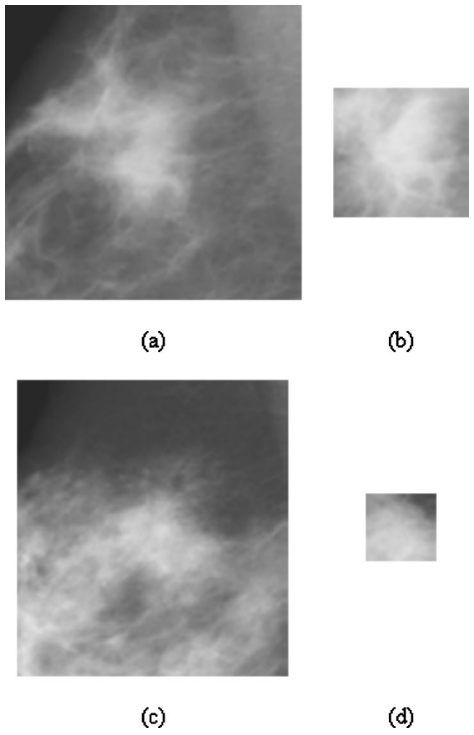


FIG. 1. Examples of templates containing the current masses and the corresponding search regions containing the prior masses for two patients: (a) The search region containing the mass in the prior mammogram (patient 1), (b) current mass template (patient 1), (c) the search region containing the mass in the prior mammogram (patient 2), (d) current mass template (patient 2).

described below. The average accuracy of a given similarity measure over the entire data set is then compared with those obtained with the other similarity measures.

### Correlation

The similarity measure that was originally used in the automated registration method[6] is Pearson's correlation coefficient

$$r = \frac{\Sigma_{i,j}(I_{\text{current}}(i,j) - \overline{I}_{\text{current}})(I_{\text{prior}}(i,j) - \overline{I}_{\text{prior}})}{\sqrt{\Sigma_{i,j}(I_{\text{current}}(i,j) - \overline{I}_{\text{current}})^2}\sqrt{\Sigma_{i,j}(I_{\text{prior}}(i,j) - \overline{I}_{\text{prior}})^2}}, \quad (1)$$

where $\overline{I}_{\text{current}}$ and $\overline{I}_{\text{prior}}$ are the mean pixel values of the mass template and the sub-region being evaluated on the prior mammogram, respectively.

### Mutual information

This similarity measure is widely considered to be highly effective for multimodal image registration.[10–12] It is a derivation from the information measure. The goal here is to maximize the information redundancy between the pixel intensity values contained in $I_{\text{current}}(i,j)$ and $I_{\text{prior}}(i,j)$. The definition for the mutual information is given as

$$S = \sum_{I_{\text{current}}(i,j),I_{\text{prior}}(i,j)} p(I_{\text{current}}(i,j),I_{\text{prior}}(i,j)) \times \log \frac{p(I_{\text{current}}(i,j),I_{\text{prior}}(i,j))}{p(I_{\text{current}}(i,j))p(I_{\text{prior}}(i,j))}, \quad (2)$$

where $p$ denotes probability. In order to calculate the probabilities in (2) we constructed a joint histogram of intensities with the pixel values of $I_{\text{current}}(i,j)$ used as the indices of the $x$ axis of the histogram and the pixel values of $I_{\text{prior}}(i,j)$ used as the indices of the $y$ axis. We studied how the similarity measure performed when the original pixel values were used to construct the joint histogram (referred to as the unscaled version) as well as when the pixel values were linearly scaled between their minimum and maximum within the subregions being matched and then used to construct the joint histogram (the scaled version). We also varied the number of bins on the histogram. We have searched for the optimal number of bins independently for two different data subsets containing the small current templates (74 templates smaller or equal to 10 mm) and the large current templates (57 templates larger or equal to 20 mm). We found that the best results for both data subsets occurred when the histogram was set to 5 bins per axis for the scaled version and to 32 bins per axis for the unscaled version. These were also the optimal number of bins when the entire data set of 318 pairs was used for the optimization.

### Increment sign correlation

This similarity measure is one that was designed to be robust for brightness change and occlusion.[13] The formula for the increment sign correlation (ISC) coefficient is given by

$$b_{ij} = \begin{cases} 1 & (I_{\text{current}}(i,j+1) \geqslant I_{\text{current}}(i,j)) \\ 0 & (I_{\text{current}}(i,j+1) < I_{\text{current}}(i,j)) \end{cases}$$
$$b'_{ij} = \begin{cases} 1 & (I_{\text{prior}}(i,j+1) \geqslant I_{\text{prior}}(i,j)) \\ 0 & (I_{\text{prior}}(i,j+1) < I_{\text{prior}}(i,j)) \end{cases}, \quad (3)$$

$$r_{\text{ISC}} = \frac{1}{(N-1)(M-1)} \sum_{i,j} \{b_{ij}b'_{ij} + (1-b_{ij})(1-b'_{ij})\}, \quad (4)$$

where $N$ and $M$ are the horizontal and vertical size of the current template. The principle here is to map the change in brightness of the template and of the corresponding subregion in the search region. This is achieved by building the two arrays $b_{ij}$ and $b'_{ij}$ each consisting of zeroes and ones. A value of one is assigned to each pixel that is greater in value than the one preceding it, and a zero is assigned to each pixel that is smaller than the one preceding it. The coefficient $r_{\text{ISC}}$ measures the similarity between the two arrays $b_{ij}$ and $b'_{ij}$. If the corresponding values in the two arrays are the same, i.e., both one or both zero, then a value of one is accumulated in the sum. On the other hand if the corresponding values are different, a value of zero is accumulated in the sum. The sum is finally divided by the number of values in the array to yield a value between zero and one.

### Gradient difference

The gradient difference measure compares the gradients of the template and the search region at each corresponding pixel[14]

$$G = \sum_{i,j} \frac{A_v}{A_v + (I_{\text{diff}V}(i,j))^2} + \sum_{i,j} \frac{A_h}{A_h + (I_{\text{diff}H}(i,j))^2}, \quad (5)$$

$$I_{\text{diff}V}(i,j) = \frac{dI_{\text{prior}}}{di} - \frac{dI_{\text{current}}}{di}, \quad I_{\text{diff}H}(i,j) = \frac{dI_{\text{prior}}}{dj} - \frac{dI_{\text{current}}}{dj}, \quad (6)$$

where $A_v$ and $A_h$ are constants, which were selected to be the vertical and horizontal variance of the prior gradient image. This similarity measure is related to increment sign correlation, except that, instead of assigning only the discrete values of zero and one, the actual derivatives are estimated in the $i$ and $j$ direction at each pixel location for both arrays. Each derivative in the template is subtracted from the derivative that is in the corresponding location and direction in the search region. Thus two new arrays are created containing information on the differences in the gradients of the two images. The goal is to find the maximum value of the coefficient $G$, which corresponds to that the differences in the gradients in the corresponding directions are at a minimum when the template has been aligned with the matching location in the search region.

### Pattern intensity

Pattern intensity is a similarity measure which utilizes the differences between the corresponding pixel values of the template and the search region.[14] The coefficient for pattern intensity is given by

$$PINT = \sum_{i,j} \sum_{v=i-r/2}^{i+r/2} \sum_{w=j-r/2}^{j+r/2} \frac{\sigma^2}{\sigma^2 + (I_{\text{diff}}(i,j) - I_{\text{diff}}(v,w))^2}, \quad (7)$$

$$I_{\text{diff}} = I_{\text{prior}} - I_{\text{current}}. \quad (8)$$

Here we create a matrix, $I_{\text{diff}}(i,j)$, consisting of the differences in the pixel values of the template and a subregion within the search region. We then take a sliding frame of dimensions $r \times r$ and move it throughout $I_{\text{diff}}(i,j)$. Each pixel value within the sliding frame is subtracted from the pixel value in the center of the sliding frame. The squared values of these differences are then added up for each location of the sliding frame. The constant $\sigma$ is used to weigh the function and plays a role in filtering out the noise. For our experiment we evaluated different values of $\sigma$ and determined that $\sigma = 10$ yields the most favorable results. For coarser images the dimensions of this frame can be increased. For the mass images, we found that $r = 3$ for the sliding window, i.e., a $3 \times 3$ sliding frame, provided the best matching.

