Towards computer-aided diagnostics of screening mammography using content-based image retrieval

Thomas M. Deserno, Michael Soiron
Department of Medical Informatics
RWTH Aachen University
Aachen, Germany
Web page: irma-project.org

Júlia E.E. de Oliveira, Arnaldo de A. Araújo
Department of Computer Science
Universidade Federal de Minas Gerais
Belo Horizonte, MG, Brazil
Mail to: deserno@ieee.org

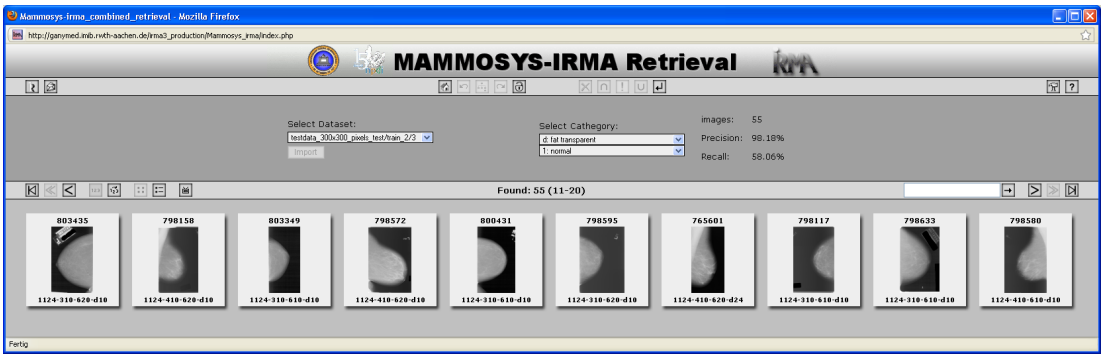


Fig. 1. Relevant images from the archive are presented by content-based image retrieval to assist computer-aided diagnosis of screening mammography.

***Abstract*—Screening mammography has been established worldwide for early detection of breast cancer, one of the main causes of death among women in occidental countries. In this paper, we aim at moving towards computer-aided diagnostics of screening mammography. Tissue and lesion are classified using the methodology of content-based image retrieval. In addition, we aim at comprehensive evaluation and have established a large database of annotated reference images (ground truth), which has been merged and unified from different sources publicly available to research.**

In total, 10,509 mammographic images have been collected from the different sources. From this, 3,375 images are provided with one and 430 radiographs with more than one chain code annotations. This data supports experiments with up to 12 classes, and 233 images per class if a equal distribution is required. Using a two-dimensional principal component analysis with four eigenvalues and a support vector machine with Gaussian kernel for feature extraction and image retrieval, respectively, the precision of computer-aided diagnosis is above 80%. It therefore may be used as second opinion in screening mammography.

***Keywords*-Content-based image retrieval; Computer-aided diagnosis; Principal component analysis; Support vector machine; Mammography; Breast lesion; Breast density**

I. Introduction

Cancer is the leading cause of death. The World Health Organization (WHO) projects deaths from cancer worldwide to continue rising, with an estimated 11 million deaths in 2030 [1]. Breast cancer is on the top five rank and one of the main causes of death among women in occidental countries (Brazilian National Cancer Institute, http://www.inca.gov.br). In the past decade, screening mammography has been established worldwide for early detection of breast cancer [2]. Using special x-ray imaging equipment, both breasts are imaged in two directions, the cranio-caudal (CC) and the medio-lateral oblique (MLO) view, and the radiographs are visually inspected by experienced radiologists.

In order to standardize mammographic reports, the Breast Imaging Reporting Data System (BI-RADS) has been established by the American College of Radiology (ACR) [3], [4]. BI-RADS define breast tissue density classes (Tab. I) as well as assessment categories (Tab. II). Despite such standardization, screening mammography is suffering until today from false positive findings, resulting in 10% unnecessary and harmful biopsies and psychological distress for many months, as well as 0.5% unnecessary treatments [4].

Computer-aided detection (CADe) has been introduced and already proven to support screening mammography [5]. However, still there are several interesting topics in cancer detection systems. In a recent review, Tang *et al.* list the following key technologies for CADe systems in mammography [6]:

- basic image enhancement,
- stochastic modeling,
- multi-scale decomposition, and
- machine learning,

which are applied for high-efficiency, high-accuracy lesion detection algorithms, including the detection of micro-

TABLE I
BI-RADS TISSUE DENSITY CLASSES.

Class	Tissue description
I	almost entirely fatty
II	scattered fibro glandular
III	heterogeneously dense
IV	extremely dense

TABLE II
BI-RADS ASSESSMENT CATEGORIES.

Class	Assessment description
0	need additional imaging evaluation and/or prior mammograms for comparison
1	no findings (negative)
2	benign
3	probably benign
4	suspicious abnormal (biopsy should be considered)
5	highly suggestive malignant
6	known biopsy-proven malignant

calcifications, masses, architectural distortion, and bilateral asymmetry. Nonetheless, CADe – by its nature – in general has a limited capacity to improve the specificity of early cancer detection.

Therefore, computer-aided diagnostics (CADx) systems currently are under development. In contrast to detection systems (CADe), where suspicious regions are marked to guide the radiologists, CADx aims at providing classification of suspicious regions, e.g., labeling the lesion according to BI-RADS assessment categories. Elter & Horsch recently have reviewed technologies successfully applied to feature extraction [7]. For clustered micro-calcifications, the following features are used:

- morphology of the cluster,
- location of the cluster,
- morphology of individual calcifications,
- optical density of individual calcifications,
- distribution of individual calcifications, and
- texture of background tissue.

In addition, features describing mammographic masses include:

- shape,
- margin characteristics,
- optical density, and
- texture.

