Preprocessing of Screening Mammograms Based on Local Statistical Models

Jiří Grim
Institute of Information Theory and Automation
Czech Academy of Sciences
P.O. BOX 18, 18208 Prague 8
Czech Republic
grim@utia.cas.cz

ABSTRACT

In a recent paper we have proposed evaluation of screening mammograms by means of local statistical model. The model describes local statistical properties of internal pixels of a small search window - as they occur when scanning the mammogram. It is defined as a mixture of Gaussian densities which can be estimated by EM algorithm from data obtained by shifting the search window. The estimated Gaussian mixture is used to compute at all window positions the log-likelihood values which can be displayed as grey levels at the respective window centers. The resulting log-likelihood image closely correlates with the structural details of the original mammogram and emphasizes potential malignant findings as untypical locations of high novelty. In this paper we discuss the possibilities to enhance the log-likelihood image for diagnostic purposes.

Categories and Subject Descriptors

I.4 [**Image Processing And Computer Vision**]: Image processing software; J.3 [**Life And Medical Sciences**]: Health

General Terms

Algorithms, Theory

1. INTRODUCTION

Mammographic screening is currently the most effective and widely accepted strategy to detect early stages of breast cancer and to reduce the related mortality rates by appropriate treatment. In many countries there are numerous screening programs producing millions of mammograms to be evaluated by specially trained radiologists. Reading of mammograms is a difficult task underlying very strong reliability requirements. Moreover, the early subtle stage of pathological finding is of key importance since palpable malignant lesions already can make metastases. One can use double reading to reduce the risk of possible false negative evaluation of mammograms but, as a standard of care, this approach becomes too resource demanding.

For obvious reasons there is a strong motivation to help radiologists by using computer-aided decision-supporting systems. However, despite of long history of research in this area, there is no satisfactory solution of the early breast cancer detection problem. From the point of view of statistical decision-making the evaluation of screening mammograms is known to be extremely complex. The malignant abnormalities may be essentially of two kinds, so called "masses" and "calcifications", but they may have different shapes, margins or distributions, may differ in size and, unlike typical pattern recognition problems, their locations in mammograms may be nearly arbitrary. The aim of breast cancer screening is to detect malignant lesions at very early stages, i.e. of very small size and not fully developed form. Simultaneously, the location of the suspect finding is of basic importance. In this connection there is a serious question if the diagnostic evaluation of screening mammograms can reasonably be solved as a classification problem.

Let us recall that, considering a statistical decision problem, we assume that some multivariate observations have to be classified with respect to a finite number of classes. In order to design a statistical decision rule we need sufficiently large set of training data and a comparably large test-data set having identical properties. According to general experience the size of training data set should be much higher than the data dimension but, unfortunately, the dimension of screening mammogram (considered as a data vector) is given by the very high number of pixels.

Since any subsampling could cause an unacceptable information loss, the widely accepted approach is to reduce the decision-making only to a limited "region of interest" (ROI). Of course, in such a case the classifier must evaluate the content of a sliding window in all possible mammogram positions to specify ROI or the suspect locations have to be chosen by radiologist or found by another special procedure. (Note that two ROI's in adjacent positions produce completely different data vectors since the internal pixel values are shifted into different dimensions.) Nevertheless, even a reasonably small ROI must contain hundreds or rather thousands of pixels in order to encompass possible greater findings and therefore some feature extraction method would be indispensable.

We recall that, even if we would succeed to extract a small number of informative features (invariant with respect to

the size, shape and location of the finding), it would be very difficult to create a reasonably representative training set (or training sets for all possible types of malignant lesions). The reason is the extreme diversity of malignant abnormalities and high natural variability of mammograms differing by age and type (dense, fatty) and additionally influenced by the "positional" noise due to manual placement of the breasts by the radiographer. It is probable that, under strongly independent test conditions and without any preselection of mammograms the classification performance would not be very optimistic.

At present the widely used computer-aided detection systems (CADe) just only place prompts on mammogram to direct the attention of the radiologists toward potential abnormalities. According to different studies (cf. [3, 2, 4, 6]), CADe systems can improve breast cancer detection by more than 20% but the false positive rates of placing prompts are high, usually in units per image [3, 11]. Thus, because of the low incidence of malignant findings (0.1 – 0.3%), standard CADe systems generate hundreds or even thousands of false positive prompts in order to correctly mark one malignant lesion.