### Ordinal measure

This is a measure of the similarity between the rankings of the pixel values of the template image and the sub-region within the search region. We used an ordinal measure of association.[15] The first step is to copy all pixel values from both images into one-dimensional arrays, $I_{\text{prior}}$ and $I_{\text{current}}$. The next step is to set up the arrays $\pi_{\text{prior}}$ and $\pi_{\text{current}}$, where $\pi^i_{\text{current}}$ is the rank of the gray level value ($I^i_{\text{current}}$) of pixel $i$ among the $I_{\text{current}}$ data. Larger gray level value will result in a larger rank value for a given pixel. Similarly $\pi^i_{\text{prior}}$ is the rank of $I^i_{\text{prior}}$ among the $I_{\text{prior}}$ data. Next we construct the vector $s$ by:

$$s^i = \pi^k_{\text{prior}}, \quad \text{where } k = (\pi^{-1}_{\text{current}})^i. \quad (9)$$

$\pi^{-1}_{\text{current}}$ is defined as the inverse permutation of $\pi_{\text{current}}$.

$$\text{If } \pi^i_{\text{current}} = j, \quad \text{then } (\pi^{-1}_{\text{current}})^j = i. \quad (10)$$

This vector $s$ represents the ranking of $I_{\text{prior}}$ with respect to the ranking $I_{\text{current}}$. Under ideal situations when the rankings of the pixel values within both images is the same, in other words when $\pi_{\text{prior}} = \pi_{\text{current}}$, then the vector $s$ should equal $\langle 1,2,3,4,\ldots,n \rangle$. The next step is to define a vector, $d_m$, which functions as a distance measure between the actual value of the vector $s$ and its ideal value of $\langle 1,2,3,4,\ldots,n \rangle$

$$d^i_m = i - \sum_{j=1}^{i} J(s^j \leqslant i), \quad (11)$$

where $J(B)$ is indicator function of event $B$, i.e., $J(B) = 1$ when $B$ is true and $J(B) = 0$ when $B$ is false. The ordinal measure of association $\kappa(I_{\text{current}}, I_{\text{prior}})$ is now calculated by

$$\kappa(I_{\text{current}}, I_{\text{prior}}) = 1 - \frac{2 \max_{i=1}^{n} d^i_m}{[n/2]}. \quad (12)$$

### Rank transform

For the rank transform similarity measure[16] a window of dimension $a \times a$ pixels, where $a$ is an odd integer greater than one and smaller than the size of the template subregion, is moved over the template $I_{\text{current}}(i,j)$ and the corresponding location $I_{\text{prior}}(i,j)$ in the search region. At each position of this window the number of pixels residing within the window that are greater in brightness than the pixel in the center of the window are counted. This number is subtracted from the total number of pixels, $a^2$, within the moving window and is defined as the pixel's rank transformation. In this way the images $I_{\text{current}}(i,j)$ and $I_{\text{current}}(i,j)$ are rank transformed to produce the arrays $r_{\text{current}}(x,y)$ and $r_{\text{prior}}(x,y)$. These transformations are given by

$$r_{\text{current}}(x,y) = a^2 - \sum_{(i,j) \in W} U[I_{\text{current}}(x+i, y+j) - I_{\text{current}}(x,y)], \quad (13)$$

$$r_{\text{prior}}(x,y) = a^2 - \sum_{(i,j) \in W} U[I_{\text{prior}}(x+i, y+j) - I_{\text{prior}}(x,y)], \quad (14)$$

$$U[t] = \begin{cases} 1, & t \geqslant 0 \\ 0 & t < 0 \end{cases}, \quad (15)$$

where $U[t]$ is a unit step function and $(i,j) \in W$ is the neighborhood of the rank window. To find the best match between the current and the prior images we find where the sum of the absolute differences of the rank transforms between the corresponding pixels is a minimum

$$RANK = \sum_{(x,y)} |r_{\text{current}}(x,y) - r_{\text{prior}}(x,y)|. \quad (16)$$

The value of $a$ in this study was selected to be 3, similar to the selection for the sliding window in the pattern intensity measure.

### Cosine measure

For the cosine measure[17] we arrange the pixel values of both images into vectors. Then in order to find the best match between the two vectors we try to find where the value of the cosine of the angle between the vectors is at a maximum. The cosine is calculated by finding the dot product and dividing it by the norm of each vector

$$\text{Cos} = \frac{\Sigma_{i,j}(I_{\text{current}}(i,j))(I_{\text{prior}}(i,j))}{\sqrt{\Sigma_{i,j}(I_{\text{current}}(i,j))^2}\sqrt{\Sigma_{i,j}(I_{\text{prior}}(i,j))^2}}. \quad (17)$$

The cosine similarity measure is very closely related to Pearson's correlation coefficient discussed earlier. The most notable difference is that the mean here is not subtracted from each value in order to center both sets of data about zero.

### Goodman and Kruskal's gamma coefficient

The Gamma coefficient[18,19] belongs to the family of ordinal measures. The Gamma coefficient is given by

$$\gamma = \frac{P_c - P_d}{1 - P_t}, \tag{18}$$

$$P_c = 2^* \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \left( \sum_{i'>i} \sum_{j'>j} p_{i'j'} \right), \tag{19}$$

$$P_d = 2^* \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij} \left( \sum_{i'>i} \sum_{j'<j} p_{i'j'} \right), \tag{20}$$

$$P_t = \sum_{i=1}^{r} \left( \sum_{j=1}^{c} p_{ij} \right)^2 + \sum_{j=1}^{c} \left( \sum_{i=1}^{r} p_{ij} \right)^2 - \sum_{i=1}^{r} \sum_{j=1}^{c} p_{ij}^2, \tag{21}$$

where $p_{ij}$ represents the probability that a pixel with gray level value $i$ on the current image $[I_{\text{current}}(x,y)=i]$ will correspond to the pixel with gray level value $j$ on the prior image $[I_{\text{prior}}(x,y)=j]$. $r$ and $c$ are the total number of possible values for the pixel gray levels of $I_{\text{current}}(x,y)$ and $I_{\text{prior}}(x,y)$, respectively. $P_c$ represents the probability that the rank ordering of the pixel values of the two images agrees, $P_d$ represents the probability that the rank ordering disagrees, and $P_t$ represents the probability of ties. The main advantage of this similarity measure over the previously discussed ordinal measure is that here we account for the case of ties between the pixel values.

### Correlation standardized by the median

We defined a similarity measure, the correlation standardized by the median, which is a variation of Pearson's correlation coefficient discussed earlier. Here instead of subtracting the mean to center and standardize the two sets of data we subtract the median

$$r_{\text{med}} = \frac{\Sigma_{i,j}(I_{\text{current}}(i,j) - \hat{I}_{\text{current}})(I_{\text{prior}}(i,j) - \hat{I}_{\text{prior}})}{\sqrt{\Sigma_{i,j}(I_{\text{current}}(i,j) - \hat{I}_{\text{current}})^2}\sqrt{\Sigma_{i,j}(I_{\text{prior}}(i,j) - \hat{I}_{\text{prior}})^2}}. \tag{22}$$

The medians of the pixel values of the template and the subregion in a corresponding location in the search region are represented by $\hat{I}_{\text{current}}$ and $\hat{I}_{\text{prior}}$.

### Extended Jaccard similarity measure

This similarity measure[20] is related to the previously discussed cosine measure. It is given by

$$Jacc = \frac{\Sigma_{i,j}(I_{\text{current}}(i,j))(I_{\text{prior}}(i,j))}{\Sigma_{i,j}(I_{\text{current}}(i,j))^2 + \Sigma_{i,j}(I_{\text{prior}}(i,j))^2 - \Sigma_{i,j}(I_{\text{current}}(i,j))(I_{\text{prior}}(i,j))}. \tag{23}$$

Unlike the cosine measure, the extended Jaccard measure also takes into account the magnitudes of the two vectors when evaluating similarity, in addition to their directions.

## DATA SET

The twelve similarity measures were evaluated on a data set consisting of 318 temporal pairs. Each pair of mammograms contained two mammograms taken at different times of the same breast. The time interval between the two mammograms ranged from 3 to 48 months. Our data set contained 510 digitized mammograms from 120 patients. Thirty-five of the mammograms were digitized with a LUMISYS DIS-1000 laser scanner at a pixel resolution of 100 $\mu$m $\times$100 $\mu$m. The digitizer had a 4096 gray level resolution and an optical density (OD) range of 0–3.5. The pixel values were linearly proportional to the OD within the range of 0.1–2.8 OD units, with a slope of 0.001 OD/pixel value. The slope of the calibration curve decreased gradually outside this optical density range. The rest of the mammograms were digitized with a LUMISCAN 85 laser scanner at a pixel size of 50 $\mu$m $\times$50 $\mu$m and again 4096 gray levels. The pixel values were linearly proportional to the OD range of 0–4 OD units, and again with a slope of 0.001 OD/pixel value. Output from both digitizers was linearly converted so that large pixel values corresponded to a low optical density. Current and prior mammograms of the same patient were digitized with the same digitizer. Since the mammographic masses are relatively large objects that do not require high resolution, we evaluated the similarity measures at a pixel size of 800 $\mu$m $\times$800 $\mu$m to reduce the processing time and reduce the image noise. The images were averaged using a filter that has constant weights over the entire filter kernel, which is referred to as a box filter, and were then down-sampled to the final resolution. The images digitized with the LUMISCAN 85 digitizer were averaged with a $16\times16$ box filter and were then down-sampled by a factor of 16. The images digitized with the DIS-1000 digitizer were averaged with an $8\times8$ box filter and were then down-sampled by a factor of 8. All images thus had a pixel size of 800 $\mu$m $\times$800 $\mu$m.

