SVM-based CBIR of Breast Masses on Mammograms

Lazaros Tsochatzidis, Konstantinos Zagoris,
Michalis Savelonas, and Ioannis Pratikakis

Visual Computing Group,
Dept. of Electrical and Computer Engineering,
Democritus University of Thrace,
67100 Xanthi, Greece
{ltsochat,kzagoris,ipratika}@ee.duth.gr
msavelonas@gmail.com
http://vc.ee.duth.gr

Abstract. Mammography is currently the dominant imaging modality for the early detection of breast cancer. However, its robustness in distinguishing malignancy is relatively low, resulting in a large number of unnecessary biopsies. A computer-aided diagnosis (CAD) scheme, capable of visually justifying its results, is expected to aid the decision made by radiologists. Content-based image retrieval (CBIR) accounts for a promising paradigm in this direction. Facing this challenge, we introduce a CBIR scheme that utilizes the extracted features as input to a support vector machine (SVM) ensemble. The final features used for CBIR comprise the participation value of each SVM. The retrieval performance of the proposed scheme has been evaluated quantitatively on the basis of the standard measures. In the experiments, a set of 90 mammograms is used, derived from a widely adopted digital database for screening mammography. The experimental results show the improved performance of the proposed scheme.

Keywords: Content-Based Image Retrieval, Mammography, Support Vector Machines

1 Introduction

The use of content-based image retrieval (CBIR) schemes for computer-aided diagnosis (CAD) has been intensively investigated in the last decade [6]. Such an approach that facilitates searching for visually similar medical images, provides radiologists with visual aid and increases their confidence in incorporating CAD-cued results in their decision making [8].

There is only a limited amount of works devoted to CBIR-based CAD for breast masses in mammograms, although mammographic CAD is one of a mature and widely adopted CAD type [8]. An early attempt towards CBIR for breast masses in mammograms was the work of Alto et al. [1], who investigated the

discriminant capability of compactness, fractional concavity, spiculation index and Haralick's textural features. In a recent work, Wang et al. [5] tested the relationship between CAD performance and the similarity level between the region of interest (ROI) of the query and the ROIs resulting as outputs of CBIR.

All the above works evaluated retrieval performance on the basis of discriminant capability between benign and malignant cases. It can be argued that a CBIR scheme is expected to retrieve cases on the basis of visual similarity, since by its own nature it cannot take into account accompanying clinical data. In a real clinical setting, the results of CBIR could be jointly assessed with all such data in the context of an integrated CAD scheme. Finally, it can be observed that all CBIR methods presented were based on simple similarity measures, which cannot optimally exploit the distribution of mammogram ROIs in the feature space.

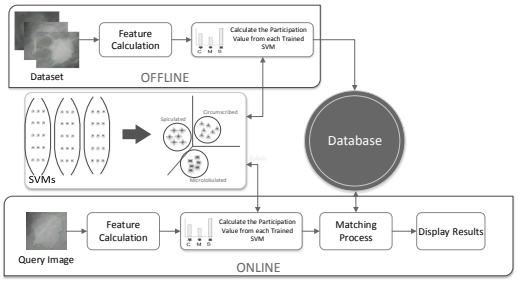

Fig. 1. The proposed system architecture.

In this work, we present a novel CBIR scheme, which utilizes a support vector machine (SVM) ensemble. The corresponding SVMs are capable of optimally exploiting the distribution of input samples in the feature space on the basis of breast imaging-reporting and data system (BI-RADS) classifications of breast masses [2], as performed by an expert radiologist. Then, based on each SVM's participation value, a new feature-vector is mapped to each feature set. The used dataset is formed by mammograms of the digital dataset for screening mammography (DDSM), which is widely adopted by the medical community.

The remainder of the paper is organized as follows: Section 2 describes the architecture of the proposed system. The experimental evaluation of the proposed CBIR scheme is presented in Section 3. Finally, conclusions and future perspectives of this work are discussed in Section 4.

2 Proposed CBIR scheme

Fig. 1 shows the proposed system architecture. At this point, it is worth to note that a mass detection and segmentation stage is applied prior to the proposed

pipeline. Since those two stages are out of the scope of the proposed system, there will be no further discussion about the methodology used. For our experiments, the segmentation information is taken from the ground truth corpora.

Initially, the following features are extracted from the mass boundaries:

1. *Solidity:* $F_{solidity} = A/H$, where A and H denote the areas of the shape and its corresponding convex hull, respectively.
2. *Compactness:* $F_{compactness} = 1-(4\pi A/P^2)$, where A and P denote the area and the perimeter of the shape, respectively.
3. *Discrete Fourier Transformation (DFT) coefficients of the Normalized Radial Length (NRL) function.* The Radial Length Function corresponds to the distance of each contour point (t) from the mass centroid (x_c, y_c): $r(t) = \sqrt{(x(t)-x_c)^2+(y(t)-y_c)^2}$. This function is sampled to a fixed number of points (256 in this work) and is normalized before its DFT computation.

Thereafter, the above feature set is supplied to three different trained SVMs that correspond to three classes of breast masses, based on BI-RADS [2], namely spiculated, micro -lobulated and circumscribed.

The Support Vector Machines (SVMs) [3] are based on statistical learning theory and have been successfully applied to several classification problems because of their discriminant ability and the fact that do not require large training sets.