2. LOCAL STATISTICAL MODEL

In a recent paper [8] we have proposed preprocessing of screening mammograms by means of local statistical models. The idea of the method is to emphasize diagnostically important details as "unusual" locations of high "novelty". We estimate the local statistical model of a mammogram as a joint probability density of pixel values in a suitably chosen search window. In particular, we have used a square window of 13x13 pixels with trimmed corners with the resulting dimension N=145 (cf. [8] for more details). Denoting by

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_N) \in \mathcal{R}^{\mathcal{N}}$$

the real vector of grey levels of the window inside in a fixed pixel arrangement, we assume the joint probability density $P(\boldsymbol{x})$ in the form of a mixture of Gaussian components

$$P(\boldsymbol{x}) = \sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m), \quad \boldsymbol{x} \in \mathcal{R}^{\mathcal{N}}. \tag{1}$$

Here we denote $\mathcal{M} = \{1, \ldots, M\}$, and $\mathcal{N} = \{1, \ldots, N\}$ the index sets of components and variables respectively, and define the mixture components as products of univariate Gaussian densities

$$F(\boldsymbol{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) = \prod_{n\in\mathcal{N}} f_n(x_n|\mu_{mn}, \sigma_{mn}), \quad \boldsymbol{x} \in \mathcal{R}^{\mathcal{N}}, \tag{2}$$

$$f_n(x_n|\mu_{mn}, \sigma_{mn}) = \frac{1}{\sqrt{2\pi}\sigma_{mn}} \exp\{-\frac{(x_n - \mu_{mn})^2}{2\sigma_{mn}^2}\}. \tag{3}$$

The standard way to estimate Gaussian mixture is to apply EM algorithm. We use the data set $\mathcal{S}$ obtained by pixelwise scanning the original mammogram with the search window

$$\mathcal{S} = \{\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}, \ldots\}, \quad \boldsymbol{x}^{(i)} \in \mathcal{R}^{\mathcal{N}}. \tag{4}$$

The corresponding log-likelihood function

$$L = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} \log \left[\sum_{m\in\mathcal{M}} w_m F(\boldsymbol{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \right] \tag{5}$$

can be maximized by the following EM iteration equations (cf. [7, 9]) ($m \in \mathcal{M}, n \in \mathcal{N}, \boldsymbol{x} \in \mathcal{S}$)

$$q(m|\boldsymbol{x}) = \frac{w_m F(\boldsymbol{x}|\boldsymbol{\mu}_m, \boldsymbol{\sigma}_m)}{\sum_{j\in\mathcal{M}} w_j F(\boldsymbol{x}|\boldsymbol{\mu}_j, \boldsymbol{\sigma}_j)}, \tag{6}$$

$$w_m^{'} = \frac{1}{|\mathcal{S}|} \sum_{x\in\mathcal{S}} q(m|\boldsymbol{x}), \tag{7}$$

$$\mu_{mn}^{'} = \frac{1}{w_m'|\mathcal{S}|} \sum_{x\in\mathcal{S}} x_n q(m|\boldsymbol{x}), \tag{8}$$

$$(\sigma_{mn}^{'})^2 = \frac{1}{w_m'|\mathcal{S}|} \sum_{x\in\mathcal{S}} (x_n - \mu_{mn}^{'})^2 q(m|\boldsymbol{x}). \tag{9}$$

Here the apostrophe denotes the new parameter values in each iteration. In our experiments we have used $M = 36$ components and the component parameters have been initialized randomly (cf. [8] for more details).

The local statistical model is estimated from each mammogram individually and therefore the method is not confronted with the natural variability of mammograms. This circumstance is relevant in view of the extreme diversity of malignant lesions. We recall that the mixture model (1) is invariant with respect to the linear transform of the underlying grey scale (cf. [8], Appendix I).

Let us remark that the estimation of Gaussian mixtures may cause computational difficulties in multidimensional spaces. For dimensions of order $N \approx 10^2$ the component values (2) may become too small to be correctly represented in memory and therefore the EM algorithm may become instable. The problem may be easily removed if we compute the components in logarithmic form. By adding a suitably chosen constant we actually multiply the components by a coefficient which can be reduced in the fraction (6). In this way only very small component values will be neglected against the current maximum.