Content-based image retrieval (CBIR) is seen as promising technology in assisting diagnosis [8], since CBIR is based on the above-mentioned visual features such as morphology, shape, and texture [9]. In general, CBIR relies on a large repository of medical images annotated with ground truth data, e.g., medical case records. The query image is presented to the system, and based on visual patterns described by features (also referred to as signature) and according distance functions, the visually most similar images are returned from the archive [10]. The physician may choose from the offered responses, and consult the electronic medical record (EMR) linked to the images for case-based reasoning (CBR) [8].

For CBIR-CAD, a second step of automatic processing is performed, where the ground truth of the retrieved images is combined with their visual similarity measures to obtain a diagnostic suggestion. So far, medical CBIR has been applied successfully for categorization of images, e.g., differing imaging modality, anatomy, field of view, or the relative positioning of patient and imaging device [11], [12]. In previous investigations, we applied the Image Retrieval in Medical Applications (IRMA) framework [13], [14] for automatic detection of BI-RADS tissue classes [15], [16].

However, CBIR has rather seldom been applied to CADx mammography. El Naqa *et al.* report a similarity learning approach to CBIR with application to digital mammography [17]. More recently, Zheng reviewed CBIR-CAD approaches in mammography [18]. The CAD performance and reliability depends on a number of factors, including the optimization of lesion segmentation, feature selection, reference database size, computational efficiency, and relationship between the clinical relevance and visual similarity.

Contributions: In this paper, we aim at moving towards CADx of screening mammography. We propose, implement, and evaluate a CBIR-CADx system addressing feature selection, computational efficiency, and reference database size. Tissue and lesion classes are classified using CBIR methodology. In addition, we aim at comprehensive evaluation and, accordingly, we have established a large database of annotated reference images (ground truth), which has been merged and unified from different sources publicly available to research.

II. STATE OF THE ART IN CBIR-BASED CAD FOR MAMMOGRAPHY

In the context of screening mammography, CBIR and CAD systems have been explored to improve knowledge and provide facilities for the radiologists. Both, tissue classes and assessment categories have been analyzed.

A. Breast density classification

Bovis & Singh considered both, all four BI-RADS categories as well as the two categories: fatty and dense [19]. Using 377 mammograms from Digital Database for Screening Mammography (DDSM), four groups of texture features were extracted. To reduce the dimensionality of the data, the principal component analysis (PCA) was used. The combination paradigm was applied by eleven component classifiers, and the authors have chosen an artificial neural network (ANN) for this task. Recognition rates of $40-71\%$ and $77-97\%$ were reported for the four and the two class problems, respectively.

Oliver *et al.* proposed a statistical technique to perform a segmentation of the breast density, which was divided into only two classes, fatty and dense [20]. The breast density was characterized using two approaches, principal component analysis (PCA) and linear discriminant analysis (LDA), in which each pixel of a new mammogram was classified as fatty or dense, taking into account its neighborhood. From the Hospital Josep Trueta, Spain, 54 regions of interest (ROI) of size 50×50 pixels were extracted, and from the Mammographic Image Analysis Society (MIAS) digital mammogram database

TABLE III
DATA DISTRIBUTION IN THE UNIFIED REFERENCE DATABASE.

Author	Year	Problem	Classes	Features	Classifier	Images	Source	Results
Bovis & Singh	2002	Tissue	2 4	ROI	ANN	377	DDSM	40 – 71% 77 – 97%
Oliver *et al.*	2009	Tissue	2	ROI	LDA/PCA	54 322	private MIAS	90%
Tagliafico *et al.*	2009	Tissue	4	Histogram	Thresholding	160	private	80 – 90%
Subashini *et al.*	2010	Tissue	3	ROI	SVM radial kernel	43	private	95%
Eltonsy *et al.*	2007	Lesion	2	ROI	Concentric layer model	540	DDSM	96%
Elter & Hasslmeyer	2008	Lesion	2	ROI, meta	Generic algorithm, Euclidean metric	360	DDSM	86% (ROC)
Tao *et al.*	2010	Lesion	1	ROI	Multi-phase pixel-level	54	private	69%
Verma *et al.*	2010	Lesion	2	ROI	Soft-clustered direct learning	200	DDSM	97%
Oliver *et al.*	2010	Both	4	ROI	LDA / PCA	184	private	92 – 94%

all the 322 images available were used. Despite the authors expectation that LDA would perform superior, PCA provided the best classification with about 90% of accuracy. However, no statistically significant differences were found.

Tagliafico *et al.* developed a method to estimate the breast density [21] based on gray scale statistics. Histogram and accumulative histograms are computed and thresholded. Image data was acquired according to the Mammography Quality Standards Act (MQSA) and digitized from the films. In total, 160 mammograms from both CC and MLO views were used wherein breast density had been evaluated according to the four BI-RADS tissue classes by two experienced radiologists. Results from 80% to 90% of agreement between the two experienced radiologists and the automatic program were reported.

Subashini *et al.* used statistical features for breast characterization [22]. In total, 43 normal mammograms from MIAS database were used and the pectoral muscle as well as labels were eliminated from the mammograms. Then, from the breast region alone, features like mean, standard deviation, smoothness, third moment, uniformity, entropy, and others were extracted and classification was performed using a support vector machine (SVM) with a radial basis kernel. The results report an accuracy of 95.44% considering three classes of breast density, namely fatty, glandular, and dense tissue.

B. Assessment categories

Eltonsy *et al.* presented a technique for automatic detection of masses using the DDSM database [23]. Morphological characteristics were extracted from 540 ROIs of images containing malignant masses (270 images), randomly selected from the DDSM cancer volumes with the same amount of negative findings, and split 50% for test and training, each. The authors have reported that 96% of the masses have been detected by the proposed concentric layers model, and only 12 masses have been failed of detection.

