# Revisiting Intensity-Based Image Registration Applied to Mammography

Yago Díez, Arnau Oliver, Xavier Lladó, Jordi Freixenet, Joan Martí, Joan Carles Vilanova, and Robert Martí

*Abstract*—**The detection of architectural distortions and abnormal structures in mammographic images can be based on the analysis of bilateral and temporal cases using image registration. This paper presents a quantitative evaluation of state-of-the art intensity based image registration methods applied to mammographic images. These methods range from a global and rigid transformation to local deformable paradigms using various metrics and multiresolution approaches. The aim of this study is to assess the suitability of these methods for mammographic image analysis. Evaluation using temporal cases based on quantitative analysis and a multiobserver study is presented which gives an indication of the accuracy and robustness of the different algorithms. Although previous studies suggested that local deformable methods were not suitable due to the generation of unrealistic distortions, in this work we show that local deformable paradigms (multiresolution B-Spline deformations) obtain the most accurate registration results.**

*Index Terms*—**Image registration, mammography, observer study, quantitative evaluation .**

## I. Introduction

DETECTION of abnormal structures or architectural distortions in mammograms can be performed by comparing images of the same patient, either the same breast taken at different times (temporal comparison) or using the left and right breast (bilateral comparison). This comparison is not straightforward mainly because the breast is a highly dynamic organ whose appearance can physiologically change significantly between sessions. Additional dissimilarities between images related to patient movement, sensor noise, different radiation exposure or variation of breast compression also make this comparison difficult. Therefore, in order to efficiently compare two mammograms and avoid nontarget dissimilarities, an initial alignment using an image registration algorithm must be carried out. Although registration of mammographic images is regarded as an ill-posed problem where the perfectly registered image can never be obtained due to the projective nature of the images, it is still an important research topic for the development of computer aided diagnosis (CAD) systems and it has not yet been included into currently commercially available CAD systems. The reasons behind this reluctance are the mentioned complexity (and computational cost) of the nonrigid registration itself, the possibility of inducing image registration artifacts, and also the relatively recent adoption of full field digital mammography systems which limits the number of temporal cases in order to be clinically evaluated at a larger scale. However, clinical evaluation of such tools is likely to be more a reality with access to recent large scale digital mammography screening trials, such as the Digital Mammographic Imaging Screening Trial (DMIST) [1], specially having in mind the strategy of moving from a single image to a patient centered CAD, which could further improve accuracy of current CAD systems.

Fig. 1 shows an example of mammographic image registration with the target, template, registered images, and image differences. A larger deviation from the mid gray level value (i.e., 127 in 8 bits) means a larger difference in the images. Note that after registration differences are less significant, and microcalcifications (bright spots on the bottom) appear closer. Registration results can be used for tasks such as visualization but also to detect changes [2], [3] or as a form of prior information for CAD systems [4]. The example in Fig. 1 illustrates the use of image registration for the detection of abnormalities. The template image (an image acquired in the last screening round) shows a spiculated lesion in the central breast region which is not visible on the target image (previous screening). Observe that the lesion is highlighted by the image difference. In that sense, the benefits of combining image registration with mass detection algorithms in order to improve detection results have been recently shown in [4].

Image registration has been widely used in medical applications for quite a while now (see for instance the surveys of [5]–[7]), and the analysis of mammographic images is not an exception [8]. Most of the published approaches on mammographic image registration use some kind of image features such as breast boundary information [2], [3], [9] as it is relatively easy to extract and provides important information about breast deformation. In addition to breast boundary, information about the deformation of internal regions has also been used in several approaches in order to obtain a more robust registration. This is the case of the pectoral muscle [9], salient regions extracted using wavelets [2], isointensity contours [9] or steerable filters [10], and internal linear structures [3], [11]. On the other hand, another and less numerous group of approaches can be classified as being intensity-based, where the deformation is recovered optimizing a measure of similarity between images, as in [12] where a regional and intensity-based variational algorithm was presented. The use of an intensity measure to recover global transformations (i.e., rigid or affine) has been reported to

Manuscript received December 31, 2010; revised March 17, 2011; accepted April 23, 2011. Date of publication May 5, 2011; date of current version September 2, 2011.

Y. Díez, A. Oliver, X. Lladó, J. Freixenet, J. Martí and R. Martí are with the Institute of Informatics and Applications, University of Girona, 17071 Girona, Spain (e-mail: {yago; aoliver; llado; jordif; joanm; marly}@eia.udg.edu).

J. C. Vilanova is with Girona Magnetic Resonance Imaging Center, Girona, Spain (e-mail: kvilanova@comg.cat).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITB.2011.2151199

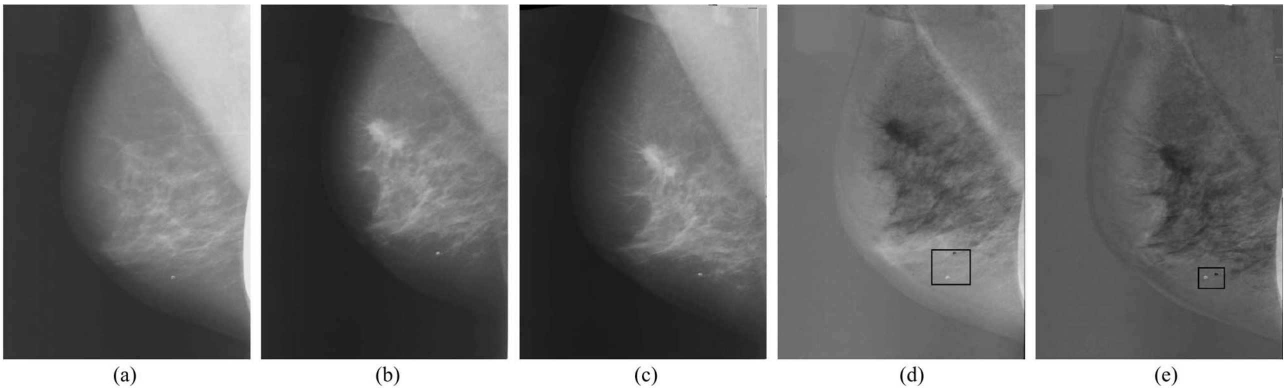


Fig. 1. Nonrigid registration example of mammographic images: (a) Target image, (b) template image, (c) registered template image, (d) difference before registration (a) and (b), (e) difference after registration (a) to (c). Squares in (d) and (e) highlight the difference in the microcalcifications area.