3. LOG-LIKELIHOOD IMAGE

Having estimated the parameters we evaluate the mixture density $P(\boldsymbol{x})$ at each window position for the underlying window-patch vector $\boldsymbol{x}$ and display the corresponding log-likelihood value $\log P(\boldsymbol{x})$ as grey level at the central reference pixel of the window. In this sense the light grey levels correspond to the "typical" highly probable parts of the image and the dark values reflect the less-probable, "untypical" or "unusual" locations. In this way, the resulting log–likelihood image should facilitate the identification of malignant abnormalities as locations of high novelty - exactly in the sense of the idea proposed by Rose and Taylor and others (cf. [13, 12]).

A frequent subject of different CADe techniques is the detection of microcalcifications usually based on some kind of thresholding (cf. e.g., [5] for extensive references). In case of log-likelihood image, the detection of microcalcifications is closely related to the underlying Gaussian mixture. At all window positions containing a "disturbing" light pixel we obtain lower log-likelihood values (cf. (2)). As a result, the log-likelihood image will contain a dark spot of window size centered on the light pixel. The size and darkness of the spot continuously depend on the size and contrast of the underlying micro-calcification, respectively.

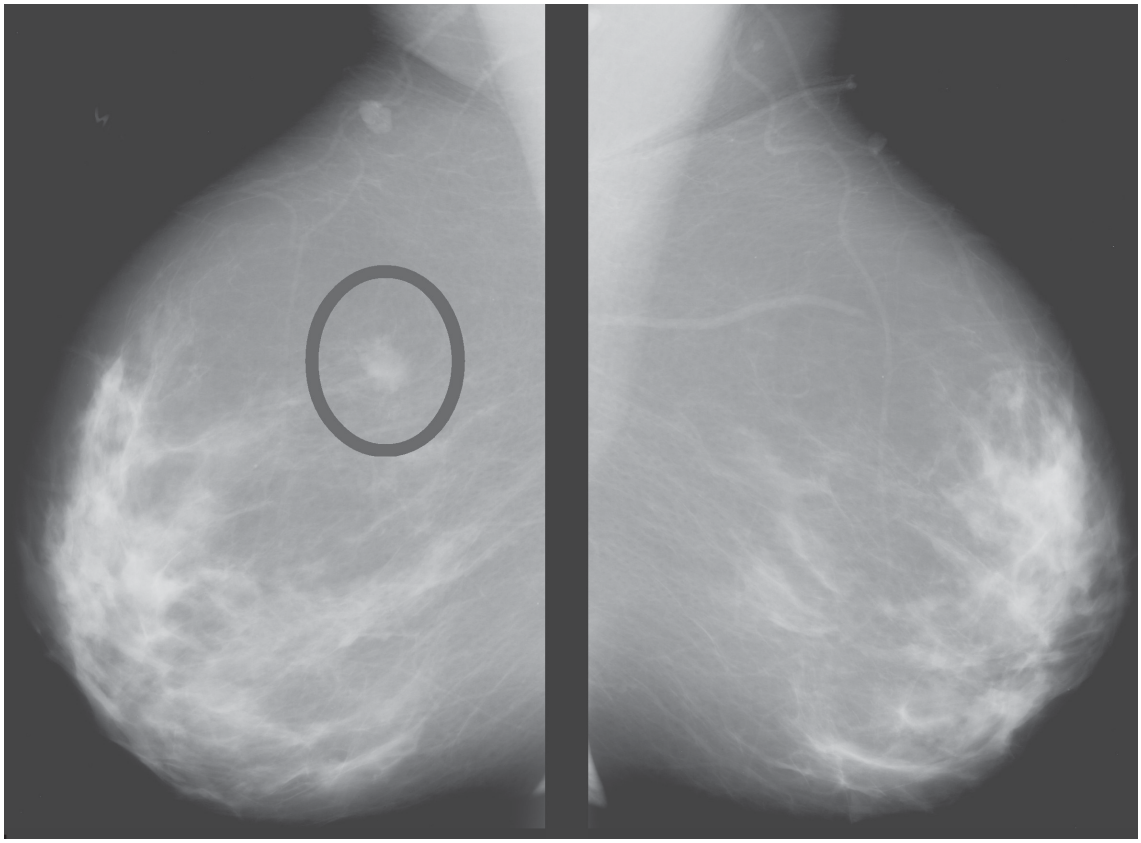

Figure 1: Original image of the mammogram B-3020-1 (right- and left medio-lateral-oblique parts) from the DDSM database [1]. In the left upper part of the image there is a malignant circumscribed mass with lobulated margins.