Of the 120 cases, 119 contained biopsy-proven masses and one was determined to be benign after a two-year follow-up. The 510 mammograms contained different mammographic views and multiple years of the masses including the year when the biopsy was performed. 172 of the 318 temporal pairs were malignant and the remaining 146 were benign. A malignant temporal pair contains the mammo-

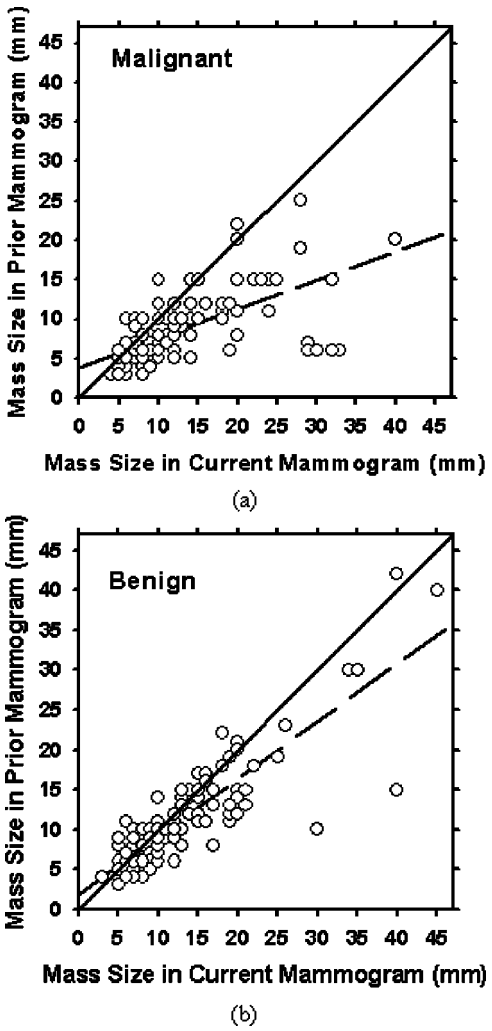


FIG. 2. Mass sizes measured by an MQSA radiologist on the current mammograms plotted against those on the prior mammograms for (a) 115 malignant and (b) 136 benign temporal pairs. The diagonal line on the graph represents the case when the current and the prior mass sizes are identical. The dashed lines are the linear regression lines defined by $y=0.366x+3.913$ for (a) and by $y=0.721x+1.935$ for (b). The correlation coefficient for the malignant masses is 0.39 and for the benign masses is 0.73.

graphic images of a biopsy proven malignant mass or a mass that was followed up and was found to be malignant when biopsy was performed in a future year. 154 of the 318 temporal pairs were CC-view pairs, 138 were MLO-view pairs, and 26 were lateral view pairs. The masses on each of the original mammograms were marked with a bounding box by a Mammography Quality Standards Act (MQSA) radiologist. The radiologist also provided a description of the characteristics of each mass and marked the nipple location on every film. The mass size, defined as the longest dimension of the mass, was measured by the radiologist on both the current and prior mammograms. Figure 2 shows the distribution of the mass sizes. Only 251 temporal pairs were plotted (115 malignant and 136 benign) due to the fact that the masses on the prior mammograms in the remaining 67 temporal pairs were too subtle for the radiologist to estimate their boundaries.

## EVALUATION METHODS

The starting point of the registration algorithm was the biopsy-proven mass location on the current mammogram, and the result of the algorithm was the location or the region of interest found by our algorithm using one of the twelve similarity measures. We evaluated the twelve similarity measures based on four criteria. The first criterion was the mean Euclidean distance. This refers to the average distance from the location where the similarity measure reports the best match between the current template and the sub-region on the prior mammogram, to the center of the bounding box of the mass, as marked by the radiologist. The second and the third criteria were based on the overlap between the current mass template at the best-match location and the bounding box of the true mass on the prior image, which is defined as:

$$\text{overlap} = \frac{\tilde{I}_{\text{current}}^{\text{best-match}} \cap \tilde{I}_{\text{prior}}^{\text{true}}}{\min(\tilde{I}_{\text{current}}^{\text{best-match}}, \tilde{I}_{\text{prior}}^{\text{true}})}, \tag{24}$$

where $\tilde{I}_{\text{current}}^{\text{best-match}}$ was the current mass template at the best-match location on prior image and $\tilde{I}_{\text{prior}}^{\text{true}}$ was the bounding box of the true mass on the prior image. The second criterion was the 50% overlap threshold criterion, defined as the percentage of pairs for which, at the best-match location, the overlap between the current mass template and the bounding box of the true mass on the prior image was 50% or more. The third criterion was the 75% overlap threshold criterion, defined as the percentage of pairs for which, at the best-match location, the overlap between the current mass template and the bounding box of the true mass on the prior image was 75% or more. These first three criteria were used to judge the accuracy of matching using a given similarity measure. We also estimated the statistical significance of the difference between the mean Euclidean distances obtained from the different similarity measures by the Student's paired t-test. The last criterion was designed to evaluate the robustness of the similarity measure. It estimated the change in the accuracy of matching using the similarity measure as a function of the search region size. To obtain a numerical representation of robustness we first calculated the slope between the successive points along the mean Euclidean distance-versus-search region size curve for each similarity measure. For a given similarity measure, a smaller slope will reflect smaller change in the Euclidean distance between two differ-

TABLE I. Number of temporal pairs used for each search region size in Experiment 1.

| Search region sidelength (mm) | Number of temporal pairs within search region |
|---|---|
| 16.8 | 234 |
| 20.0 | 249 |
| 24.8 | 269 |
| 28.0 | 280 |
| 32.8 | 287 |
| 40.8 | 303 |
| 48.8 | 309 |

TABLE II. Mean Euclidean distance and standard deviation of twelve similarity measures using a different subset of pairs for each search region size (refer to Table I). (Experiment 1). The rank of each similarity measure (between 1 and 12) at each window size is also shown.

| Size (mm) | 16.8 | | 20 | | 24.8 | | 28 | | 32.8 | | 40.8 | | 48.8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | err | rnk | err | rnk | err | rnk | err | rnk | err | rnk | err | rnk | err | rnk |
| Correlation | 3.1±3.2 | 1 | 3.4±3.9 | 1 | 4.1±4.9 | 2 | 4.5±5.5 | 1 | 4.9±6.1 | 1 | 6.3±8.3 | 1 | 7.5±9.8 | 1 |
| Cosine | 3.3±3.5 | 3 | 3.7±4.1 | 3 | 4.0±4.6 | 1 | 4.6±5.8 | 2 | 5.2±6.5 | 2 | 7.0±9.0 | 2 | 8.9±11.5 | 2 |
| Median | 3.2±3.3 | 2 | 3.7±4.1 | 2 | 4.4±5.3 | 3 | 5.4±6.5 | 4 | 6.9±8.0 | 6 | 10.5±11.1 | 6 | 15.8±14.2 | 9 |
| Gamma | 3.5±3.7 | 4 | 3.8±4.2 | 4 | 4.6±5.1 | 4 | 5.0±5.6 | 3 | 5.6±6.4 | 3 | 7.0±8.4 | 3 | 9.8±11.2 | 3 |
| Mutual Information (Scaled) | 3.5±3.7 | 5 | 4.0±4.2 | 5 | 4.6±5.3 | 5 | 5.4±6.5 | 5 | 6.4±7.6 | 5 | 9.5±10.7 | 4 | 12.1±12.7 | 4 |
| Ordinal Measure | 3.7±3.8 | 6 | 4.0±4.4 | 6 | 4.9±5.6 | 6 | 5.6±6.4 | 6 | 6.3±7.3 | 4 | 10.1±11.3 | 5 | 15.0±15.2 | 7 |
| Mutual Information (Unscaled) | 4.0±4.0 | 7 | 4.7±4.7 | 7 | 5.7±6.1 | 7 | 7.0±7.4 | 7 | 8.1±8.3 | 7 | 11.2±10.9 | 7 | 14.4±12.7 | 6 |
| Increment Sign Correlation | 4.9±4.0 | 10 | 5.4±4.7 | 8 | 6.3±5.8 | 8 | 7.2±6.5 | 8 | 8.2±7.3 | 8 | 11.2±9.8 | 8 | 15.1±12.7 | 8 |
| Pattern Intensity | 4.6±4.4 | 9 | 5.4±5.3 | 9 | 6.5±6.5 | 9 | 7.3±7.2 | 9 | 8.8±8.3 | 9 | 11.2±10.1 | 9 | 13.4±11.6 | 5 |
| Rank Transform | 4.2±4.0 | 8 | 5.5±5.5 | 10 | 7.5±7.1 | 10 | 9.7±8.6 | 11 | 11.6±9.7 | 11 | 17.3±12.9 | 11 | 24.0±14.2 | 12 |
| Extended Jaccard | 5.6±4.9 | 11 | 6.2±5.5 | 11 | 7.6±6.5 | 11 | 8.6±7.6 | 10 | 10.0±8.8 | 10 | 13.2±10.9 | 10 | 16.7±13.3 | 10 |
| Gradient Difference | 7.9±4.4 | 12 | 9.4±5.3 | 12 | 11.7±6.2 | 12 | 13.4±7.1 | 12 | 15.6±8.2 | 12 | 19.6±9.7 | 12 | 22.6±11.2 | 11 |

ent search region sizes, thereby indicating that the similarity measure is less dependent on the search region size. To summarize the absolute change of the Euclidean distance for several successive search region sizes, we computed the sum-squares of the slopes along the curve. The sum of the squared slopes provided a measure of how sensitive the similarity measure was to a change in the search region size and thus could serve as an index for robustness.