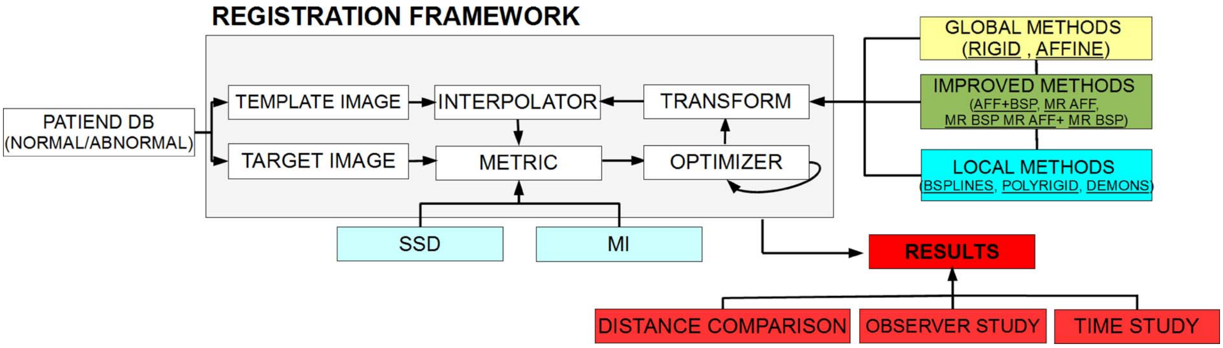

REGISTRATION FRAMEWORK
PATIEND DB (NORMAL/ABNORMAL)
TEMPLATE IMAGE
TARGET IMAGE
INTERPOLATOR
METRIC
TRANSFORM
OPTIMIZER
SSD
MI
GLOBAL METHODS (RIGID , AFFINE)
IMPROVED METHODS (AFF+BSP, MR AFF, MR BSP MR AFF+ MR BSP)
LOCAL METHODS (BSPLINES, POLYRIGID, DEMONS)
RESULTS
DISTANCE COMPARISON
OBSERVER STUDY
TIME STUDY

Fig. 2. General image registration framework used. The template image is transformed in order to optimize a similarity metric computed using the transformed template and the target image. Different transforms and metrics evaluated in this work are shown as subclasses of *metric* and *transform*. Results are evaluated using the resulting similarity metric, TLE error (distance comparison), an observer study and execution time.

obtain robust results for mammographic registration specially compared to nonrigid methods [13]. In their study, nonrigid methods obtained significantly worse results mainly due to the fact that they induced nonrealistic deformations to the images. However, authors evaluated only one nonrigid registration algorithm based on automatic point correspondence using Thin Plate Splines. In this paper, we aim to provide a more thorough insight on the use of intensity-based rigid and nonrigid image registration algorithms to mammographic images.

The main goal of this work is to review and evaluate the applicability of eight state-of-the art intensity based image registration algorithms to mammographic images, more specifically to temporal studies: two images of the same breast acquired at different time intervals, normally different screening rounds which are 2–3 years a part. To our knowledge, this is the first attempt to quantitatively compare rigid and nonrigid intensity-based registration in mammographic images, specially taking into account the quantitative and complementary evaluation criteria used in this work: metric comparison, an observer study, running time analysis, and target localization error (TLE) for a subset of images in terms of the microcalcification distance before and after registration. In total, our data comprise 250 digitized mammograms from 22 patients in both CC (craniocaudal) and MLO (mediolateral oblique) views. Initial results of this work were presented in [14] but a more extensive evaluation (namely observer study and TLE evaluation) and discussion are addressed here. The code of the methods evaluated in the paper is freely available at eia.udg.edu/%7emarly/registration.html. This paper is structured as follows: In Section II, we briefly describe the image registration algorithms and their implementation. Subsequently, registration results are presented, providing details on the data, experiments, and quantitative analysis in Section III. Finally, discussions and conclusions are provided in Section IV.

## II. Methodology

In this section, we provide details on the methodology used, which is essentially based on the widely known general image registration framework [7], as shown in Fig. 2. However, for the sake of completeness, we will discuss in more detail here the metrics and transformations used in the different proposed algorithms, which are also depicted in Fig. 2. In total, we evaluate eight algorithms from the perspective of the transformation used: Rigid (RIG), Affine (AFF), B-Spline Free-Form Deformations (BSP), Polyrigid (PRIG) and Demons (DEM), and combinations of them (I1: multiresolution (MR) BSP; I2: AFF + BSP; I3: MR AFF + MR BSP). Although other algorithm combinations have been tested, we present here the ones which significantly improved results compared to the individual methods.

### A. *Similarity Metric*

The measure of similarity between images or regions is a crucial component in image registration [7] along with the selection of the transformation function. In addition to using the similarity to drive the optimization process of the registration, similarity metrics are also often used to evaluate the performance of image registration proposals under the assumption that a higher similarity between images after registration means better alignment. Although those metrics provide an objective way of computing apparent similarity between images, in some cases they do not completely agree with human visual perception. This will be exemplified in Section III how sometimes better metric values are not perceived by expert observers and can even be neglected in front of other considerations. Consequently, we will use similarity metrics as only one of the criteria for evaluating the most adequate registration method among those analyzed. In this section, we present the two metrics used in this work, sum of squared differences (SSD) and mutual information (MI).

*1) SSD:* The SSD metric computes the squared differences between intensity values for corresponding pixels. This is a simple, widely used metric that assumes a linear relationship between intensities in the images to be compared and its optimal value is 0 (images are identical). Equation 1, where $A$ and $B$ stand for the images and $i$ iterates over the $I$ pixels in the images, shows how to compute this metric.

$$SSD(A, B) = \frac{1}{I}\sum_{i=1}^{N}(A_i - B_i).^2 \tag{1}$$

*2) MI:* MI [15] provides a measure of probabilistic mutual dependence between two intensity distributions. MI allows to account for nonlinear differences in intensity (a feature often useful in multimodality registration) and is defined as

$$MI(A, B) = H(A) - H(B|A) = H(A) + H(B) - H(A, B) \tag{2}$$

where again $A$ and $B$ are the images to be compared, $H(X, Y) = -\sum_{x,y=0}^{N} p_{x,y}\log_2(p_{x,y})$, $H(X) = -\sum_{i=0}^{N} p_x \log_2(p_x)$ represents the joint and individual entropies, respectively, of random variables $X$, $Y$ associated to the images to be compared. Here, $N$ stands for the number of intensity levels and $p_x$ ($p_{xy}$) is the probability of value $x$ ($x$,$y$) in the (joint) probability distribution of variable $X$ ($X$ and $Y$). Here a larger MI value means more similar images. MI can be computed using different approaches; in this paper, we have adopted the implementation of MI devised by Mattes *et al.* [16].