Instead of utilizing the sign of the SVM decision function, we propose to normalize it, based on [7], in order to calculate a participation value of each feature vector to each trained SVM, which corresponds to each BI-RADS class. The normalized decision function is calculated by the following equation:

$$R(x) = \begin{cases} \max\left\{\frac{1}{1+\frac{1}{3}e^{f(x)}}, \frac{1}{1+\frac{1}{3}e^{-f(x)}}\right\} & if\ f(x) > 0 \\ 1 - \max\left\{\frac{1}{1+\frac{1}{3}e^{f(x)}}, \frac{1}{1+\frac{1}{3}e^{-f(x)}}\right\} & if\ f(x) < 0 \end{cases} \quad (1)$$

where $f(x)$ denotes the SVM decision function. The output of the Eq. 1 represents the membership value of the data x to the corresponding class and ranges in the interval [0,1]. Finally, the outputs of the SVMs construct the new three-element feature vector, used in the remainder of the retrieval process. In the sequel, the euclidean distance between the query and each indexed samples is computed leading to a ranked list of similar objects.

3 Experimental evaluation

In this study, 90 regions of interest (ROI) were used, extracted from various mammograms of DDSM, which contain masses. Each case is accompanied with ground-truth delineations and additional information, such as the biopsy-proven pathology of the lesion, its shape and margin types, the overall breast density and the assessment, of an expert radiologist [4] based on the BI-RADS standards. The margin types that were taken into consideration in this work are circumscribed, micro-lobulated and spiculated, as they are highly correlated with the

mass' pathology. For each margin type, masses of various shapes (oval, round, lobulated and irregular - Fig.2) were included. For each selected ROI, the contour of the depicted mass was acquired, by an expert radiologist indicating the exact position of its margin.

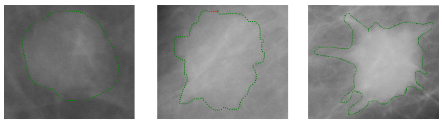

Fig. 2. Example ROIs for each margin type: A circumscribed (left), a micro-lobulated (center) and a spiculated (right) mass.

Two performance evaluation metrics are employed, which measure the system's ability to retrieve masses of similar margin type to the query. The first one is the Precision at Top 5 Retrieved items (P@5), which defines how successfully the algorithms produce relevant results to the first 5 position of the ranking list. The second metric used, is the Mean Average Precision (MAP) which is a typical measure for the performance of information retrieval systems and it is defined as the average of the precision value obtained after each relevant retrieved item.

The BI-RADS SVMs used a radial basis function (RBF) kernel, they were trained by the 2/3 and evaluated with the remainder portion of the dataset.

The proposed method was evaluated against a typical, unsupervised, state-of-the-art retrieval system employing the euclidean distance, calculated directly from the features instead of the participation values from each SVM. Comparative results are presented in Table 1 and show that the proposed method outperformed the typical unsupervised euclidean-based retrieval system.

Table 1. Experimental Results

Classes	**Unsupervised CBIR**		**SVM-based CBIR**	
	P@5	**MAP**	**P@5**	**MAP**
Circumscribed	0.9	0.915	0.9	0.92
Microlobulated	0.71	0.723	0.71	0.763
Spiculated	0.654	0.608	0.745	0.73
Average	0.75	0.743	0.78	0.8

4 Conclusions

This work introduced a CBIR scheme, which utilizes a support vector machine (SVM) ensemble. The retrieval performance of the proposed scheme has been

evaluated on the basis of BI-RADS classifications of breast masses. The used dataset is formed by 90 cases of the DDSM, which is a dataset widely adopted by the medical community. The experimental results lead to the conclusion that the proposed CBIR scheme outperforms standard euclidean-based retrieval, while greatly reducing the feature vector dimension and, consequently, the computational cost.

Future perspectives of this work include: 1) the integration of the proposed CBIR scheme within the context of a mammographic CAD system, which will also consider accompanying clinical and textual data, 2) the development of a similar CBIR scheme to facilitate CAD of breast microcalcifications.

Acknowledgements

This work is funded by the Hellenic Republic, Ministry of Education and Religious Affairs, General Secretariat of Research and Technology (GSRT), and particularly the National Programme "SYNERGASIA 2011" (11SYN_10_1546) in the National Strategic Reference Framework (NSRF) 2007-2013.

References

1. Alto, H., Rangayyan, R.M., Desautels, J.E.L.: Content-based retrieval and analysis of mammographic masses. J. Electronic Imaging 14(2), 023016 (2005)
2. Berg, W.A., Campassi, C., Langenberg, P., Sexton, M.J.: Breast imaging reporting and data system: inter- and intraobserver variability in feature analysis and final assessment. Am. J. Roentgenol. 174(6), 1769–1777 (2000)
3. Cortes, C., Vapnik, V.: Support vector networks. Machine Learning 20, 273–197 (1995)
4. Heath, M., Bowyer, K., Kopans, D., Kegelmeyer Jr, P., Moore, R., Chang, K., Munishkumaran, S.: Current status of the digital database for screening mammography. In: Digital mammography, pp. 457–460. Springer (1998)
5. Wang, X., Park, S., Zheng, B.: Assessment of performance and reliability of computer-aided detection scheme using content-based image retrieval approach and limited reference database. Journal of Digital Imaging 24(2), 352–359 (2011)
6. Welter, P., Fischer, B., Günther, R.W., Deserno (Né Lehmann), T.M.: Generic integration of content-based image retrieval in computer-aided diagnosis. Comput. Methods Prog. Biomed. 108(2), 589–599 (2012)
7. Zagoris, K., Ergina, K., Papamarkos, N.: Image retrieval systems based on compact shape descriptor and relevance feedback information. Journal of Visual Communication and Image Representation 22(5), 378 – 390 (2011)
8. Zheng, B.: Computer-aided diagnosis in mammography using content-based image retrieval approaches: Current status and future perspectives. Algorithms 2(2), 828–849 (2009)