## RESULTS

For this study we used a total of 318 temporal pairs. The average size of the current templates, was 17 mm × 17 mm. We conducted three experiments with different subsets of the 318 temporal pairs and different sizes of the search regions (Table I). The results for the 12 similarity measures for all the search region sizes are given in Tables II–VI and Figs.

TABLE III. Percentage of pairs that surpass the 50% overlap threshold, using a different subset of pairs for each search region size (refer to Table I). (Experiment 1). The rank of each similarity measure (between 1 and 12) at each window size is also shown.

| Size (mm) | 16.8 | | 20 | | 24.8 | | 28 | | 32.8 | | 40.8 | | 48.8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | rnk | % | rnk | % | rnk | % | rnk | % | rnk | % | rnk | % | rnk |
| Correlation | 92 | 1 | 90 | 1 | 87 | 1 | 84 | 1 | 83 | 1 | 78 | 1 | 76 | 1 |
| Cosine | 89 | 3 | 87 | 3 | 84 | 3 | 82 | 2 | 80 | 2 | 75 | 2 | 70 | 2 |
| Median | 91 | 2 | 88 | 2 | 84 | 2 | 79 | 4 | 73 | 6 | 60 | 6 | 47 | 6 |
| Gamma | 88 | 6 | 87 | 4 | 81 | 4 | 79 | 3 | 76 | 3 | 73 | 3 | 65 | 3 |
| Mutual Information (Scaled) | 89 | 4 | 85 | 6 | 81 | 5 | 77 | 6 | 74 | 5 | 65 | 4 | 59 | 4 |
| Ordinal Measure | 87 | 7 | 86 | 5 | 80 | 6 | 78 | 5 | 76 | 4 | 63 | 5 | 51 | 5 |
| Mutual Information (Unscaled) | 83 | 8 | 79 | 8 | 74 | 7 | 66 | 7 | 62 | 7 | 54 | 7 | 47 | 7 |
| Increment Sign Correlation | 79 | 10 | 73 | 10 | 69 | 8 | 64 | 8 | 61 | 8 | 52 | 8 | 43 | 9 |
| Pattern Intensity | 79 | 9 | 75 | 9 | 68 | 10 | 64 | 9 | 59 | 9 | 50 | 9 | 46 | 8 |
| Rank Transform | 88 | 5 | 79 | 7 | 68 | 9 | 57 | 10 | 50 | 11 | 35 | 11 | 24 | 11 |
| Extended Jaccard | 70 | 11 | 67 | 11 | 59 | 11 | 55 | 11 | 52 | 10 | 45 | 10 | 39 | 10 |
| Gradient Difference | 55 | 12 | 45 | 12 | 32 | 12 | 28 | 12 | 24 | 12 | 18 | 12 | 16 | 12 |

TABLE IV. Percentage of pairs that surpass the 75% overlap threshold, using a different subset of pairs for each search region size (refer to Table I). (Experiment 1). The rank of each similarity measure (between 1 and 12) at each window size is also shown.

| Size (mm) | 16.8 | | 20 | | 24.8 | | 28 | | 32.8 | | 40.8 | | 48.8 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | % | rnk | % | rnk | % | rnk | % | rnk | % | rnk | % | rnk | % | rnk |
| Correlation | 82 | 1 | 81 | 1 | 78 | 1 | 75 | 2 | 74 | 2 | 70 | 1 | 68 | 1 |
| Cosine | 81 | 2 | 80 | 2 | 77 | 2 | 76 | 1 | 74 | 1 | 69 | 2 | 65 | 2 |
| Median | 79 | 3 | 78 | 3 | 74 | 3 | 69 | 3 | 64 | 5 | 52 | 6 | 41 | 7 |
| Gamma | 76 | 5 | 76 | 5 | 70 | 5 | 69 | 4 | 67 | 3 | 64 | 3 | 58 | 3 |
| Mutual Information (Scaled) | 76 | 6 | 73 | 6 | 69 | 6 | 66 | 6 | 64 | 6 | 56 | 4 | 50 | 4 |
| Ordinal Measure | 77 | 4 | 76 | 4 | 70 | 4 | 68 | 5 | 66 | 4 | 55 | 5 | 46 | 5 |
| Mutual Information (Unscaled) | 71 | 7 | 67 | 7 | 63 | 7 | 59 | 7 | 55 | 7 | 49 | 7 | 42 | 6 |
| Increment Sign Correlation | 61 | 10 | 59 | 10 | 54 | 9 | 51 | 9 | 49 | 9 | 43 | 9 | 36 | 9 |
| Pattern Intensity | 68 | 8 | 64 | 8 | 59 | 8 | 55 | 8 | 49 | 8 | 43 | 8 | 40 | 8 |
| Rank Transform | 68 | 9 | 60 | 9 | 50 | 10 | 42 | 11 | 39 | 11 | 29 | 11 | 19 | 11 |
| Extended Jaccard | 56 | 11 | 53 | 11 | 48 | 11 | 46 | 10 | 43 | 10 | 38 | 10 | 32 | 10 |
| Gradient Difference | 18 | 12 | 13 | 12 | 9 | 12 | 8 | 12 | 8 | 12 | 6 | 12 | 5 | 12 |

3–9. The results for the mean Euclidean distance were plotted as two groups of six similarity measures each, for clarity of the presentation. The division into these two groups was based on the mean Euclidean distances of the similarity measures at the 24.8 mm $\times$ 24.8 mm search region size. The measures with the six lowest mean Euclidean distances for this size were shown in Fig. 3 and those with the six highest were shown in Fig. 4. In the tables we also presented the performance ranks of the similarity measures for each specific size of the search region.

In order to study the accuracy and robustness of the twelve similarity measures we used seven search region sizes (Table I). The original size of the search region, chosen in our previous study based on the performance of the three-

TABLE V. Comparison of the performance of the twelve similarity measures in terms of the mean Euclidean distance and standard deviation, percentages of pairs with the 50% overlap and 75% overlap threshold criteria for the 269 and 318 temporal pairs of using a search region size of 24.8 mm $\times$ 24.8 mm.

| | Mean Euclidean distance (mm) | | 50% overlap threshold (%) | | 75% overlap threshold (%) | |
|---|---|---|---|---|---|---|
| Experiment # | 2 | 3 | 2 | 3 | 2 | 3 |
| No. of pairs | 269 | 318 | 269 | 318 | 269 | 318 |
| Correlation | 4.1±4.9 | 6.4±8.9 | 87 | 79 | 78 | 70 |
| Cosine | 4.0±4.6 | 6.4±8.9 | 84 | 76 | 77 | 69 |
| Median | 4.4±5.3 | 6.9±9.3 | 84 | 76 | 74 | 66 |
| Gamma | 4.6±5.1 | 6.9±8.9 | 81 | 74 | 70 | 64 |
| Mutual Information (Scaled) | 4.6±5.3 | 7.0±9.1 | 81 | 73 | 69 | 61 |
| Ordinal Measure | 4.9±5.6 | 7.3±9.5 | 80 | 72 | 70 | 62 |
| Mutual Information (Unscaled) | 5.7±6.1 | 8.4±10.2 | 74 | 66 | 63 | 56 |
| Increment Sign Correlation | 6.3±5.8 | 8.9±9.7 | 69 | 62 | 54 | 48 |
| Pattern Intensity | 6.5±6.5 | 8.7±9.09 | 68 | 61 | 59 | 53 |
| Rank Transform | 7.5±7.1 | 10.2±10.6 | 68 | 60 | 50 | 44 |
| Extended Jaccard | 7.6±6.5 | 9.4±8.9 | 59 | 54 | 48 | 44 |
| Gradient Difference | 11.7±6.2 | 13.8±9.1 | 32 | 30 | 9 | 9 |

TABLE VI. Robustness index of the twelve similarity measures in the three different experiments.

| Experiment # | 1 | 2 | 3 |
|---|---|---|---|
| No. of pairs | Region size dependent | 269 | 318 |
| Pearson's Correlation | 0.08 | 0.04 | 0.03 |
| Cosine | 0.16 | 0.09 | 0.08 |
| Median Correlation | 0.84 | 0.63 | 0.64 |
| Gamma | 0.18 | 0.12 | 0.11 |
| Mutual Information (Scaled) | 0.36 | 0.22 | 0.23 |
| Ordinal Measure | 0.66 | 0.47 | 0.49 |
| Mutual Information (Unscaled) | 0.53 | 0.41 | 0.34 |
| Increment Sign Correlation | 0.49 | 0.34 | 0.32 |
| Pattern Intensity | 0.32 | 0.29 | 0.17 |
| Rank Transform | 1.85 | 1.55 | 1.4 |
| Extended Jaccard | 0.53 | 0.41 | 0.37 |
| Gradient Difference | 0.88 | 0.76 | 0.66 |

stage registration technique,[6] was 24.8 mm $\times$ 24.8 mm (31 $\times$ 31 pixels). We defined six additional search region sizes. Four of them: 28 mm $\times$ 28 mm, 32.8 mm $\times$ 32.8 mm, 40.8 mm $\times$ 40.8 mm, and 48.8 mm $\times$ 48.8 mm, were larger than the original size of 24.8 mm $\times$ 24.8 mm. The remaining two search region sizes: 16.8 mm $\times$ 16.8 mm, and 20 mm $\times$ 20 mm, were smaller than the original size. The new search regions were defined by centering at the centroids of the original search regions by changing the sidelength of the square region.