### B. Transformations

The choice of the transformations that images are allowed to undergo influences the whole registration process. In the case of breast registration, two main types of deformations are usually considered [13]. On one hand, the two images to be registered can be globally aligned so that the main anatomical structures (nipple, breast contour, pectoral muscle) match. On the other hand, internal breast structures also carry important information that could be taken into account in order to recover a more local deformation. To try to account for both problems and in line with other authors [13], we propose to evaluate the following image registration methods in this work, divided into global and local methods,

*1) Global Transformations. Rigid and Affine:* We refer to global methods as the ones in which all pixels suffer the same transformation, which often results in simple and fast computation due to its small number of parameters. However, this simplicity makes it hard to account for some of the nonrigid deformation that occur during breast acquisition (i.e., breast compression and/or tissue movement). Rigid and affine transformations are proposed as global transformations [7]. A 2-D rigid transform is composed of a rotation of angle $\theta$, and a translation of vector ($t_x$ and $t_y$), $x$ and $y$ are the original points and $x'$, $y'$ refer to the transformed points.

$$\begin{aligned} x' &= (x\cos(\theta) - y\sin(\theta)) + t_x \\ y' &= (x\sin(\theta) - y\cos(\theta)) + t_y. \end{aligned} \tag{3}$$

Affine transformations allow additional shearing for a total of six parameters in 2-D

$$\begin{aligned} x' &= t_x + a_1 x + a_2 y \\ y' &= t_y + b_1 x + b_2 y \end{aligned} \tag{4}$$

where $t_x$, $t_y$, $x$, $y$, $x'$ and $y'$ refer to the same as in the rigid equation and $a_1$, $a_2$, $b_1$, and $b_2$ represent affine parameters.

*2) Local Transformations. B-Splines, Polyrigid and Demons:* Local methods (also known as deformable registration) include methods where pixels are transformed locally, having a different transformation depending on their local similarity and position. These approaches enable to consider more complex deformations than global methods but, in some cases, this additional deformation capability makes them difficult to capture the global transformations of the image. Moreover, as we will see in Section III-D, sometimes these types of deformations can induce nonrealistic deformations which are not well received by expert observers, as shown in [13]. However, it is our assertion that correct parametrization of these algorithms and in particular choosing transformations with strong regularization constraints minimizes these undesirable effects.

Many methods and variations have been proposed under the local deformation paradigm. In addition to the local computation of the metric, they usually incorporate aspects such as regularization in order to ensure smoothness and continuity, which can be implicit in the transformation or as an added constraint to the transformation function. Among these methods we have selected B-Spline (BSP) free form deformations (FFD) [17], polyrigid transformations [18], and Thirion's Demons algorithm [19], due to its wide popularity in medical applications although not widely tested in x-ray mammographic images.

The B-Spline FFD [17] algorithm is based on deforming an image by modifying a mesh of control points following a maximization of a similarity measure. These control points define a mesh of smooth and continuous B-Spline functions with limited support (modifying a control point only affects neighboring points). The degree of deformation of the mesh can be modeled with the resolution of the mesh (coarse meshes are more suited for large scale transforms and finer meshes for local deformations). This also represents a trade-off between the capacity to account for finer deformations and computation time.

Polyrigid transformations [18] were proposed as a novel type of transformations in order to provide a higher degree of flexibility compared to rigid transformations but a less deformable nature as for instance found in the B-Splines formulation. They exhibit a locally rigid behavior and continuous and diffeomorphic properties by integrating the infinitesimal displacements of each rigid transformation into an ordinary differential equation formulation.

Finally, the Demon's algorithm [19] is based on viewing the registration as a diffusion process, inspired by optical flow formulation, where the diffusivity is related to the local characteristics of the image (i.e., second order derivatives). This method differs from the rest of global and local methods used in this paper by the fact that it does not minimize a global objective function (usually based in distance) as such, but it works locally following the optical flow principle.

### C. Multiresolution and Algorithm Combination

Although those methods could be used independently, it is commonly accepted that results can be improved in terms of accuracy and robustness by using a multiresolution (MR) approach or by combining different approaches [7]. The former is based on registering the images in a lower resolution, propagating parameter estimation into a higher resolution and performing registration again. This often avoids local minima in the parameter search space and reduces computational time. Algorithm combination exploits the benefits of the different methods, for instance using a global and a local method, i.e., affine registration with a B-Spline deformation. In this case, the global method recovers for main pose and scale differences and the local method accounts for localized nonlinear deformations. Following this idea, and after evaluating the registration methods individually (see Section III), we chose to evaluate BSP (which we considered had the best results) in combination with affine (AFF) registration and different MR combinations (I1: MR BSP, I2: AFF + BSP and I3: MR AFF + MR BSP).

### D. Optimizer and Interpolator

As is usual in a registration framework, parameters are recovered by optimizing a similarity measure. This optimization is guided by an objective function, usually based on the metric between images. Consequently, the two metrics presented are used in two different contexts: First, as part of the objective function to be optimized during registration process and, second, as a measure on the success of this process. This is true except for the Demons method, as has already been stated. The rest of the methods have been implemented using both distance functions, except for the polyrigid case which due to implementation limitations uses the SSD distance only.

Regarding the optimizer method, we used the gradient descent optimizer as experimental tests have proved to offer acceptable convergence times as well as enhanced reliability. The only exception is the polyrigid algorithm, where we maintained the Levenberg Marquardt optimizer as stated by its authors. In terms of interpolation, linear interpolation has been used in all cases as it provided the right trade-off between accuracy and execution time.

### E. Implementation Details

All registration methods have been implemented using the Insight Toolkit (itk) libraries [20]. The code for the polyrigid method was obtained following the instructions provided in [18]. We used 128 histogram bins and 10 000 samples for the computation of MI metric. A minimum step length stopping criteria was also used. For practical reasons, we also fixed a maximum number of iterations for all methods to a maximum of 1000 iterations for each registration. In combined or MR methods, these iterations were evenly distributed between the methods or MR levels.

These decisions, as well as the optimizer and interpolator described in Section II-D, were adopted after extensive testing as the ones that obtained the best results. The main criteria used were to keep acceptable running times (see details in Section III-G) while retaining the highest algorithm performance.