Elter & Hasslmeyer proposed a CADx system to differ malignant from benign masses [24]. From the DDSM database, 360 ROIs were selected, 164 and 196 containing a proven malignant and a benign mass, respectively. In total, 75 features were extracted including some meta-information such as gender and age. To escape the *curse of dimensionality*, a genetic algorithm was applied reducing the feature space as well as finding suitable weights. An Euclidean distance was applied as metric. The performance was evaluated in terms of the area under the receiver operating characteristic (ROC) curve, and 86% was obtained.

Aiming the segmentation of breast masses in mammograms, Tao *et al.* proposed an algorithm involving multi-phase pixel-level classification [25]. A non-specified dataset provided 54 mammograms (51 malignant and 3 benign) from which ROIs were verified and had the ground truth set by five experienced radiologists. The following features were extracted: texture (from co-occurrence matrix), shape (Hessian matrix), and a group of gray-level statistics, in a total of 30 dimensional features for each ROI. After this, in order to obtain the potential mass pixels, the Otsu thresholding method was used, and for the spiculation detection, a steerable edge detection approach was employed. Results showed an accuracy of 87.81% for the pixel-scale classification task and an area overlapping ratio of 68.90% and 54% for segmenting entire mass and margin portion only, respectively.

Verma *et al.* proposed a new algorithm for breast lesions classification [26]. From DDSM database, 200 mammograms had the ROI containing the lesion extracted and then characterized using the following features: density, mass shape, mass margin, abnormality assessment rank, patient age, and subtlety value. A soft-clustered direct learning classifier was proposed for the classification of suspicious areas and over 97% of accuracy was obtained, a high value when compared with other techniques such as standard multi-layer perceptron, for instance.

Combining CBIR and CAD, Tao *et al.* established a mass retrieval platform [27] with the performance of two experiments. The first one used a database consisting of 415 mammograms containing masses (244 malignant and 171 benign) from University of Michigan, and the second one used a reference library non identified containing 476 masses (219 malignant and 257 benign). Prior to feature extraction, a segmentation of the masses was performed using the multi-level learning-based segmentation method [25]. Regular, lobulated, and irregular shapes were computed with the following features: third order moments, curvature scale space descriptors (CCSD), radial length statistics, and region-based shape. Regarding mass margin, they considered circumscribed, microlobulated, indistinct, and spiculated, extracting texture features from them.

Similarity was calculated using a locally linear embedding (LLE) method with the performance indicated by a ROC analysis. For the first database, an area under the ROC curve of 75% was obtained, and for the second dataset, an average area of 80% was achieved.

C. Tissue and assessment combined

Continuing their previous work, Oliver *et al.* have used 92 mammograms of MLO and CC views each to verify if the breast density information improves the results of a CAD system for breast masses detection [28]. These 184 mammograms were obtained with a digital mammographic unit (Mammomat Novation, Siemens AG, Munich, Germany). The reason for combining MLO and CC views is that sometimes a lesion is hidden and only seen in one of the views. Texture attributes were used to characterize each BI-RADS tissue type and results showed that although a better performance was obtained for MLO mammograms, the performance of the CAD system is improved and this information is advantageous. An accuracy of 92% was obtained without considering breast density information and one of 94% using automatic density estimation. However, the authors do not report on false negative findings.

D. Resumee

Table III summarizes previous publications. It becomes obvious, that all experiments have been performed on different, rather small, and also selective datasets. Selection procedures as well as separation of test and training data are ambiguous, in particular with respect to the small sample sizes. Furthermore, most authors use private data, not being available for others. Results are reported by precision, accuracy, or ROC curves, which also are incomparable quantitatively. Another problem is hidden in the small number of categories, disregarding whether tissue, assessment, or mixed classes are considered. In most papers, a simple two-class problem is investigated, which reflects clinical practice insufficiently. Recently, this need of unified data repositories and evaluation study protocols has also been claimed by Horsch *et al.* [29].

III. Material & Methods

A. Development of reference database

In our previous work, mammographic images have been unified and imported from several sources available to the public [30]. In total, 10,509 images have been made available from the following repositories:

- Mammographic Image Analysis Society (MIAS)
- Digital Database for Screening Mammography (DDSM)
- Lawrence Livermore National Laboratory (LLNL)
- Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, Aachen, Germany Department of Radiology.

All radiographs are classified reliably according to the BI-RADS tissue classes (Tab. I) and assessment categories (Tab. II) as well as the type of lesion (Tab. IV).

TABLE IV
IRMA lesion classes [30].

Class	Type of Lesion
a	unspecified
b	calcification, unspecified
c	calcification, micro-sized
d	calcification, macro-sized
e	mass, unspecified
f	mass, circumscribed
g	mass, spiculated
h	architectural distortion
i	asymmetry

To create a ground truth for diagnostics, the IRMA mammography database was extended by additionally importing spatial annotations that were (partly) delivered with the images. Figure 2 depicts the conversion of annotations, which are provided as center coordinate and size (circle), masking image (overlay), or polygon/free-hand annotation (drawing), and transformed into a chain code definition in the extensible markup language (XML). Within the IRMA database, many-to-one standardized annotations are linked to the images, respectively. The IRMA framework intrinsically supports consistent scaling of both, images and chain-coded annotations.

Centered to the annotated lesion outline, quadratic patches are extracted at different scales and size (128 pixel with respect to a full scale of 1,024 pixel bounding box) and used within the experiments. For BI-RADS assessment class 1 (negative), we extracted the patches from the center of the breast tissue, which can be computed automatically combining contour with muscle detection in both CC and MLO images.