In the first experiment for each search region size we analyzed the subset of the 318 temporal pairs that had the mass centroids on the prior mammogram located inside the search region. We studied the performance of the similarity measures with the seven different search region sizes. The centroid locations of the original search regions were defined after the first two stages of the registration procedure.[6] The number of temporal pairs used for each search region size therefore varied and was given in Table I. The performance results of the twelve similarity measures are presented in Tables II–IV and Figs. 3–6. The mean Euclidean distance for Pearson's correlation coefficient was 3.1 mm for the smallest search region size, 4.1 mm for the original search region size (24.8 mm $\times$ 24.8 mm) and 7.5 mm in the largest search region size, while for mutual information (scaled) the results were 3.5, 4.6, and 12.1 mm for the smallest, original and largest search regions, respectively (Table II and Fig. 3). The percentage of pairs surpassing the 50% overlap threshold were 92%, 87%, and 76% in the smallest, original, and largest search regions, respectively, using correlation, and 89%,

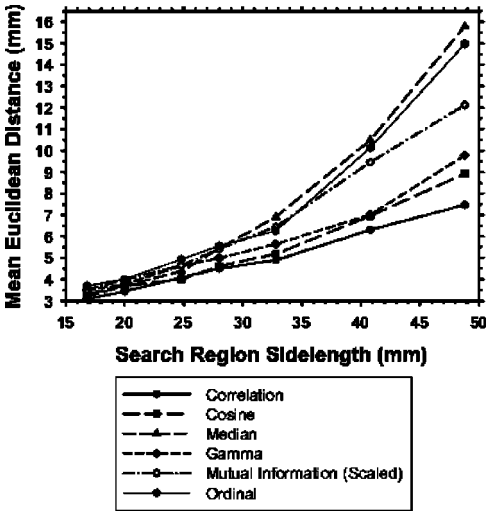


FIG. 3. Mean Euclidean distance of six of the twelve similarity measures for Experiment 1, in which the number of prior masses inside the search region varied with the region size (refer to Table I).

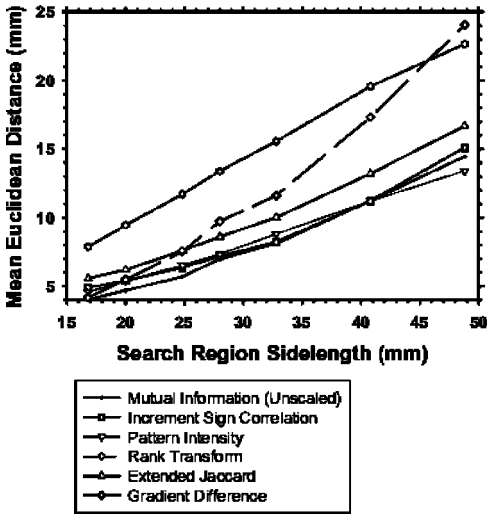


FIG. 4. Mean Euclidean distance of the remaining six (out of twelve) similarity measures for Experiment 1, in which the number of prior masses inside the search region varied with the region size (refer to Table I).

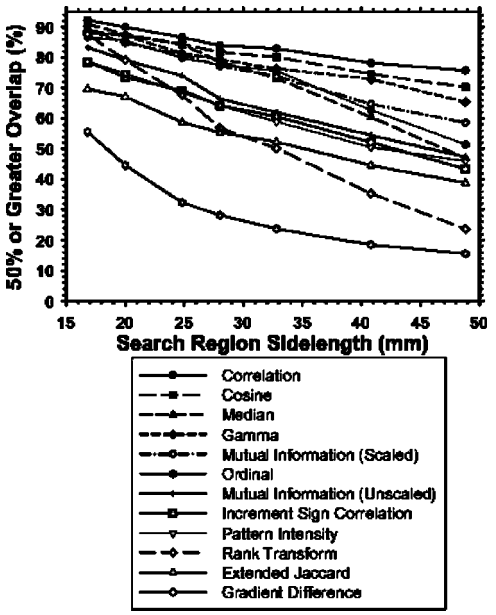


FIG. 5. Percentage of pairs that surpass the 50% overlap threshold for Experiment 1, in which the number of prior masses inside the search region varied with the region size (refer to Table I).

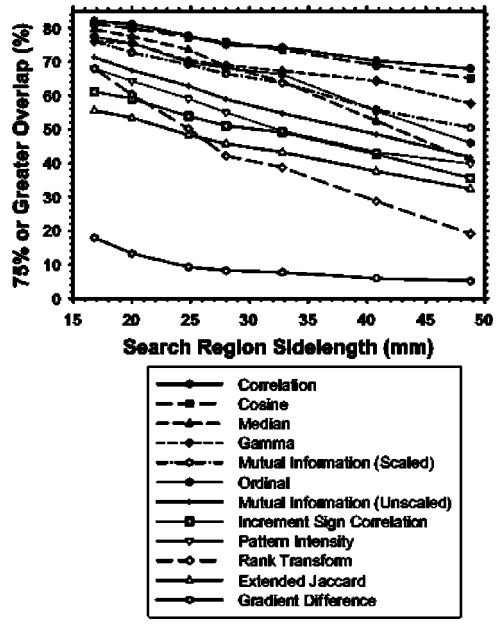


FIG. 6. Percentage of pairs that surpass the 75% overlap threshold for Experiment 1, in which the number of prior masses inside the search region varied with the region size (refer to Table I).

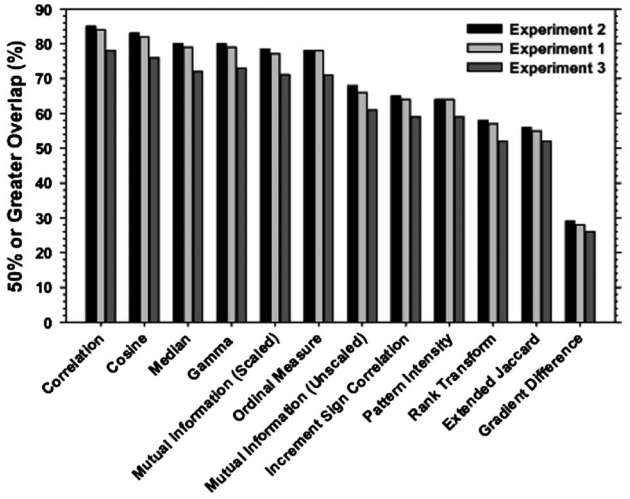


FIG. 8. Percentage of pairs that surpass the 50% overlap threshold for search region size of 28 mm × 28 mm in Experiment 1 (number of prior masses inside the search region varied with the region size: for 28 mm × 28 mm, 280 temporal pairs were used), Experiment 2 (269 temporal pairs), and Experiment 3 (318 temporal pairs).

81%, and 59%, respectively, using mutual information (Table III and Fig. 5). For the 75% overlap threshold the percentage of pairs were 82%, 78%, and 68% using correlation, and 76%, 69%, and 50% using mutual information (Table IV and Fig. 6).

In the second experiment we used a fixed subset of 269 pairs of the total 318 pairs applied to five different search region sizes. In this way we studied the effect of the increasing search area over the performance of the similarity measures for the same data set. The 269 temporal pairs were the ones for which the centroid of their mass on the prior mammogram was inside the 24.8 mm × 24.8 mm search regions. The location of the search regions was defined by the first two stages of the registration procedure.[6] By using these 269 temporal pairs we studied the performance of the similarity measures with the additional four larger different search region sizes. The results are presented in Table V and Figs. 7–9. In order to limit the number of tables and figures in this paper we presented only the results for the original search region size of 24.8 mm × 24.8 mm (Table V) and the search region of 28 mm × 28 mm (Figs. 7–9). These search regions were selected because they were found to be in the range of best performance when we plotted the entire curves. For the discussion of the performance of the similarity measures

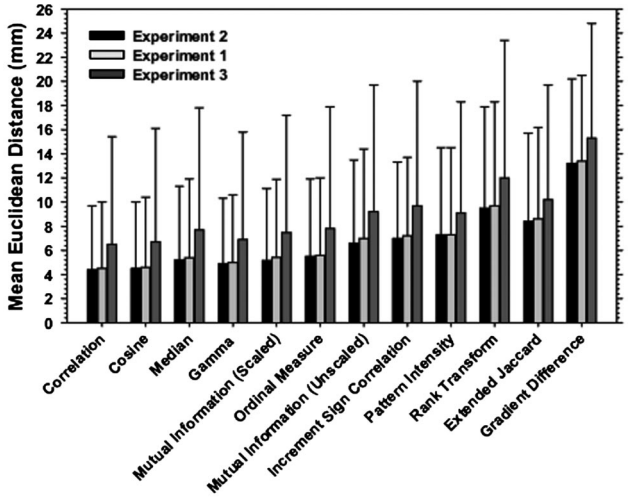


FIG. 7. Comparison of mean Euclidean distance of the twelve similarity measures for search region size of 28 mm × 28 mm in Experiment 1 (number of prior masses inside the search region varied with the region size: for 28 mm × 28 mm, 280 temporal pairs were used), Experiment 2 (269 temporal pairs), and Experiment 3 (318 temporal pairs).