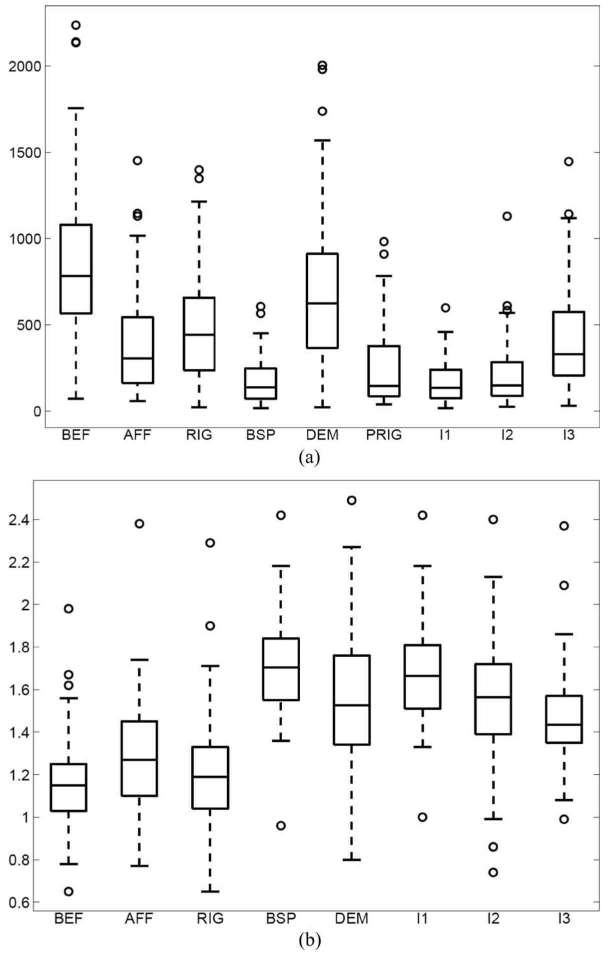
## III. Results

### A. Data Used

This section shows the evaluation results for the registration algorithms described in the previous section. The data used in this paper are a local database of 10 normal (no suspicious region was found) and 12 abnormal (a lesion in terms of a mass or architectural distortion was detected in the last screening round) patients with temporal information (images of the same breast taken at three different time intervals usually two or three years) assessed by experienced radiologists. In total, 125 cases are available (250 images for temporal comparison) in both CC and MLO views (not all patients presented both CC and MLO images). For each case, images were registered using the methods described in Section II setting the earliest mammogram as the target image and the follow up mammograms as the template images. These mammograms were originally on film and scanned using a Lumisys scanner at a resolution of 50 microns and rescaled up to 200 microns for computational purposes.

### B. Quantitative Analysis Experiments

Evaluating the results of registration methods in mammographic images is not an easy task. One could initially compute similarity metrics before and after registration to obtain an indication of how similar images are. A higher similarity is expected after image registration and the method with the highest similarity would be expected to be the most accurate, as shown in Section III-C. However, metric does not always tell the full story as sometimes images that are "closer" in terms of metric functions are perceived to be more different by human observers. In order to analyze the correlation between similarity metric and visually correct registration, we also reviewed our methods using an observer study (Section III-D), where registration results were evaluated by 11 observers with a different degree of expertise in both medical image analysis and radiology: one expert radiologist, one trainee radiologist, and nine



(a)

(b)

Fig. 3. Boxplots for metric evaluation. BEF = before registration, AFF = affine, RIG = rigid, BSP = B-Splines FFD, DEM = Demons, PRIG = polyrigid, I1 = MR BSP, I2 = AFF + BSP, I3 = MR AFF + MR BSP. (a). SSD metric. (b). MI metric.

computer vision experts with over ten years of experience (4), five to ten years (3), and less than five years (2) in mammographic image analysis. An additional evaluation criteria is also presented in Section III-E computing the TLE between salient points (e.g., microcalcifications) before and after registration. We also present results on how abnormality presence and views (CC/MLO) affect registration results. Finally, computation time results are shown in Section III-G.

### C. Metric Evaluation

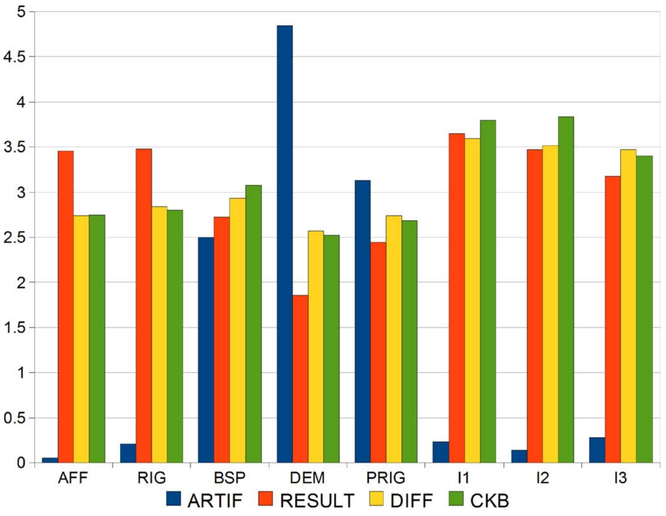


Fig. 4. Summary of the observer study. Bars show observer perception, in all cases higher means better except for the number of artifacts (ARTIF) were lower means better. AFF = affine, RIG = rigid, BSP = B-Splines FFD, DEM = Demons, PRIG = polyrigid, I1 = MR BSP, I2 = AFF + BSP, I3 = MR AFF + MR BSP.

Our assertion is that higher similarity metric means more similar images, hence, better registration. For all the images in the database we calculated SSD and MI metrics. Fig. 3 presents boxplot charts for the complete database for both metrics. The metric value (SSD or MI) used is computed after registration between the target image and the registered template images using the same metric for both registration and evaluation. This final value is related to the metric optimized by the algorithm which makes the metric comparison (SSD versus MI) difficult to assess, but nevertheless, results are useful to compare registration methods using the same metric. Through this paper, outliers were defined as observations outside interval (Q1 − 1.5IQR, Q3 + 1.5IQR, where IQR refers to interquartile range and Q1, Q3 stand for the first and third quartile, respectively).

All the methods used improved metric measurements in both distances. Concerning SSD distance, B-Splines works better among individual methods although polyrigid and affine obtain good results too. The use of MR and combination of methods generally perform better than individual ones. MR B-Splines is seen to be the best method overall. We also observe that, concerning this metric, the difference between the two best observed methods (i.e., B-Splines with and without multiresolution) is small. However, the method that used multiresolution (MR) obtained much better rating in the observation study (see Section III-D). As for the MI metric, B-Splines methods again obtain the best results and MR and method combination keep on performing generally better than individual ones. In this case, however, rigid and affine methods do not perform too well and the Demons method obtains better results.

Concerning statistical significance, we carried out several hypothesis tests (HT) to support the statements just presented. For example, the affirmation "registration reduces the error between images" was reinforced by the HT

$$H_O : \mu_{Err\_Before} = \mu_{Err\_After}$$
$$H_1 : \mu_{Err\_Before} > \mu_{Err\_After}.$$

Our assertion is validated in all tests by p-values of less than 0.01. Similar HT were carried out for affirmations of the type "BSP reduces error more than AFF"

$$H_O : \mu_{Err\_Bsplines} = \mu_{Err\_Affine}$$
$$H_1 : \mu_{Err\_Bsplines} > \mu_{Err\_Affine}$$

with p-value always less than 0.03 which shows the significance of the presented results.