B. Feature extraction and distance measure

Breast tissue of different density with or without a lesion may differ in brightness and texture attributes, since it contains information about the spatial distribution and variation of gray levels. Since high-dimensional feature vectors may limit computational efficiency and accuracy (*curse of dimensionality*), a technique that combines the representation of texture with the reduction of dimensionality is desirable. The two-dimensional (2D) principal component analysis (2DPCA) technique is able

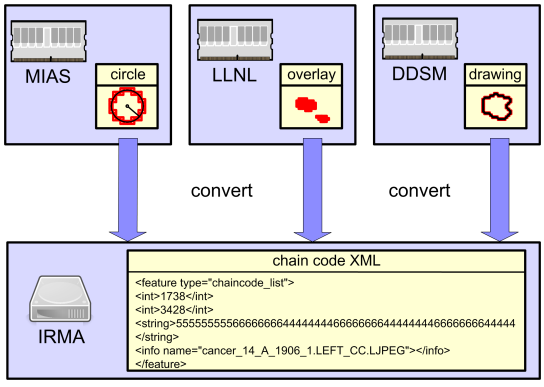


Fig. 2. Conversion of annotations

to satisfy these requirements. Unlike PCA, where the principal component is a scalar, with 2DPCA each principal component is a vector [31]. Previous experiments have shown that $d = 4$ of such vectors, each having m components, are sufficient to characterize the tissue patterns [15].

For a binary classification, the support vector machine (SVM) can be described as follows: given two classes and a set of points that belong to these classes, the SVM determines the hyperplane in the feature space that separates the points in order to place the highest number of points of the same class on the same side, while maximizing the distance of each class to that hyperplane. The hyperplane generated is determined by a subset of items from the two classes, called support vectors.

For more than two classes, the problem turns into a multi-class problem, which opens a variety of solutions [32]. According to previous investigations, we apply the one-against-one method [15], where an SVM is built for a pair of classes through its training in the discrimination of two classes. In this way, the number of SVMs used in the method is $M(M-1)/2$, where M is denoting the number of classes. Each SVM belongs to only the two corresponding classes.

During the training phase, each class is matched against all other classes. The obtained parameters are stored in a matrix-like model file (Fig. 3, bottom).

Figure 3 (top) depicts how the prediction method is yielded from the one-against-one solutions. Corresponding steps are indicated by respective colors. For each step, all of which being two-class problems, the site of the hyperplane is determined on which the feature is lying. All hyperplanes are delivered by the respective models ("C" in Fig. 3). This way, only M steps are needed for classification.

C. CADx and CAD experiments with CBIR

The experiments were performed in several consecutive steps, which are visualized in Figure 4:

1) *ROI location:* Conserving their aspect ratio, all mammographies have been reduced in pixel resolution to fit a bounding box of 1024×1024 pixel, and 128×128 pixel patches where extracted centering the lesion that is indicated by a chain code in the ground truth data. For BI-RADS category 1, the patches were positioned arbitrarily within the breast tissue, excluding the pectoral muscle, nipple and background areas.

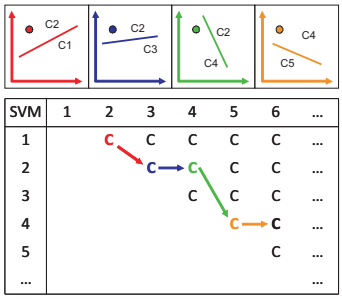


Fig. 3. Process of classification in the SVM model.

2) *Feature extraction:* The 2DPCA method was performed on all the patches. The principal components related to the first d largest eigenvalues of the covariance matrix were used in the experiments.
3) *Spit & Scale:* The data is divided into 60% for training and 40% for testing. To improve the performance of the SVM classifiers, the training features are scaled into the interval $[-1, \ldots, 1]$, and the scaling factor is used to normalize the testing features, too.
4) *First Model:* Using the LIBSVM library, 5-fold cross validation was performed on the 60% training data in order to obtain the best parameters of the SVM, which then are used to build the first model. All hyperplanes are stored in a matrix according to Figure 3.
5) *Selection:* This first SVM model is used to indicate the relevance of images to a certain query (classifier experiment). In a final system, the selection can be done using a relevance feedback loop. Since this has not jet been implemented, we applied the ground truth information so far.
6) *Second Model:* It has been observed that radiologists, when performing CBIR experiments, pay most attention on the top ten images retrieved. Hence, the top ten retrieved images were used for training the second SVM in order to obtain the final model.
7) *Retrieval:* Using the second model, the query image finally is classified. For quantitative experiments, the classification of the query image is compared to the ground truth.

The first $d = 4$ principal components of 2DPCA technique were considered for lesion (malignant or benign) characterization and SVM was used for classification with 10-fold cross validation of all the patches. The evaluation was performed with measures of accuracy, which is the percentage of correctly classified images over the ground truth of all images in that category (normal or malignant lesion and normal or benign lesion).

To increase the comparability with results published by others (Tab. III), the performance of the 2DPCA technique was compared to PCA and SVD for breast and lesions characterization, as these two techniques were already reported as being able to represent texture and reduce the dimensionality of the feature vector. SVM was evaluated for the task of image retrieval.

D. Implementation

Our system was implemented using MatLab through the image processing toolboxes and the LIBSVM library [33]. Feature extraction was executed on an IntelCore2Quad 2.66 GHz processor with 8 GB of RAM operated with Microsoft Windows 64 bit system. Image retrieval was performed on an Intel-Core2Duo 2 GHz processor with 3 GB of RAM under Microsoft Windows 32 bit.