Comparing the original mammogram and the corresponding log-likelihood image, we can see that structural details closely correlate in both images but the boundaries of different regions are emphasized. The tendency of the local statistical model to create "contour lines" is an artefact having a simple theoretical reason. The log-likelihood values $\log P(\boldsymbol{x})$ are typically "dominated" by a single component of the mixture, which is most adequate to the underlying region. A detailed numerical observation shows that the "switching" of dominating mixture components is responsible for the arising dark contour lines at the boundaries of regions having different textural properties.

The most apparent demonstration of this mechanism can be seen at the margins of the breast region which are characterized by continuously decreasing grey levels (cf. Fig. 1 and 2). We observe that, unlike iso-intensity contours (cf. e.g., [10]), the contour lines produced by mixture components should generally emphasize regions of similar textural properties.

4. MODIFIED STATISTICAL MODEL

Let us recall that the log-likelihood image is a purely statistical tool based on the local statistical model. There is no specific relation between the log-likelihood image and screening mammography, except for the hypotheses that the malignant lesions can be identified as unusual locations of high novelty. The advantage of this principle is that the method need not be trained, it is generally applicable and, unlike statistical pattern recognition, the high natural variability of mammograms and diversity of malignant abnormalities do not cause any specific problems. The resulting preprocessed mammogram may be helpful to the radiologists because any asymmetry of structural details becomes well visible, the regions of identical textural properties are emphasized by contour lines and microcalcifications are displayed as dark well visible spots.

On the other hand, the most dominant features of the log-likelihood image typically correspond to the light structural details which represent a small part of the mammogram and therefore occur as untypical. Consequently, they correspond to dark grey levels but they are usually quite normal without any pathological meaning. We would like modify the log-likelihood image with the aim to suppress the irrelevant features and increase the sensitivity of the method with respect to the potentially pathological findings. There are different possibilities to modify the method by changing the structure of the search window or by including additional variables to the data vector (cf. [8]) but also the EM iteration equations can be modified.

A promising approach provides the weighted version of EM algorithm. In particular, denoting $\gamma(\boldsymbol{x})$ the relative frequency of $\boldsymbol{x}$ in the training data set $\mathcal{S}$:

$$\gamma(\boldsymbol{x}) = \begin{cases} N(\boldsymbol{x})/|\mathcal{S}|, & \text{for } \boldsymbol{x} \in \mathcal{S}, \\ 0, & \text{for } \boldsymbol{x} \notin \mathcal{S} \end{cases} \tag{10}$$