### D. Observer Study

In this part of the study, we aimed at evaluating the performance of registration methods using the subjective

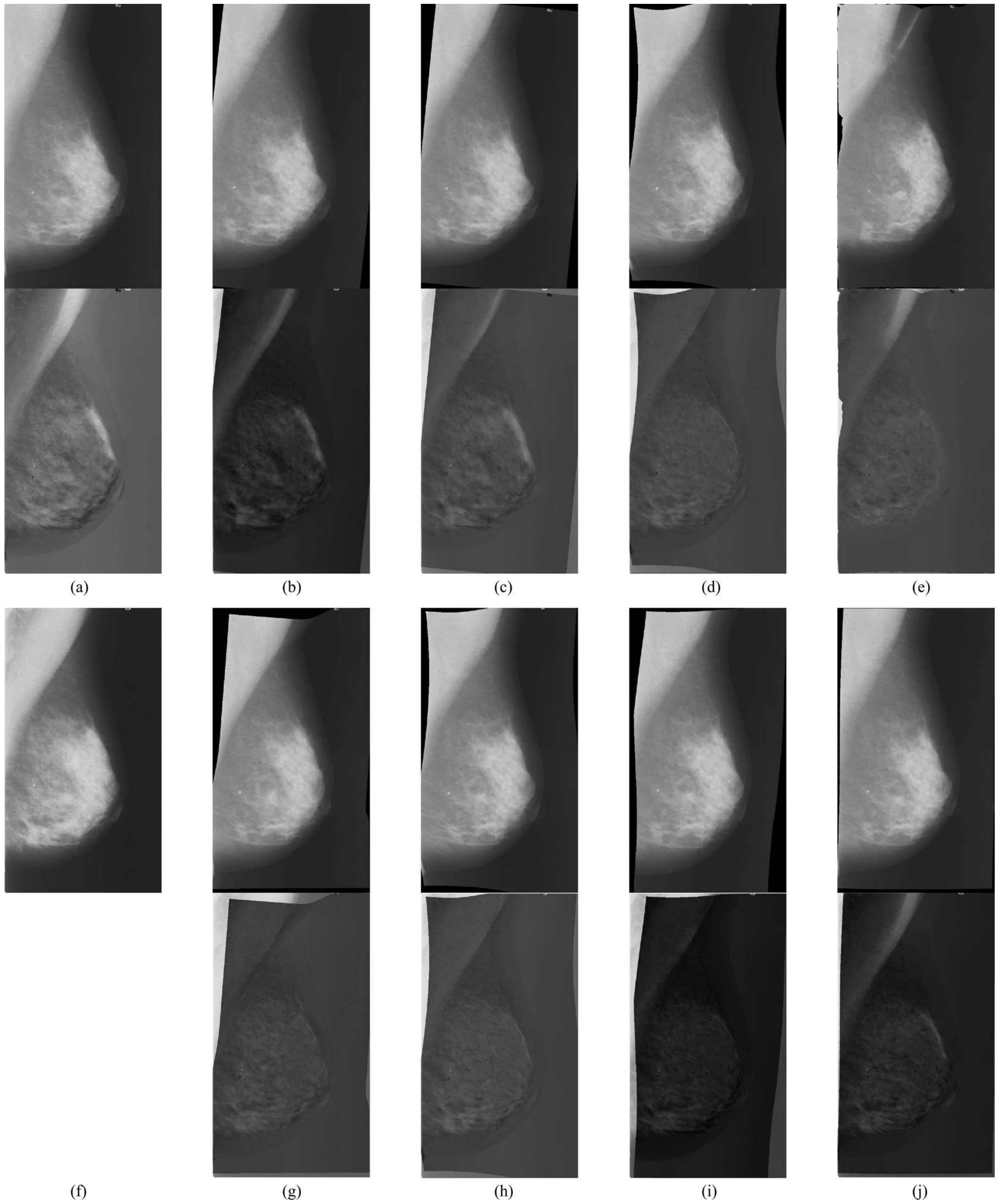


Fig. 5. Qualitative results. First row: (a) Original template and transformed template images with (b) AFF, (c) RIG, (d) BSP, (e) DEM algorithms. Second row: Difference using the (a) original and transformed template images with each respective method (b-e). Similarly, third row: (f) Target (reference) and transformed template images with (g) PRIG, (h) I1, (i) I2, (j) I3 algorithms. Fourth row: Difference using the original and transformed template images with each respective method (g-j).

perception of experts. Experts were randomly presented with mammographic images registered using the different methods. For each image, they provided a subjective integer evaluation value comprised between 1 (worst) and 5 (best) for each of the criteria considered. Scoring of each expert was standardized (using mean and standard deviation) in order to be comparable with each other. This experiment helped to find correspondences and disagreements between metric and subjective evaluation. Fig. 4 shows the mean results for all observers in the observer study.

In order to take into account as much aspects of registration as possible, we used several **criteria**:

1) *Registration artifacts (ARTIF)*: For instance, features that were not present in the template image but are present in the registered image or unrealistic deformations. The number of cases when such artifacts were detected was

counted. Results were scaled to match the 0-5 tabulation in line with the rest of criteria.

2) *Visual similarity (RESULT):* Between registered template and target images using the 1 (worst) to 5 (best) scale.
3) *Difference image (DIFF):* Allows to evaluate dissimilarities and registration performance from a global point of view.
4) *Checkerboard image (CKB):* A mosaic composed using the registered template and target images also helps to assess registration results, specially the continuity in the tissue and the smoothness of breast outlines.

An example of those criteria is observed in Fig. 5 that shows registration results with all the eight methods for a single case. Registration artifacts were evaluated in the registered image [see for instance Demons results of Fig. 5(e)], visual similarity by visualizing both target and template images, and registration accuracy by the difference and checkerboard images (checkerboard not included due to paper space limitations).

A detailed discussion on the specific results of the observation experiment is given below.

*1) Individual Methods:* As expected, rigid and affine methods produce very few artifacts. BSP method produces quite a lot (about 50%), and this percentage increases for the polyrigid method (62.5%). Demons method is perceived to produce artifacts "almost always" (96%). Registration artifacts seem to have a negative effect in the perception that experts have on the quality of registration. We will go deeper into this later. Concerning the resulting images, all methods obtain good scores except for the Demons method. Affine and rigid seem to be better. As for difference image evaluation, BSP method performs best. BSP, Demons, and polyrigid improve clearly their results from the previous criteria while affine and rigid get worse results. BSP also performs best in the checkerboard image evaluation. Only the results for the last two criteria in the observer study match the metric measurements. This illustrates the fact that metric computation does not suffice for assessing image registration. It seems that observers are greatly influenced by the presence of artifacts; even when some methods perform better according to the rest of criteria, their result is perceived as worse due to the presence of artifacts. Table I presents those results; we compare mean and variance for individual the methods in several situations for the *RESULT* criteria; all the cases, cases where artifacts are found and artifact-free cases. We observe how variances are much higher in the BSP and polyrigid methods, which are methods that sometimes present artifacts. It is clear from the results that in the presence of artifacts nonrigid methods obtain a significantly poor performance, whereas if no artifacts are produced, both BSP and polyrigid perform similarly to affine and rigid, obtaining mean values of 3.46, 3.48, 3.16, and 3.09 for affine, rigid, B-Splines, and polyrigid, respectively. Hence, a first conclusion would be that although B-Spline methods obtain good results overall, they produce a significant number of artifacts, whereas affine and rigid methods perform reasonably well taking into account their lower complexity and computation time. Multiresolution and combination methods are not included in Table I as they shown a similar trend (i.e., lower results with artifacts) but with a smaller and hence less representative number of artifacts overall, as those methods significantly decrease them, as discussed next.