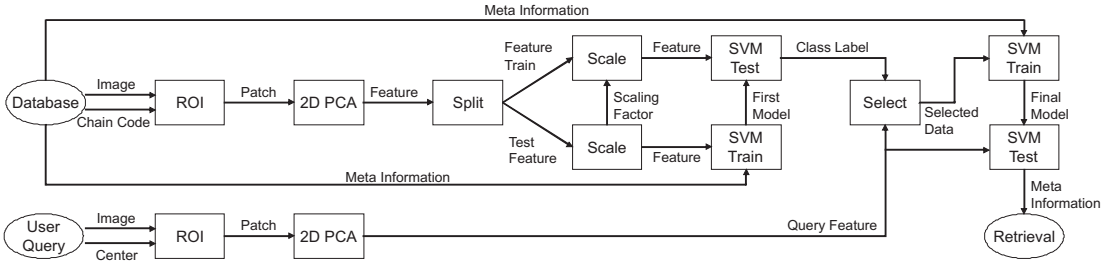


Fig. 4. Flowchart of the algorithm.

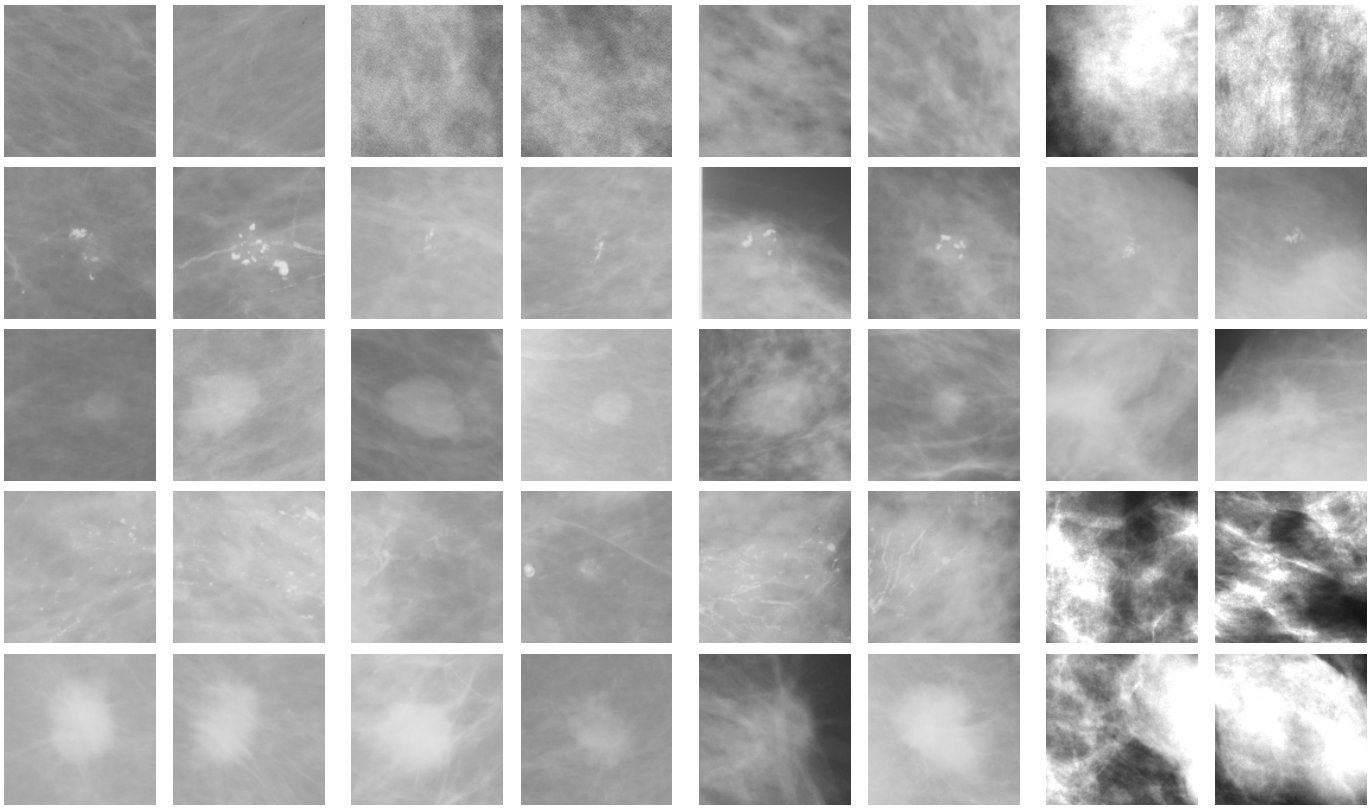


Fig. 5. Patches from the IRMA mammography reference database

IV. Results

A. Database and annotation

From the IRMA database, a total of 9,870 mammographic images are available according to the relevant BI-RADS codes (Tab. V). From this, 3,375 images are provided with one and 430 radiographs with more than one chain code annotations. The smallest sample size of 233 images is recorded with BI-RADS IV-5 (Tab. V). Since our implementation is based on the LIBSVM library, equal-sized classes are needed so far, and 233 images were selected randomly from each class in the database, resulting in 2,796 individual images of 12 classes.

Figure 5 exemplifies the resulting patches. Each two horizontally neighbored patches are from the same class. The four double-columns correspond to the BI-RADS tissue density classes I to IV. From top to bottom, the five rows depicts BI-RADS assessment categories 1, 2, 2, 5, 5, respectively, where rows 2, 4 and 3, 5 show examples from IRMA lesion classes (b-d) calcification and (e-g) mass, respectively.

This figure impressively demonstrate the classification problem to handle. On the one hand side, one may feel to see the

TABLE V
Data distribution in the unified reference database.

BI-RADS		Tissue classes I	II	III	IV	Sum
Assessment categories	1	2,518	1,855	1,295	834	6,501
	2	676	515	383	237	1,811
	5	591	471	263	233	1,558
	Sum	3,785	2,841	1,940	1,304	9,870

TABLE VI
RESULTS MATRIX OBTAINED WITH $d = 4$.