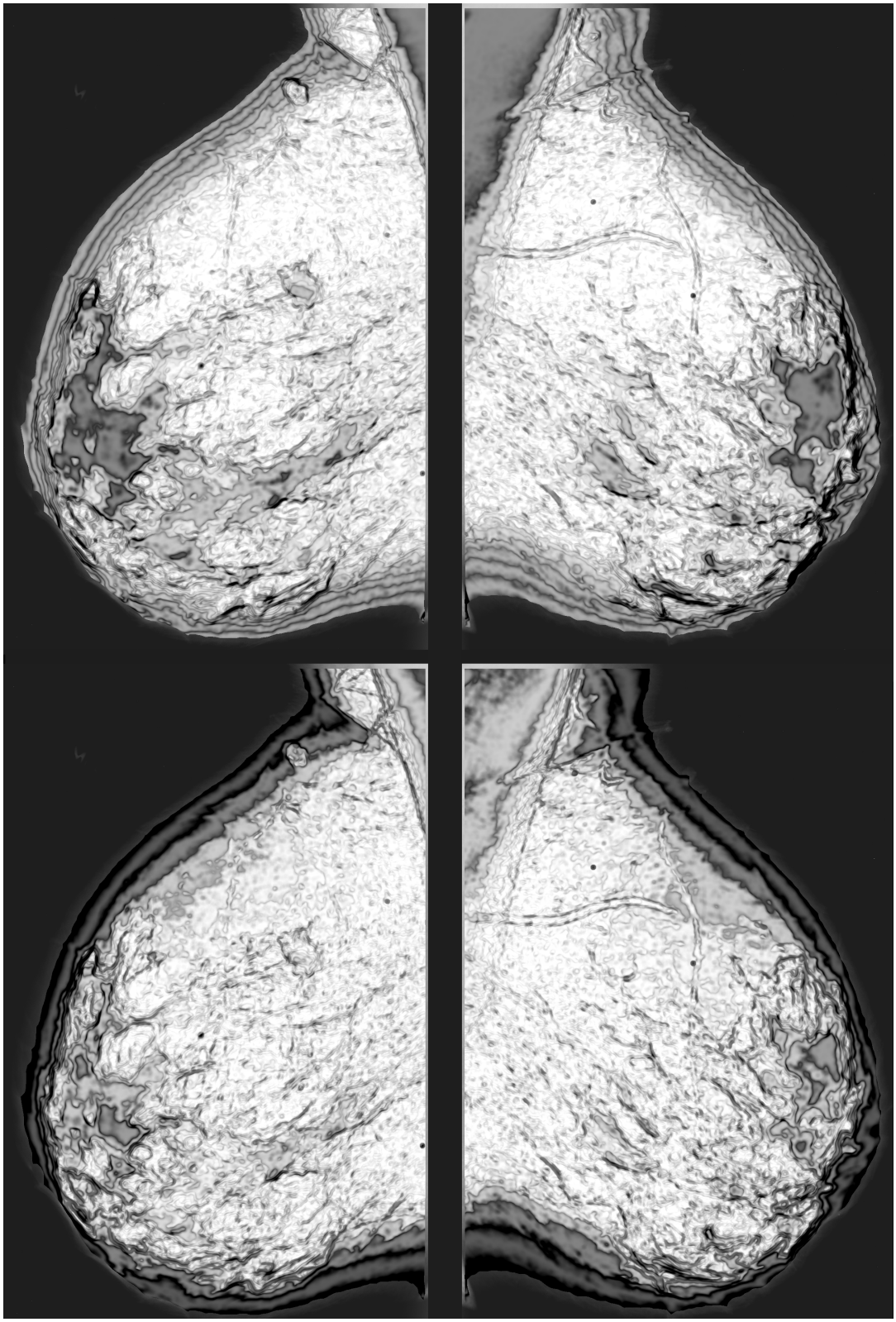

Figure 2: The log-likelihood image of the original mammogram from Fig.1. (in the upper part) has the same resolution. Each pixel value is defined by log-likelihood value $\log P(\boldsymbol{x})$ where $\boldsymbol{x}$ is the 145-dimensional vector defined by the search window. The malignant lesion is partly emphasized by contour lines. The lower part of the figure shows the modified (weighted) version of the log-likelihood image. The resulting "weighted" log-likelihood image (in the lower part) contains finer details and additional contour lines in the light region.

and $\boldsymbol{X}$ the subset of $\mathcal{R}^{\mathcal{N}}$ with $\gamma(\boldsymbol{x}) > 0$:

$$\boldsymbol{X} = \{\boldsymbol{x} \in \mathcal{R}^{\mathcal{N}} : \gamma(\boldsymbol{x}) > 0\}, \quad (\sum_{\boldsymbol{x} \in \boldsymbol{X}} \gamma(\boldsymbol{x}) = 1)$$

we can rewrite the criterion (5) and the EM iteration equations (7) - (9) equivalently in the following form:

$$L = \sum_{\boldsymbol{x} \in \boldsymbol{X}} \gamma(\boldsymbol{x}) \log \left[\sum_{m \in \mathcal{M}} w_m F(\boldsymbol{x} | \boldsymbol{\mu}_m, \boldsymbol{\sigma}_m) \right], \tag{11}$$

$$w_m^{'} = \sum_{x \in \boldsymbol{X}} \gamma(\boldsymbol{x}) q(m|\boldsymbol{x}), \tag{12}$$

$$\mu_{mn}^{'} = \frac{1}{w_m^{'}} \sum_{x \in \boldsymbol{X}} x_n \gamma(\boldsymbol{x}) q(m|\boldsymbol{x}), \tag{13}$$

$$(\sigma_{mn}^{'})^2 = \frac{1}{w_m^{'}} \sum_{x \in \boldsymbol{X}} (x_n - \mu_{mn}^{'})^2 \gamma(\boldsymbol{x}) q(m|\boldsymbol{x}). \tag{14}$$

Note that, in the above equations, we may specify arbitrary weights $\gamma(\boldsymbol{x})$ which can be interpreted as a repeated occurrence of the respective vectors in the data set. In this way the weighted EM algorithm can be used to increase the meaning of different data vectors individually.