TABLE I
RESULT IMAGE RATING: MEAN AND VARIANCE OBSERVER SCORES FOR INDIVIDUAL METHODS DEPENDING ON ARTIFACT PRESENCE. N.A. MEANS THAT NO IMAGES FALL INTO THE CATEGORY.

| Algorithm | All | | Artifact | | Artifact Free | |
|---|---|---|---|---|---|---|
| | Mean | Var | Mean | Var | Mean | Var |
| Affine | 3.46 | 0.69 | N.A. | N.A. | 3.46 | 0.69 |
| Rigid | 3.48 | 0.67 | N.A. | N.A. | 3.48 | 0.67 |
| B-Splines | 2.73 | 1.42 | 1.88 | 0.89 | 3.16 | 1.15 |
| Demons | 1.85 | 0.93 | 1.85 | 0.93 | N.A. | N.A. |
| Polyrigid | 2.45 | 1.26 | 2.13 | 1.16 | 3.09 | 0.86 |

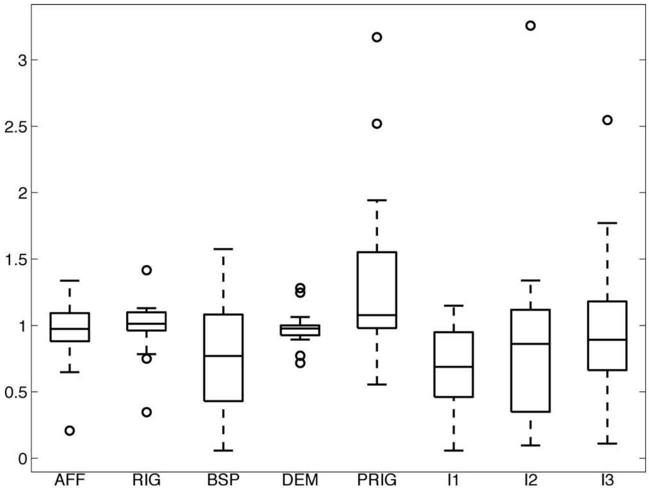
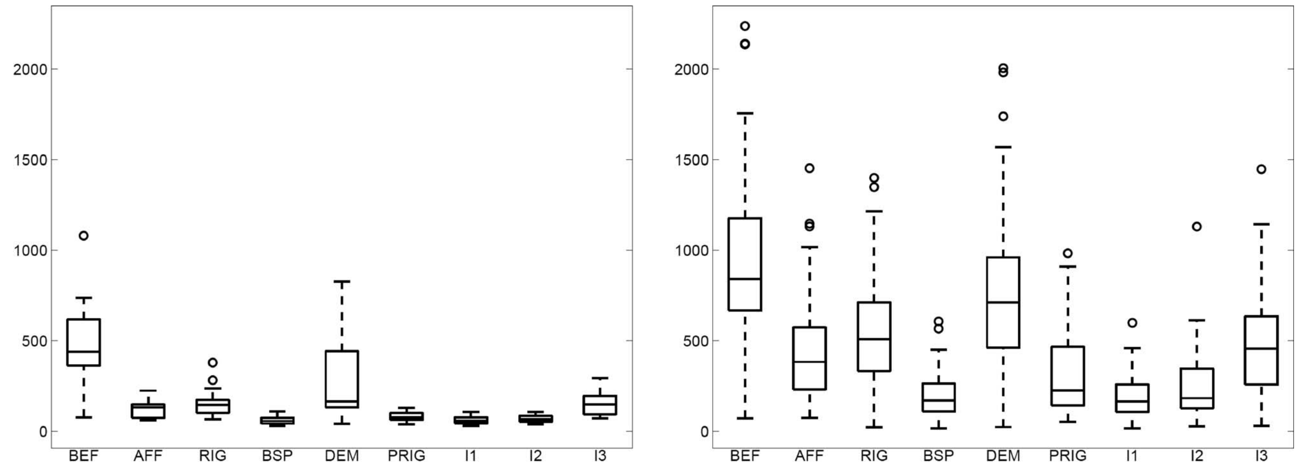


Fig. 6. Relative Microcalcifications distance. AFF = affine, RIG = rigid, BSP = B-Splines FFD, DEM = Demons, PRIG = polyrigid, I1 = MR BSP, I2 = AFF + BSP, I3 = MR AFF + MR BSP.

*2) Multi-resolution and Method Combination.:* As shown in Fig. 4, these methods reduce drastically the presence of registration artifacts and, virtually, never produce them. The rest of criteria are generally improved, although this is more visible for the difference and checkerboard tests. Overall, these methods range among the best in all criteria and MR BSP is clearly the top method overall. Surprisingly, using an affine combination did not obtain significantly better results, the same conclusion was obtained in the evaluation using metric computation. In conclusion, MR BSP (I1) obtained the best results in the observer study.

*3) Observer Variability:* We studied data for each observer separately, and for the groups (*radiologists* and *computer vision researchers*). Observer standard deviation for each observer type ranged between 0.8 and 1.2. Certain minor differences were observed (i.e., slightly better but no significative results for rigid methods given by radiologists) but general tendencies presented here were still present. In general and regarding computer vision researchers, observer agreement was almost perfect (with kappa values greater than 0.8) compared to one or the other radiologist. This is specially the case for more experienced researchers. However, and due to space limitations we do not exceed further in this respect.



(a) (b)

Fig. 7. SSD Boxplots for (a) normal and (b) abnormal cases. BEF = before registration, AFF = affine, RIG = rigid, BSP = B-Splines FFD, DEM = Demons, PRIG = Polyrigid, I1 = MR BSP, I2 = AFF + BSP, I3 = MR AFF + MR BSP.