Feature extraction	SVM kernel shape polynomial	sigmoidal	Gaussian
2DPCA	72.31%	67.84%	**80.07%**
PCA	66.69%	66.61%	68.83%
SVD	66.32%	**60.56%**	69.96%

differences between malignant and benign masses, or between the patches from the BI-RADS I-IV tissue classes of negative findings (BI-RADS 1, first row), but, for instance, deciding malignant or benign calcification seems rather difficult. It is worth mention that neither mean gray scale nor contrast is a sufficient feature for robust tissue density nor assessment category classification.

B. Classification

Table VI lists the best average precision obtained with $d = 4$, comparing breast density and lesion characteristics using 2DPCA, PCA, and SVD for feature extraction and SVM with polynomial, Gaussian, and sigmoidal kernels for the retrieval task. In general, polynomial kernels perform better than sigmoidal kernels, and Gaussian kernel outperform both, polynomial and sigmoidal kernels. With respect to the feature extraction and selection methods, 2DPCA performs superior to SVD, and PCA is worst. The overall best result of 80,07% is obtained for 2DPCA with a Gaussian kernel, which is in consonance with previous experiments [15]. Worst result is only 60,56% of averaged precision, which is obtained with SVD and a sigmodial kernel of the SVM classifier.

The time of execution of the CBIR system was 6,200 seconds using the polynomial kernel, 2 seconds using the sigmoidal kernel and 4.3 seconds using the Gaussian kernel. As a CBIR system that takes several minutes to execute the retrieval process is not viable for clinical practice of radiologists, the use of polynomial kernels cannot be recommended.

Figure 1 depicts ten relevant images from the archive presented to the radiologist by CBIR to assist CAD of screening mammography. With the validated meta-information linked to these images, this set can be regarded as a second reading, providing additional confidence. So far, the physician is enabled to switch the view of the images between the thumbnail (Fig. 6a), thumbnail and chain code fused (Fig. 6b), and patch (Fig. 6c). The corresponding image is shown in a separate browser window when the thumbnail is hit with the mouse. Here, a link to the patient's EMR could be placed, too.

V. DISCUSSION

Table VII shows the confusion matrix of the 12 class experiments. The matrix is sparsely occupied, which indicates that some paths of the one-against-one classification process (Fig. 3) are less likely than others. Unsurprisingly, it appears that classes within the same category of pathology are harder to differentiate from each other than from different categories. Furthermore, categories of more translucent tissue are easier to classify. This is compliant with findings by other authors.

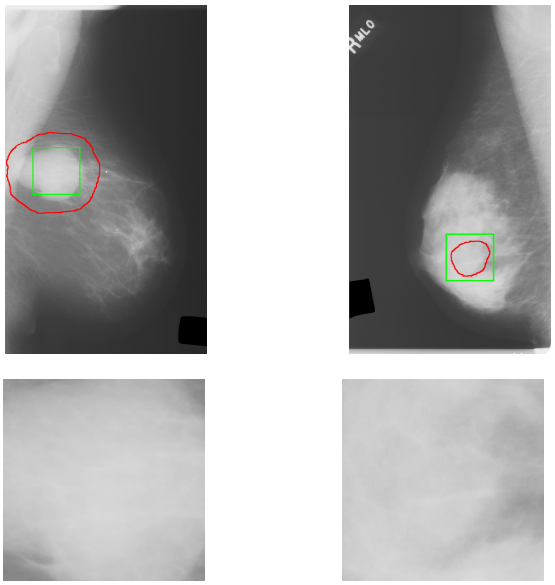
TABLE VII
CONFUSION MATRIX. CRITICAL ERRORS IN BOLDFACE.

	I-1	**I-2**	**I-5**	**II-1**	**II-2**	**II-5**	**III-1**	**III-2**	**III-5**	**IV-1**	**IV-2**	**IV-5**
I-1	89	0	0	0	0	0	3	0	0	2	0	**1**
I-2	0	80	0	0	0	0	0	0	0	0	7	**7**
I-5	0	0	62	0	**6**	4	0	**7**	11	0	0	4
II-1	0	0	0	78	0	0	0	0	0	0	0	**16**
II-2	0	0	**10**	0	60	11	0	7	**6**	0	0	0
II-5	0	0	24	0	**13**	34	0	**5**	0	**18**	0	0
III-1	15	0	0	0	0	0	40	0	0	38	0	1
III-2	0	0	**19**	0	5	**9**	0	48	**12**	0	60	**1**
III-5	0	0	19	0	20	14	0	**9**	32	0	0	0
IV-1	16	0	0	0	0	0	20	0	0	58	0	0
IV-2	0	6	0	0	0	0	0	0	0	0	66	**22**
IV-5	0	**7**	0	0	0	0	0	0	0	0	**39**	48

For dense tissue (BI-RADS class IV), most critical errors occur (Tab. VII, highlighted in red). Here, benign (BI-RADS category 2) is confused with malignant (BI-RADS category 5) and vice versa. Again, this finding has been reported in the literature, too.

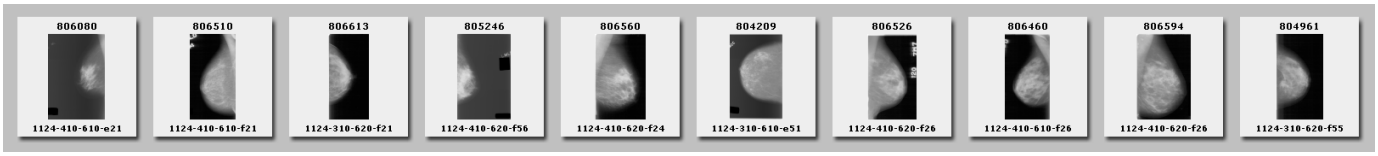
Rather surprisingly, another medical meaningful confusion has occurred frequently (19 times) between BI-RADS III-2 and I-5. This can be explained however due to the fact that, currently, patched of fixed resolution-related size have been extracted from the mammographies. If the patch is too small with respect to the actual lesion, it contains only tumorous tissue, and this may be erroneously classified as heterogeneously dense tissue (Fig. 7). Therefore, adapting the patch size to the size of lesions might improve the retrieval results. It is worth mention that this information is captured in the compiled ground truth database. Also, separating MLO from CC images might further improve the results.