In this paper we derive the weight function directly from the original mammogram (cf. Fig.1). In order to address the problem of the too emphasized light normal regions we define the weight function by Eq.

$$\gamma(\boldsymbol{x}) = \frac{1}{N} \sum_{n \in \mathcal{N}} x_n. \tag{15}$$

In other words, if we use the mean value of the grey levels x_n in the window as a weight, then the "presence" of light locations in the data set will artificially increase. Consequently, they will be less unusual and therefore less emphasized. In the experiment we have used the modified EM algorithm (11) - (14) with the weight function (15). It can be seen that the resulting "weighted" log-likelihood image (cf. Fig.2, lower part) contains finer details and additional contour lines in the light region. In this sense the weighted modification of our method could be useful in case of "dense" mammograms when the identification of malignant masses is particularly difficult.

5. CONCLUSION

We propose a weighted modification of a recently proposed method of diagnostic enhancement of screening mammograms. In our experiments we have used the mean grey level of the window inside as a weighting function. In comparison with the original log-likelihood image the modified image shows finer structural details and additional contour lines. Again, there is no specific relation between the modified method and screening mammography, only the underlying "high-novelty" hypotheses has been slightly modified. Nevertheless, we recall that there is wide variety of other possibilities to specify the weight function which could improve the diagnostic conspicuity of the final log-likelihood image.

Acknowledgment

This work was supported by the MŠMT projects 2C06019 ZIMOLEZ and 1M0572 DAR.

6. REFERENCES

[1] K. W. Bowyer, D. Kopans, W. P. Kegelmeyer, R. Moore, and M. Sallam. *The digital database for screening mammography*. Kluwer Academic Publishers, 1998.

[2] R. F. Brem, J. K. Baum, S. Kaplan, S. Sounders, L. G. Naul, and J. Hoffmeister. Improvement in sensitivity screening mammography with computer-aided detection (cad): a multi-institutional trial. *American Journal of Roentgenology*, 181:687–693, 2003.

[3] R. F. Brem, J. A. Rapelyea, G. Zisman, J. W. Hoffmeister, and M. P. DeSimio. Evaluation of breast cancer with a computer-aided detection system by mammographic appearance and histopathology. *Cancer*, 104(5):931–935, 2005.

[4] W. L. J. Burhenne, S. A. Wood, and C. J. D. et al. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology*, 215:554–562, 2000.

[5] H. D. Cheng, X. J. Shi, R. Min, L. M. Hu, X. P. Cai, and H. N. Du. Approaches for automated detection and classification of masses in mammograms. *Pattern Recognition*, 39(11):646–668, November 2006.

[6] T. W. Freer and M. J. Ulissey. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology*, 220:781–786, 2001.

[7] J. Grim, M. Haindl, P. Somol, and P. Pudil. A subspace approach to texture modelling by using gaussian mixtures. In *Proceedings of the 18th IAPR International Conference on Pattern Recognition ICPR 2006*, volume 2, pages 235–238, Los Alamitos, CA, USA, 2006. IEEE Computer Society.

[8] J. Grim, P. Somol, M. Haindl, and J. Danes. Computer-aided evaluation of screening mammograms based on local texture models. *Image Processing, IEEE Transactions on*, 18(4):765 –773, april 2009.

[9] J. Grim, P. Somol, M. Haindl, and P. Pudil. A statistical approach to local evaluation of a single texture image. In *Proceedings of the 16-th Annual Symposium PRASA 2005*, pages 171–176, 2005.

[10] D. Guliato, R. M. Rangayyan, J. A. Zuffo, , and J. E. L. Desautels. Detection of breast tumor boundaries using iso-intensity contours and dynamic thresholding. In *Proc. 4th Int. Workshop on Digital Mammography IWDM 1998*, pages 253–260, Nijmegen, The Netherlands, 1998.

[11] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels. Detection of breast masses in mammograms by density slicing and texture flow-field analysis. *IEEE Trans. Med. Imaging*, 20:1215–1227, November 2001.

[12] C. Spence, L. Parra, and P. Sajda. Detection, synthesis and compression in mammographic image analysis with a hierarchical image probability model. In *Proc. of the IEEE Workshop on Mathematical Methods in Biomedical Image Analysis*, 2001.

[13] L. Tarassenko, P. Hayton, and N. C. et al. Novelty detection for the identification of masses in mammograms. In *Proceedings of the 4th IEE International Conference on Artificial Neural Networks*, page 442Ű447, Cambridge, UK, 1995.