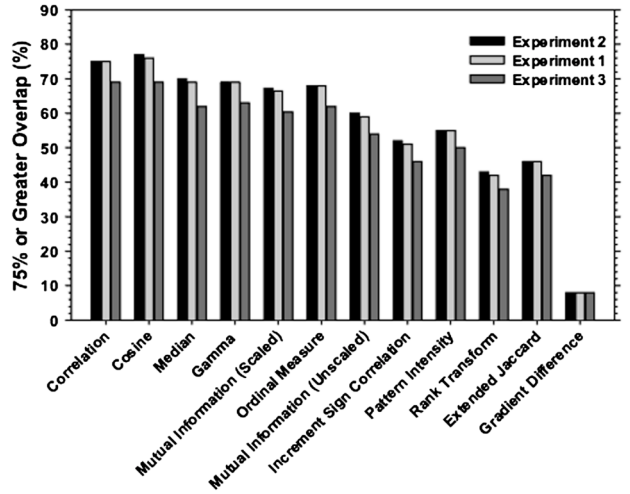


FIG. 9. Percentage of pairs that surpass the 75% overlap threshold for search region size of 28 mm × 28 mm in Experiment 1 (number of prior masses inside the search region varied with the region size: for 28 mm × 28 mm, 280 temporal pairs were used), Experiment 2 (269 temporal pairs), and Experiment 3 (318 temporal pairs).

with the remaining search region sizes we will use the relative ranks of the similarity measures compared to Experiment 1. The mean Euclidean distance for Pearson's correlation coefficient was 4.1 mm in the smallest search region size for this experiment (24.8 mm × 24.8 mm) (Table V) and 6.6 mm in the largest search region size (48.8 mm × 48.8 mm), while for mutual information the results were 4.6 and 10.6 mm for the smallest (Table V) and largest search regions, respectively. The percentage of pairs with the 50% overlap threshold criterion were 87% and 78% in the smallest and largest search regions, respectively, using correlation, and 81% and 62%, respectively, using mutual information (Table V). For the 75% overlap threshold criterion, the percentage of pairs were 78% and 70% using correlation, and 69% and 54% using mutual information (Table V). The mean Euclidean distance for all the similarity measures was slightly smaller than the corresponding one in Experiment 1, except for those of the 24.8 mm × 24.8 mm region size which were identical. The percentage of pairs with the 50% and the 75% overlap threshold criteria were also slightly increased compared to the Experiment 1. However, the ranking of the similarity measures was largely the same as in the case of the Experiment 1. Occasionally, there was a swap in the ranking by one position for a small number of the similarity measures and for some of the search region sizes.

In the third experiment we used all 318 temporal pairs when evaluating the similarity measures. When using this complete set of 318 pairs it is important to note that some of the masses were actually located outside of the search region. The smaller the search region was, the more masses would be located outside of it. When a mass falls outside of the search region the function of the similarity measure is futile, since there is no chance for it to match with the prior mass, and as a result large errors may occur. Here we tested all seven search region sizes as in the Experiment 1. The results are presented again as in the above experiment for the original search region size of 24.8 mm × 24.8 mm (Table V) and the search region of 28 mm × 28 mm (Figs. 7–9). The mean Euclidean distance slightly increased at all corresponding search region sizes when compared to the results from the first two experiments, i.e., with the different subset of pairs for each search region size (Experiment 1) and the fixed subset of 269 pairs (Experiment 2), respectively. However, the ranking of the similarity measures was very similar to the ranking of the first experiment, as can be seen from Figs. 7–9.

The performance for the three experiments are presented in Figs. 7–9, using the search region size of 28 mm × 28 mm. In addition, a comparison between the results for the Experiment 2 and Experiment 3 for the original search region size of 24.8 mm × 24.8 mm are shown in Table V.

The robustness results for all similarity measures and the three experiments are presented in Table VI. A smaller value of the robustness index indicates that the similarity measure is less dependent on the search region size and thus is more robust. The correlation similarity measure shows the best robustness among all twelve similarity measures in all experiments.

For all experiments described above, the current mass template was determined by the bounding box marked by the radiologist. We performed additional experiments to investigate the effects of a change in the current mass template size and in the position on the similarity measure results. For this purpose, we increased the size of the current mass template and also shifted its center within a specified range from the center of the bounding box marked by the radiologist. The size of the template was increased in two different ways. The first methods was to add a fixed number of pixels to each side of the bounding box—five and 10 pixels (4 and 8 mm, respectively) were added to the bounding box in each direction. The second method was to increase the size of the bounding box of the current mass by a percentage of its dimensions—an increase of 15% and 30% were evaluated. Thus, overall, five different template sizes were generated, including the original bounding box marked by the radiologist. The current template centers were shifted uniformly within specified bounds in both the $x$ and $y$ directions. The bounds in the $x$ and $y$ directions were determined as 20% and 30% of the respective dimensions of the original bounding box. Including the original unshifted bounding box, there were three experiments based on shifting the current template. Thus, these three experiments in combination with the five experiments for increasing the size of the template yielded a total of fifteen trials. The trials were carried out for a 28 mm × 28 mm prior search region. The results for these fifteen experiments for five of the similarity measures (correlation, cosine, gamma, mutual information, and increment sign correlation) are given in Table VII. A relatively small change in the Euclidean distance error was observed when the template was enlarged. The largest change was obtained when 10 pixels were added to each side of the current template. A larger effect was observed when the centroids of the current templates were shifted. A combination of an enlarged template and a larger shift gave better results than a smaller template and the same amount of shift. A possible reason for this observation is that when it is shifted, the enlarged template more likely will enclose a larger portion of the mass, which will result in a more successful matching within the search region on the prior mammogram.

## DISCUSSION

Each one of the three experiments serves a different purpose. In the first experiment only the temporal pairs that had the prior mass inside the search region were included in the analysis and the number of pairs will increase with increasing search region size. The second experiment was similar to the first in that all temporal pairs would have the prior mass inside the search region. However, it differed in that we had fixed the masses being analyzed to be those inside the search region of 24.8 mm × 24.8 mm. In other words, the same 269 temporal pairs of masses were analyzed for all sizes of the search regions. This experiment separated the increased

TABLE VII. Mean Euclidean distance and standard deviation of correlation, cosine, gamma, mutual information and increment sign of correlation for the increased size of the current mass template and shifted template center within a specified range from the center of the bounding box marked by the radiologist. The current mass template was increased by adding 5 and 10 pixels (4 mm and 8 mm, respectively) to the bounding box in each direction and by increasing the bounding box by 15% and by 30% of the template dimension. The current template centers were shifted using a shift amount uniformly distributed within 20% and 30% of the respective dimensions of the original bounding box in both the *x* and *y* directions. Additionally, the results of the original unshifted bounding boxes are included.