The results presented in this paper have been obtained on a by far increased database as compared to previously published

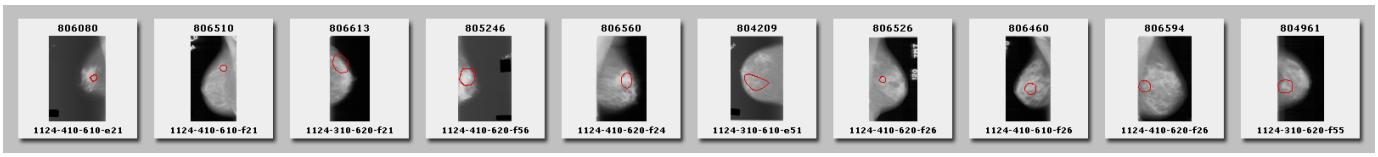


(c) BI-RADS I-5 (d) BI-RADS III-2

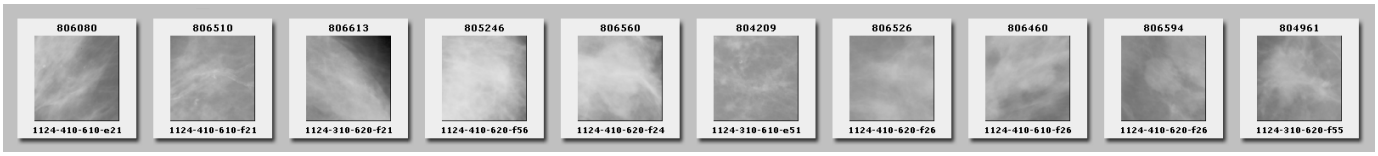
Fig. 7. Extracting patches of fixed size.



(a) Normal view



(b) Chain code overlay



(c) Extracted patch

Fig. 6. View options from the CBIR result browser

work in terms of both, the number of classes and the number of samples. Handling different sample sizes per class and extracting patches from those images in the database, which are associated with more that one annotation, will further increase the number of patches for even more comprehensive experiments. Then, however, overlaps may occur and must be handled, as well as linear dependencies, accordingly.

Further test have been performed using 20 classes, where – supported by the ground truth that we have established in the database – the type of lesion is differentiated between calcifications and masses. Since both patterns differ visually, one may expect an improvement that might be able to compensate the performance loss due to more classes. However, the SVM library in use is forcing an equal class distribution, which further reduces the data to a total of 900 images; 45 images are in the smallest of the 20 classes. Hence, the limitation of equal-sized classes must be overcome before further experiment become generalizable.

VI. Conclusion

In this study, a CBIR system was presented that uses breast density together with the existence of a breast lesion as pattern for image retrieval. We have continued comprehensive system evaluation based on a significantly enlarged database of, so far, 3,375 images of 12 classes. This work may contribute to CBIR-CAD of mammographies, providing a system able to aid radiologists in their diagnosis or a system that is useful as pre-processing stage for computer-aided systems for breast lesions classification. Future work will include the characterization of breast lesions individually, through morphological features. Also, combining CC and MLO views for classification is expected to increase the system's reliability. We also plan to conduct experiments with even more classes of pathology, e.g., separating calcification from masses. Nonetheless, the precision obtained ($> 80\%$) already indicates that CADx has the potential to significantly reduce the number of unnecessary breast biopsies in clinical practice.

Acknowledgment

This work is supported by CAPES and CNPq, the Brazilian research funding agencies, and the Federal Ministry of Education and Research Germany (DLR-BRA 009/045). The IRMA project has been funded by the German Research Foundation (DFG), Le 1108/4 and Le 1108/9.

References

[1] WHO, "Cancer," World Health Organization Fact Sheet No. 297, available at http://www.who.int/mediacentre/factsheets/fs297/en/, 2011.

[2] P. C. Gotzsche and M. Nielsen, "Screening for breast cancer with mammography," *Cochrane Database Syst Rev*, vol. 7, p. CD001877, 2009.

[3] ACR, "American college of radiology: Illustrated breast imaging reporting and data system (bi-rads)," Reston, VA: ACR, 4th edition, 2003.

[4] S. Obenauer, K. P. Hermann, and E. Grabbe, "Applications and literature review of the bi-rads classification," *Eur Radiol*, vol. 15, pp. 1027–1036, 2005.

[5] R. L. Birdwell, "The preponderance of evidence supports computer-aided detection for screening mammography," *Radiology*, vol. 253, pp. 9–16, 2009.

[6] J. Tang, R. M. Rangayyan, J. Xu, I. E. Naqa, and Y. Yang, "Computer-aided detection and diagnosis of breast cancer with mammography: recent advances," *IEEE Trans Inf Technol Biomed*, vol. 13, pp. 236–251, 2009.

[7] M. Elter and A. Horsch, "CADx of mammographic masses and clustered micro-calcifications: a review," *Med Phys*, vol. 36, pp. 2052–2068, 2009.

[8] H. Müller, N. Michoux, D. Bandon, and A. Geissbuhler, "A review of content-based image retrieval systems in medical applications-clinical benefits and future directions," *Int J Med Inform*, vol. 73, pp. 1–23, 2004.

[9] T. M. Lehmann, M. O. Güld, T. Deselaers, D. Keysers, H. Schubert, K. Spitzer, H. Ney, and B. B. Wein, "Automatic categorization of medical images for content-based retrieval and data mining," *Comput Med Imaging Graph*, vol. 29, pp. 143–155, 2005.