### *E. Target Localization Error (TLE): Microcalcifications Distance*

To have a complementary view on what has been said so far we have also computed Euclidean distances between distinguished points (e.g., microcalcifications) before and after registration, in order to compute the TLE, a well-known measure in image registration [21]. A total of 18 cases from 8 patients showed visible microcalcifications. The main reason to choose microcalcifications as distinguished points is that they appear as small white regions with distinctive shapes so they are relatively easy to track manually. The last two images in Fig. 1 illustrate one of the cases we have considered where the microcalcifications are closer after registration (BSP is used in the example). When more than one lesion was present, the average distance difference for all lesions was calculated. We normalized all obtained distances by the distance before registration. Consequently, when a method obtained a final measure smaller than 1, it meant that it improved the average distance between microcalcifications. Fig. 6 presents the obtained results. Affine, rigid and Demons methods achieve small distance reductions of 5% or less on average, whereas BSP methods improve at about 20%. However, BSP is greatly affected by artifacts as image deformation might deform microcalcifications. In a very small number of cases (2 for the Demons method and 1 for B-Splines), registration artifacts make it impossible to locate the position of microcalcifications in the registered image. We have excluded these cases from this study. Using multiresolution, once more, helps to reduce the presence of artifacts and also helps to improve the results of this experiment. The MR BSP algorithm ranked better overall with an average 30% reduction of distances between microcalcifications. The polyrigid method does not perform well in this test and tends to increase the distance between microcalcifications due to misregistration. The combinations of affine and BSP with and without multiresolution get mixed results. In some cases they achieve notable improvements and in others they increase the distances. Overall they reach a small average reduction of 5–10% with a large variance. All the affirmations on this section were tested using hypothesis tests similar to those presented in Section III-C with significance values less than 0.1 which show the high degree of significance of our results.

### *F. Views and Lesions*

As described previously, the database has the particularity of incorporating temporal information of CC and MLO mammograms for both normal and abnormal patients. Results have also been evaluated taking into account the influence of the mammographic views (CC and MLO) and abnormality. In the former case, the view does not seem to have an impact on the registration results, obtaining not significative differences in the similarity metrics, observer studies, and TLE errors (graphs not included). Regarding abnormality, experiments show that error measures in the similarity metric significantly increase for abnormal cases, as shown in Fig. 7, although similar conclusions are reached regarding registration algorithm evaluation. Although SSD is only shown, MI obtained similar results. This larger error can be explained by the fact that abnormal images present a higher degree of dissimilarity due the presence of the lesion, specially in the last screening round when the abnormality was detected. This indicates that the lower degree of similarity could be used, with further analysis and processing, to develop methods for detecting normal and abnormal patients, as suggested, for instance, in [3]. However, this is not further investigated here and will be explored in future work.

### *G. Time Study*

In this section, we briefly discuss the computational cost of the registration methods evaluated, as the time needed to execute an algorithm might be a limiting factor in certain contexts. Fig. 8 presents the mean execution times for 25 registrations (represented by bars) and the standard deviation for these execution times (depicted as error bars).

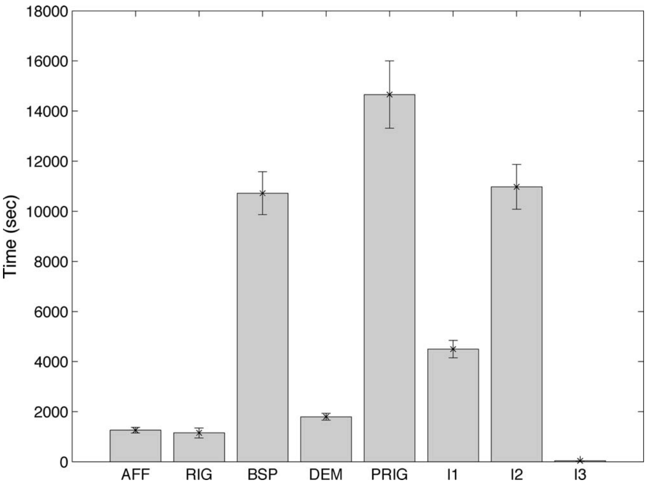


Fig. 8. Average time (bars)/standard deviation (error bars). AFF = affine, RIG = rigid, BSP = B-Splines FFD, DEM = Demons, PRIG = Polyrigid, I1 = MR BSP, I2 = AFF + BSP, I3 = MR AFF + MR BSP.

We observe how polyrigid registration is the slowest. B-Spline registration methods are also quite slow (as can be seen in the bars corresponding to the B-Spline method alone as well as the combination of affine and B-Spline registration), although their running times are greatly reduced if MR is used. Affine, rigid, and Demons methods are much faster, but not as fast as the combination of MR affine and MR BSP methods. At this point, we considered letting this method run for more iterations, but we observed how, although results improved, this improvement was not too significant compared to MR BSP so, for the sake of concretion, we present only data where all methods run the same number of iterations. The faster computation of MR AFF + MR BSP compared to MR BSP can be explained by the fact that the AFF (faster than BSP) already recovered for global misalignment, hence subsequent BSP algorithm had a faster convergence. On the other hand, BSP alone had to perform a larger number of iterations to converge.

## IV. Conclusion

We have quantitatively evaluated eight state-of-the-art registration methods for mammographic image registration using several criteria such as similarity metric computation, an observer study, TLE, and computational time. Overall, we obtained significant reductions in the metric measurements between images prior and after registration as well as positive subjective evaluation on all methods. BSP method obtained the best results from the numeric as well as the subjective point of view. This method has the problem that it produces registration artifacts (in our tests, this happened in half the cases). However, we have seen how this problem can be minimized by combining it with affine registration or by using multiresolution. In conclusion, the MR BSP method obtained the best results overall. Future work will be focused on a larger clinical validation of the methods with the aim of incorporating registration results into a computer aided detection system (CADe) using full field digital mammograms.

## Acknowledgment

The authors thank the reviewers for their comments which improved the content and readability of the paper.

## References

[1] E. D. Pisano, C. A. Gatsonis, M. J. Yaffe, R. E. Hendrick, A. N. A. Tosteson, D. G. Fryback, L. W. Bassett, J. K. Baum, E. F. Conant, R. A. Jong, M. Rebner, and C. J. D'Orsi, "American college of radiology imaging network digital mammographic imaging screening trial: Objectives and methodology," *Radiology*, vol. 236, no. 2, pp. 404–412, Aug. 2005.

[2] K. Marias, C. P. Behrenbruch, S. Parbhoo, A. Seifalian, and M. Brady, "A registration framework for the comparison of mammogram sequences," *IEEE Trans. Med. Imaging*, vol. 24, no. 6, pp. 782–790, Jun. 2005.

[3] R. Marti, D. Raba, A. Oliver, and R. Zwiggelaar, "Mammographic registration: Proposal and evaluation of a new approach," in *Digital Mammography/IWDM*, Lecture Notes in Computer Science, vol. 4046. New York, Springer, 2006, pp. 213–220.

[4] M. Tortajada, A. Oliver, Y. Díez, R. Martí, J. C. Vilanova, and J. Freixenet, "Improving a CAD system using bilateral information," *Proc. IEEE Eng. Med. Biol. Soc. Conf.*, Buenos Aires, Argentina, 2010, pp. 5054–5057.