| Displacement from the centroid | Change in template size | Correlation | Cosine | Gamma | Mutual Information | ISC |
|---|---|---|---|---|---|---|
| 0% | 0 | 4.5±5.5 | 4.6±5.8 | 5.0±5.6 | 5.4±6.5 | 7.2±6.5 |
| | 5 pixels | 4.7±5.5 | 5.2±5.8 | 5.0±5.8 | 6.3±6.5 | 7.6±6.6 |
| | 10 pixels | 5.4±5.9 | 5.7±6.0 | 5.7±5.8 | 6.7±6.5 | 8.0±6.5 |
| | 15% | 4.4±5.1 | 4.9±6.0 | 4.5±5.2 | 5.5±6.1 | 7.5±6.7 |
| | 30% | 4.7±5.3 | 5.0±5.5 | 4.9±5.4 | 5.9±6.2 | 7.5±6.4 |
| 20% | 0 | 5.4±5.7 | 5.7±6.1 | 6.1±6.2 | 6.9±6.6 | 7.5±6.3 |
| | 5 pixels | 5.4±5.5 | 5.7±5.8 | 5.9±6.0 | 7.3±7.0 | 7.7±6.1 |
| | 10 pixels | 5.9±5.8 | 6.1±5.8 | 6.1±5.6 | 7.5±6.8 | 8.2±6.3 |
| | 15% | 5.4±5.5 | 5.7±6.1 | 5.6±5.8 | 6.7±6.6 | 7.9±6.3 |
| | 30% | 5.2±5.2 | 5.6±5.6 | 5.8±5.8 | 6.8±6.7 | 8.0±6.3 |
| 30% | 0 | 6.5±5.8 | 6.8±6.4 | 7.0±6.2 | 8.1±6.9 | 8.4±6.5 |
| | 5 pixels | 6.2±5.7 | 6.6±6.1 | 6.7±6.1 | 7.7±6.8 | 8.5±6.3 |
| | 10 pixels | 6.5±5.7 | 6.8±5.9 | 6.7±5.7 | 8.0±6.7 | 9.0±6.4 |
| | 15% | 6.2±5.5 | 6.9±6.4 | 6.9±6.1 | 7.7±6.7 | 8.2±6.1 |
| | 30% | 6.1±5.4 | 6.6±5.9 | 6.6±6.0 | 7.6±6.6 | 8.5±6.3 |

chance of false matching due to the increased region size from the increasing number of masses included in the search region. The third experiment in which all 318 temporal pairs were tested modeled a more realistic situation since in real life one cannot guarantee that the prior mass will always lie inside the search region, due to the fact that the first and the second stages of the detection procedure may determine an incorrect search region.

The results from the first and the second experiments differed from those of the third in two main ways. First, the accuracy of all of the similarity measures was higher when the masses were guaranteed to be inside of the search region. As a result the mean Euclidean distances from the first and second experiments were lower than those from the third experiment, and conversely the percentage of pairs surpassing the 50% and 75% overlap thresholds were higher. Second, in the first and second experiments it was observed that the mean Euclidean distance decreased monotonically with decreasing search region size. However, in the third experiment several of the similarity measures had a minimum for the Euclidean distance at the search region size of 24.8 mm $\times$ 24.8 mm (graph not shown). For smaller search region sizes the error actually increased. The explanation for this is that as the search region size was reduced below this size an increasing number of prior masses was left outside of the search region and could not be matched, regardless the performance of the similarity measures, thus contributing more significantly to the error. Despite these differences the results of the three experiments lead to basically the same conclusions about the effectiveness of the different similarity measures. In the discussion that follows we simply chose to use the first experiment to arrive at our conclusions.

The three similarity measures that performed best (lowest mean Euclidean distance and highest percentage of pairs exceeding the overlap thresholds) were Pearson's correlation coefficient, the cosine coefficient, and Goodman and Kruskal's Gamma coefficient. The correlation and cosine coefficients slightly outperformed the Gamma coefficient for all search region sizes. The correlation coefficient performed significantly (paired t-test $p < 0.05$) better than the Gamma coefficients for 4 search region sizes and could not reach statistical significance for the 20 mm $\times$ 20 mm, 28 mm $\times$ 28 mm, and 40.8 mm $\times$ 40.8 mm sizes. The difference between the cosine and the Gamma measures was statistically significant only for the search region size of 24.8 mm $\times$ 24.8 mm. The difference between the correlation coefficient and the cosine coefficient was statistically significant only for the 48.8 mm $\times$ 48.8 mm search region size and not statistically significant ($p$ value ranged from 0.13 to 0.69) for the remaining tested search region sizes. It may be noted, however, that the mean Euclidean distance for the cosine measure was slightly smaller than that for correlation (Table II, Fig. 3) when the search region size was 24.8 mm $\times$ 24.8 mm, and slightly higher for all remaining search region sizes. This gap between the performance of the correlation and cosine measures continued to grow as the size of the search region increased. The $p$ value decreased with increasing search region sizes larger than 28 mm $\times$ 28 mm. The robustness of these three measures was also the highest compared to those of the remaining nine similarity measures.

Among the three similarity measures, Pearson's correlation had the highest robustness index (smallest value) of 0.08, while the indices of the cosine and gamma coefficients were 0.16 and 0.18, respectively (Table VI).

After this first group of three similarity measures, which performed best, six of the remaining nine similarity measures can be grouped together, based on their performances. The second group consists of the correlation standardized by the median, the ordinal measure, both scaled and unscaled versions of mutual information, increment sign correlation, and pattern intensity. All six of these similarity measures had a mean error in the range of 12.1–15.8 mm in the largest search region size used. It can be observed that the correlation standardized by the median started out with a relatively low error for the smallest search region and could even compete with the three similarity measures of the first group for this search region size. However, as the search region size was increased, the error of the correlation standardized by the median increased at a faster rate than did the errors of the other measures in this group of six. On the other hand, both the scaled and unscaled versions of mutual information tended to show a relatively high level of robustness. The scaled version of the mutual information had a smaller mean Euclidean distance than the ordinal measure for all search region sizes except for 32.8 mm $\times$ 32.8 mm. However, the difference was statistically significant only for the largest search region size (48.8 mm $\times$ 48.8 mm). The scaled version of the mutual information had a larger mean Euclidean distance for search region sizes up to 28.0 mm $\times$ 28.0 mm than the median correlation. The unscaled version of mutual information had a larger mean Euclidean distance for search region sizes up to 40.8 mm $\times$ 40.8 mm than the ordinal measure and the median correlation. However, for the largest search region size, the errors of both the ordinal measure and the median correlation surpassed the errors of the two mutual information measures since their growth rates were higher than the growth rates of the unscaled and scaled versions of mutual information.

A direct comparison between the scaled and the unscaled versions of mutual information shows that the scaled versions seemed to outperform the unscaled version at all search region sizes. The difference was statistically significant for all of the search region sizes. For both versions the growth rates of their mean Euclidean distances were relatively linear when the search region size increased. Mutual information proved to be a relatively robust similarity measure, however, since it is a probability-based measure its accuracy is easily influenced by the quality of statistics used to calculate it. In this case, the quality is determined by the number of template pixels used to construct the joint histogram. The number of the template pixels in this application is relatively small which resulted in sparse joint histograms when the number of bins was determined by all (4096) gray scale values. The best results were obtained with very small number of bins and, therefore, not so sparse joint histogram. However, a small number of bins determine a small number of effective grey levels used to calculate the similarity measure, resulting in a coarser matching.

The other two members of this second group of six similarity measures are increment sign correlation and pattern intensity. The performances of these two similarity measures were very similar. The difference between their mean Euclidean distance was not statistically significant at any of the tested search region sizes. However, pattern intensity showed a higher level of robustness than the increment sign correlation, as their mean Euclidean distances grew in a relatively linear fashion.

We place the remaining three similarity measures in the third group. These measures are the rank transform, the extended Jaccard measure, and the gradient difference measure. Of these three measures the extended Jaccard proved to be the most effective. At smaller search region sizes the extended Jaccard had a larger mean Euclidean error than the rank transform. At the search region size of 24.8 mm $\times$ 24.8 mm it had an error of 7.6 mm, about 1 mm larger than the measures from the previous groups, and at the largest search region size it yielded an average error of 16.7 mm, again about 1 mm higher than the measures from the previous group. The rank transform measure showed better results compared to the extended Jaccard measure at the first three search region sizes. However, its mean Euclidean error increased much faster than that of the extended Jaccard and became significantly larger than those of the extended Jaccard when the search region size was larger than 28 mm $\times$ 28 mm. At the largest search region size, 48.8 mm $\times$ 48.8 mm, the rank transform similarity measure had a mean Euclidean distance of 24 mm. This was the highest recorded error among all similarity measures for that search region size. The rank transform with a robustness index of 1.85 was clearly the least robust of the twelve similarity measures, and was, therefore, highly ineffective for search region sizes with dimension larger than about 30 mm $\times$ 30 mm. Finally, we can conclude that the gradient difference measure is the least useful similarity measure for the task of temporal-pair matching. For the original search region size of 24.8 mm $\times$ 24.8 mm, the mean Euclidean distance for the gradient difference was 11.7 mm, which is more than 4 mm greater than the measure that had the second largest error. For all the other tested search region sizes except the largest one (48 mm $\times$ 48 mm), the gradient difference continued to have the highest mean Euclidean distance. For the largest search region size, the error of the gradient difference was slightly smaller than that of the rank transform, however, this difference was not statistically significant.

The comparisons using the 50% and 75% overlap thresholds agreed with those using the mean Euclidean distance. The correlation and the cosine coefficients yielded the highest percentages of images to surpass the 50% overlap threshold as well as the 75% overlap threshold for all the tested search region sizes. Again the Goodman and Kruskal's Gamma followed closely behind these two measures. These results were also consistent with the robustness of the similarity measures, which was described previously with respect to mean Euclidean distance. The percentage of images that

exceeded the 50% overlap threshold and the 75% overlap threshold decreased as the search region size increased for all similarity measures. However, this decrease was noticeably steeper for the similarity measures that were previously described as less robust with respect to the mean Euclidean distance. The ordinal measure, the median correlation, and the rank transform are the clearest examples of this trend. The overlap threshold criteria further confirmed the ineffectiveness of the gradient difference as a similarity measure for this task.