[10] H. D. Tagare, C. C. Jaffe, and J. Duncan, "Medical image data-bases: a content-based retrieval approach," *J Am Med Inform Assoc*, vol. 4, pp. 184–198, 1997.

[11] T. Deselaers, H. Müller, P. Clough, H. Ney, and T. M. Lehmann, "The clef 2005 medical annotation task," *Int J Computer Vis*, vol. 74, pp. 51–58, 2007.

[12] T. Deselaers, T. M. Deserno, and H. Müller, "Automatical medical image annotation in imageclef 2007: overview, results, and discussion," *Pattern Recognit Lett*, vol. 29, pp. 1988–1995, 2008.

[13] T. M. Lehmann, M. O. Güld, C. Thies, B. Fischer, K. Spitzer, D. Keysers, H. Ney, M. Kohne, H. Schubert, and B. B. Wein, "Content-based image retrieval in medical applications," *Methods Inf Med*, vol. 43, pp. 354–361, 2004.

[14] M. O. Güld, C. Thies, B. Fischer, and T. M. Lehmann, "A generic concept for the implementation of medical image retrieval systems," *Int J Med Inform*, vol. 76, pp. 252–259, 2007.

[15] J. E. E. de Oliveira, A. M. C. Machado, G. C. Chavez, A. P. B. Lopes, T. M. Deserno, and A. de A. Araujo, "MammoSys: a content-based image retrieval system using breast density patterns," *Comput Methods Programs Biomed*, vol. 99, pp. 289–297, 2010.

[16] J. E. E. de Oliveira, A. de A. Araujo, and T. M. Deserno, "Content-based image retrieval applied to BI-RADS tissue classification in screening mammography," *World J Radiol*, vol. 3, no. 1, pp. 24–31, 2011.

[17] I. El-Naqa, Y. Yang, N. P. Galatsanos, R. M. Nishikawa, and M. N. Wernick, "A similarity learning approach to content-based image retrieval: application to digital mammography," *IEEE Trans Med Imaging*, vol. 23, pp. 1233–1244, 2004.

[18] B. Zheng, "Computer-aided diagnosis in mammography using content-based image retrieval approaches: current status and future perspectives," *Algorithms*, vol. 2, pp. 828–849, 2009.

[19] K. Bovis and S. Singh, "Classification of mammographic breast density using a combined classifier paradigm," *Proc MIUA (Medical Image Understanding and Analysis)*, 2002.

[20] A. Oliver, X. Lladó, E. Pérez, J. Pont, E. Denton, J. Freixenet, and J. Marti, "A statistical approach for breast density segmentation," *Journal of Digital Imaging*, vol. 23, pp. 55–65, 2009.

[21] A. Tagliafico, S. Tagliafico, F. Tosto, C. Chiesa, L. E. Martinoli, and M. Calabrese, "Mammographic density estimation: comparison among birads categories, a semi–automated software and a fully automated one," *The Breast*, vol. 18, pp. 35–40, 2009.

[22] T. S. Subashini, V. Ramalingam, and S. Palanivel, "Automated assessment of breast tissue density in digital mammograms," *Comput Vis Image Underst*, vol. 114, no. 1, pp. 33–43, 2010.

[23] N. H. Eltonsy, G. D. Tourassi, and A. S. Elmaghraby, "A con-centric morphology model for the detection of masses in mammography," *IEEE Trans on Med Imag*, vol. 26, no. 527–537, pp. 880–889, 2007.

[24] M. Elter and E. Hasslmeyer, "A knowledge-based approach to CADx of mammographic masses," *Proc SPIE*, vol. 6915, p. 0L, 2008.

[25] Y. Tao, S.-C. B. Lo, M. T. Freedman, E. Makariou, and J. Xuan, "Multilevel learning-based segmentation of ill-defined and spiculated masses in mammograms," *Med Phys*, vol. 37, no. 11, p. 5993, 2010.

[26] B. Verma, P. McLeod, and A. Klevansky, "Classification of benign and malignant patterns in digital mammograms for the diagnosis of breast cancer," *Expert Syst Appl*, vol. 37, no. 4, pp. 3344–3351, 2010.

[27] Y. Tao, S.-C. B. Lo, L. Hadjiski, H.-P. Chan, and M. T. Freedman, "Birads guided mammographic mass retrieval," *Proc SPIE*, vol. 7963, no. 70632, 2011.

[28] A. Oliver, X. Lladó, J. Freixenet, R. Martí, E. Pérez, J. Pont, and R. Zwiggelaar, "Influence of using manual or automatic breast density information in a mass detection cad system," *Acad Radiol*, vol. 17, no. 7, pp. 877–883, 2010.

[29] A. Horsch, A. Hapfelmeier, and M. Elter, "Needs assessment for next generation computer-aided mammography reference image databases and evaluation studies," *Int J Comput Assist Radiol Surg*, pp. 1–19, 2011.

[30] J. E. E. de Oliveira, M. O. Guld, A. de A. Araujo, B. Ott, and T. M. Deserno, "Towards a standard reference database for computer-aided mammography," *Proc SPIE*, vol. 6915, p. 69, 2008.

[31] J. Yang, D. Zhang, A. F. Frangi, and J. Yang, "Two-dimensional pca: a new approach to appearance-based face representation and recognition," *IEEE Trans Pattern Anal Mach Intell*, vol. 26, pp. 131–137, 2004.

[32] C. W. Hsu and C. J. Lin, "A comparison of methods for multi-class support vector machines," *IEEE Trans Neural Netw*, vol. 13, pp. 415–425, 2002.

[33] C. C. Chang and C. J. Lin, "LIBSVM: a library for support vector machines," Software available at http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.