[5] J. B. A. Maintz and M. A. Viergever, "A survey of medical image registration," *Med. Image Anal.*, vol. 2, no. 1, pp. 1–36, 1998.

[6] D. Hill, P. Batchelor, M. Holden, and D. Hawkes, "Medical image registration," *Phys. Med. Biol.*, vol. 46, 2001, pp. R1–R45.

[7] B. Zitova, "Image registration methods: A survey," *Image Vision Comput.*, vol. 21, no. 11, pp. 977–1000, Oct. 2003.

[8] Y. Guo, R. Sivaramakrishna, C.-C. Lu, J. Suri, and S. Laxminarayan, "Breast image registration techniques: A survey," *Med. Biol. Eng. Comput.*, vol. 44, no. 1, pp. 15–26, Mar. 2006.

[9] S. Kok-Wiles, M. Brady, and R. Highnam, "Comparing mammogram pairs for the detection of lesions," *Proc. 4th Int. Workshop Digital Mammography*, N. Karssemeijer, M. Thijssen, J. Hendriks, and L. van Erning, Eds. Kluwer Academic, 1998, pp. 103–110.

[10] M. Sallam and K. Bowyer, "Registration and difference analysis of corresponding mammogram images," *Med. Image Anal.*, vol. 3, no. 2, pp. 103–118, 1999.

[11] N. Vujovic and D. Brzakovic, "Establishing the correspondence between control points in pairs of mammographic images," *IEEE Trans. Image Process.*, vol. 6, no. 10, pp. 1388–1399, Oct. 1997.

[12] F. Richard, "A new image registration technique with free boundary constraints: Application to mammography," *Comput. Vision Image Understanding*, vol. 89, no. 2/3, pp. 166–196, Mar. 2003.

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

[14] Y. Díez, R. Martí, A. Oliver, and X. Lladó, "Comparison of registration methods using mamographic images," in *Proc. IEEE Int. Conf. Image Process.*, Hong Kong, 2010, pp. 4421–4424.

[15] J. P. W. Pluim, J. B. A. Maintz, and M. A. Viergever, "Mutual-information-based registration of medical images: A survey," *IEEE Trans. Med. Imaging*, vol. 22, no. 8, pp. 986–1004, Jul. 2003.

[16] D. Mattes, D. Haynor, H. Vesselle, T. Lewellen, and W. Eubank, "Pet-ct image registration in the chest using free-form deformations," *IEEE Trans. Med. Imaging*, vol. 22, no. 1, pp. 120–128, Jan. 2003.

[17] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. Hill, M. O. Leach, and D. J. Hawkes, "Nonrigid registration using free-form deformations: Application to breast MR images," *IEEE Trans. Med. Imaging*, vol. 18, no. 8, pp. 712–721, Aug. 1999.

[18] V. Arsigny, X. Pennec, and N. Ayache, "Polyrigid and polyaffine transformations: A novel geometrical tool to deal with non-rigid deformations application to the registration of histological slices," *Med. Image Anal.*, vol. 9, no. 6, pp. 507–523, Dec. 2005.

[19] J. P. Thirion, "Image matching as a diffusion process: An analogy with Maxwell's demons," *Med. Image Anal.*, vol. 2, no. 3, pp. 243–260, Sep. 1998.

[20] L. Ibanez, W. Schroeder, L. Ng, and J. Cates, *The ITK Software Guide*, 1st ed., Kitware, Inc. ISBN 1-930934-10-6, http://www.itk.org/ItkSoftwareGuide.pdf, 2003.

[21] J. M. Fitzpatrick, "The role of registration in accurate surgical guidance," *Proc. Institution Mech. Eng. Part H, J. Eng. Med.*, vol. 224, no. 5, pp. 607–622, 2010.

**Yago Díez** received the B.Sc. and Ph.D. degrees in mathematics from the Universitat Politècnica de Catalunya, Spain, in 2002 and 2008, respectively.

His research areas are computational geometry and computer vision, and his research interests include algorithms and data structures, point cloud matching, and medical image registration.

**Arnau Oliver** received the M.S. degree in physics from Universitat Autònoma de Barcelona, Spain, in 1999, and the Ph.D. degree in information technology from Universitat de Girona, Spain, in 2007.

He is with the Computer Vision and Robotics Group at the University of Girona, Girona, Spain, where he is a Lecturer. His research is mainly focused on pattern recognition and the development of automatic tools for breast cancer detection.

**Xavier Lladó** received the B.S. degree in computer science, in 1999, and the PhD degree in computer engineering, in 2004, both from the University of Girona, Girona, Spain.

Currently, he is a Lecturer at the University of Girona, Girona, Spain. His research interests are in the field of image processing and computer vision, focusing on colour and texture analysis, structure from motion, and object recognition and their applications to medical imaging focusing on mammography, prostate, and brain analysis.

**Jordi Freixenet** received the M.S. degree in computer science from the Polytechnical University of Catalonia, Barcelona, Spain, in 1994, He received the PhD in Computer Engineering for his research on object recognition in outdoor scenes from the University of Girona, in 2000.

He joined the Computer Vision and Robotics Group at the University of Girona, Spain, where currently, he is the Director of the Institute of Informatics and Applications. His research interests are in the field of computer vision, focusing on medical image analysis, object recognition, image classification, and segmentation. An important focus of his research is on improving detection and diagnosis of breast cancer.

**Joan Martí** received the B.Sc. degree in physics from Autonomous University of Barcelona, Spain, in 1986, and the Ph.D. degree in computer vision from the Polytechnical University of Catalonia, Spain, in 1998.

He is currently a Senior Lecturer at the University of Girona, Girona, Spain. His research interests include knowledge-based systems, segmentation of biomedical images, and color- and texture-based segmentation for object recognition purposes. He has chaired the 10th International Workshop on Digital Mammography (IWDM) and the 3rd Iberian Conference on Pattern Recognition and Image Analysis (IbPRIA).

**Joan Carles Vilanova** received the Medical Doctor degree, in 1988, Board in Radiology, in 1993, and the Ph.D. degree, in 2008, all from the University of Barcelona, Spain.

Currently, he is Chief of the Magnetic Resonance Imaging Department in Clinica Girona and Hospital Sta Caterina, and Lecturer at the University of Girona, Girona, Spain. His research interests are in the field of MRI, focusing on oncology imaging, brain, and musculoskeletal disease.

**Robert Martí** received the M.Sc. degree in computer science from the Universitat de Girona, Girona, Spain, in 1999, and the Ph.D. degree, in 2002, from the University of East Anglia, Norwich, U.K., for his work on image registration applied to multimodal mammography.

He is currently an Associate Professor at the Universitat de Girona, Girona, Spain. His main research interests include medical image analysis, image registration, pattern recognition, and feature extraction techniques, specially focusing on mammographic and prostatic data. He has authored more than 50 publications in international journals and conferences.