The change of the size and position of the current mass template had relatively small effect on the similarity measures performance. The effect was smaller when the template was enlarged and slightly larger when the current template was shifted from the center of the bounding box marked by the radiologist.

We additionally studied whether the comparison of the similarity measures that was obtained using the search regions found by the first and the second stages of our registration algorithm is still valid if the search regions were found by a different method. For this purpose we generated uniform distributions of locations which served as the estimated mass centers for the search regions in the third stage of the registration algorithm. The use of the uniform distribution would represent a pessimistic distribution of the centers of the search regions. Three different experiments were performed. In the first experiment, the centroids of all search regions coincided with the true mass centroids on the prior mammograms. This represented the ideal situation. In the second experiment, the centroids of the search regions were uniformly distributed within a radius of 20 pixels from the true location of the mass centroids on prior mammograms. In the third experiment, the centroids were uniformly distributed within a radius of 30 pixels from the true location of the mass centroids. The similarity measures that were evaluated included correlation, cosine, gamma (which were the three best-performed measures), mutual information and increment sign correlation. The results of these experiments showed that the relative performance of the similarity measures is essentially independent of the method used in the original study to find the location of the search region. Even though the mean Euclidean error increased as the radius of the uniform distribution increased from 0 to 30 pixels, the order of the similarity measures was not changed (Fig. 10). The mean Euclidean error at a search region size of 24.8 mm $\times$ 24.8 mm for correlation was 3.5, 4.2, and 5.0 mm for the experiments one, two, and three, respectively, and 4.1 mm for the original distribution from the first and second stages. The mean Euclidean error was higher for all similarity measures when the radius of the uniform distribution was increased from 20 pixels to 30 pixels.

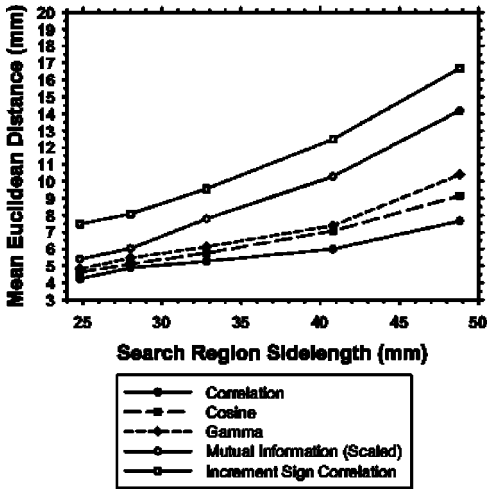


FIG. 10. Mean Euclidean distance of the five (out of twelve) similarity measures for the experiment in which the centroids of the search regions at the third stage were uniformly distributed within a radius of 16 mm (20 pixels) from the true locations of the mass centroids on the prior mammograms.

## CONCLUSION

The results of our study indicate that the best similarity measure which can be used in the third stage of the registration technique developed by Hadjiiski *et al.*[6] is the Pearson's correlation coefficient. The two other similarity measures that can match the performance of the Pearson's correlation, without significantly sacrificing accuracy and robustness, are the cosine coefficient and Goodman and Kruskal's Gamma coefficient. The mean Euclidean distance for the Pearson's correlation was 3.1 mm in the smallest search region, 4.1 mm for the original search region (24.8 mm $\times$ 24.8 mm) and 7.5 mm in the largest search region, respectively. The cosine coefficient achieved very similar results as the differences between their mean Euclidean distances were not statistically significant, except for the largest tested search region size. The Gamma coefficient was slightly behind the correlation and cosine measures but the difference between the Gamma coefficient and the cosine measure achieved statistical significance only for the search region size of 24.8 mm $\times$ 24.8 mm. The main disadvantage of the Gamma coefficient was its slow execution when compared to that of the correlation and cosine. Furthermore we found that the widely used mutual information similarity measure is not among the most effective measures for the task of matching masses on serial mammograms using small size templates. The mean Euclidean distance for the version of mutual information that had the better performance, i.e., the scaled version, was 3.8, 5.7, and 13.2 mm for the smallest, original and largest search region sizes, respectively. The robustness of the Pearson's correlation, which had a robustness index of 0.08, was also superior to that of the mutual information (scaled version), which had an index of 0.39. Lastly we identified three similarity measures, the extended Jaccard measure, the rank transform, and the gradient difference, that were least effective for the task of matching masses on temporal pairs of mammograms. Although matching masses on serial mammograms was the task of interest in this study, we expect that our results will have implication for similar template-matching tasks that use small-sized templates.

## ACKNOWLEDGMENT

This work was supported by a U.S. Army Medical Research and Materiel Command (USAMRMC) Grant DAMD17-02-1-0489.

[a]Author to whom correspondence should be addressed. Telephone: (734) 647-7428; Fax: (734) 615-5513. Electronic mail: lhadjisk@umich.edu

[1]M. Sallam and K. Bowyer, "Detecting abnormal densities in mammograms by comparison with previous screenings," in *Digital Mammography '96*, edited by K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt (Elsevier, Amsterdam, 1996).

[2]N. Vujovic and D. Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," IEEE Trans. Image Process. **6**, 1388–1399 (1997).

[3]S. Sanjay-Gopal, H. P. Chan, T. Wilson, M. Helvie, N. Petrick, and B. Sahiner, "A regional registration technique for automated interval change analysis of breast lesions on mammograms," Med. Phys. **26**, 2669–2679 (1999).

[4]L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, M. A. Helvie, and S. Sanjay-Gopal, "Automated identification of breast lesions in temporal pairs of mammograms for interval change analysis," Radiology **213**(P), 229–230 (1999).

[5]W. F. Good, B. Zheng, Y. H. Chang, Z. H. Wang, and G. S. Maitz, "Generalized procrustean image deformation for substraction of mammograms," Proc. SPIE **3661**, 1562–1573 (1999).

[6]L. M. Hadjiiski, H. P. Chan, B. Sahiner, N. Petrick, and M. A. Helvie, "Automated registration of breast lesions in temporal pairs of mammograms for interval change analysis—local affine transformation for improved localization," Med. Phys. **28**, 1070–1079 (2001).

[7]B. Zheng, W. F. Good, D. R. Armfield, C. Cohen, T. Hertzberg, J. H. Sumkin, and D. Gur, "Performance change of mammographic CAD schemes optimized with most-recent and prior image databases," Acad. Radiol. **10**, 283–288 (2003).

[8]S. Van Engeland, P. Snoeren, J. Hendriks, and N. Karssemeijer, "A comparison of methods for mammogram registration," IEEE Trans. Med. Imaging **22**, 1436–1444 (2003).

[9]L. M. Hadjiiski, H. P. Chan, B. Sahiner, C. Zhou, M. A. Helvie, and M. A. Roubidoux, "Computerized Regional Registration of Corresponding Masses and Microcalcification Clusters on Temporal Pairs of Mammograms for Interval Change Analysis," Presented at the 89th Scientific Assembly and Annual Meeting of the Radiological Society of North America, Chicago, IL, November 30–December 5, 2003, 389.

[10]C. R. Meyer, J. L. Boes, B. Kim, P. Bland, K. R. Zasadny, P. C. Kison, K. Koral, K. A. Frey, and R. L. Wahl, "Demonstration of accuracy and clinical versatility of mutual information for automatic multimodality image fusion using affine and thin plate spline warped geometric deformations," Med. Image Anal **3**, 195–206 (1997).

[11]F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," IEEE Trans. Med. Imaging **16**, 187–198 (1997).

[12]J. P. W. Pluim, J. B. Antoine Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," IEEE Trans. Med. Imaging **22**, 986–1004 (2003).

[13]S. Kaneko, I. Murase, and S. Igarashi, "Robust image registration by increment sign correlation," Pattern Recogn. **35**, 2223–2234 (2001).

[14]G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. Hill, and D. J. Hawkes, "A comparison of similarity measures for use in 2-D-3-D medical image registration," IEEE Trans. Med. Imaging **17**, 586–595 (1998).

[15]D. N. Bhat and S. K. Nayar, "Ordinal measures for image correspondence," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 415–423 (1998).

[16]J. Banks and M. Bennamoun, "Reliability analysis of the rank transform for stereo matching," IEEE Trans. Syst. Man Cybern. **31**, 870–880 (2001).

[17]B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl, "Item-based collaborative filtering recommendation algorithms," Proc. of the 10th International World Wide Web Conference, Hong Kong, May 2001, pp. 285–295 (2001).

[18]L. A. Goodman and W. H. Kruskal, "Measures of association for cross-classification," J. Am. Stat. Assoc. **49**, 732–764 (1954).

[19]A. Agresti, "The effect of category choice on some ordinal measures of association," J. Am. Stat. Assoc. **71**, 49–55 (1976).

[20]L. Hamers, Y. Hemeryck, G. Herweyers, M. Janssen, H. Keters, R. Rousseau, and A. Vanhoutte, "Similarity measures in scientometric research: The Jaccard index versus Salton's cosine formula," Inf. Process. Management **25**, 315–318 (1989).