Relevance Vector Machine for Automatic Detection of Clustered Microcalcifications

Liyang Wei, *Student Member, IEEE*, Yongyi Yang*, *Senior Member, IEEE*, Robert M. Nishikawa, Miles N. Wernick, *Senior Member, IEEE*, and Alexandra Edwards

***Abstract*—Clustered microcalcifications (MC) in mammograms can be an important early sign of breast cancer in women. Their accurate detection is important in computer-aided detection (CADe). In this paper, we propose the use of a recently developed machine-learning technique – relevance vector machine (RVM) – for detection of MCs in digital mammograms. RVM is based on Bayesian estimation theory, of which a distinctive feature is that it can yield a sparse decision function that is defined by only a very small number of so-called relevance vectors. By exploiting this sparse property of the RVM, we develop computerized detection algorithms that are not only accurate but also computationally efficient for MC detection in mammograms. We formulate MC detection as a supervised-learning problem, and apply RVM as a classifier to determine at each location in the mammogram if an MC object is present or not. To increase the computation speed further, we develop a two-stage classification network, in which a computationally much simpler linear RVM classifier is applied first to quickly eliminate the overwhelming majority, non-MC pixels in a mammogram from any further consideration. The proposed method is evaluated using a database of 141 clinical mammograms (all containing MCs), and compared with a well-tested support vector machine (SVM) classifier. The detection performance is evaluated using free-response receiver operating characteristic (FROC) curves. It is demonstrated in our experiments that the RVM classifier could greatly reduce the computational complexity of the SVM while maintaining its best detection accuracy. In particular, the two-stage RVM approach could reduce the detection time from 250 s for SVM to 7.26 s for a mammogram (nearly 35-fold reduction). Thus, the proposed RVM classifier is more advantageous for real-time processing of MC clusters in mammograms.**

***Index Terms*—Breast cancer detection, computer-aided diagnosis, mammography, microcalcifications, relevance vector machine.**

I. Introduction

Breast cancer is a common form of cancer diagnosed in women. One of the important early signs of breast cancer in mammograms is the appearance of microcalcification (MC) clusters, which appear in 30%–50% of mammographically diagnosed cases [1]. MCs are calcium deposits of very small dimension and appear as a group of granular bright spots in a mammogram. As an example, Fig. 1 shows a mammogram image with a cluster of MCs. Individual MCs are sometimes difficult to detect because of the surrounding breast tissue, their variation in shape and small dimension.

Because of its importance in breast cancer diagnosis, accurate detection of MC clusters is an important problem. In recent years, there has been a great deal of research in development of computerized methods for automatic and accurate detection of MC clusters, which could potentially assist radiologists in diagnosis of breast cancer. A thorough review of various methods for MC detection reported in the literature can be found in [2].

In our previous work [3], we developed a support vector machine (SVM) approach for detection of clustered MCs in mammograms, and demonstrated that such an approach could outperform several well-known methods in the literature, such as the image difference technique (IDT) in [4], the difference of Gaussian (DoG) method in [5], the wavelet-decomposition based method in [6], and the two-stage multilayer neural network (TMNN) method in [7]. It was demonstrated that the SVM approach could achieve the best detection performance when evaluated using the free-response receiver operating characteristic (FROC) curves. For the ease of reference and comparison, we show in Fig. 2 the FROC evaluation results for the different methods obtained in [3].

While the SVM approach achieves the best detection performance, the computational complexity of the SVM classifier may prove to be burdensome in real-time or near real-time applications. In an SVM classifier, the decision function is determined by a subset of training samples [called support vectors (SVs)]; the computational complexity of the decision function (which is nonlinear) is linearly proportional to the number of SVs. At the root of the problem is that no sparsity constraint is explicitly imposed on the SVM classifier model; as a consequence, too many a SV can lead to a classifier that is computationally expensive. This issue is especially important for MC detection, as modern digital mammography scanners can produce images at high resolutions, which may require significant computation time to process.

In this work, we propose to improve the computational efficiency of our previously developed SVM classifier by using an alternative, recently developed machine learning technique –

Manuscript received March 3, 2005; revised July 14, 2005. This work was supported in part by the National Institutes of Health (NIH) under NCI Grant CA89668. The Associate Editor responsible for coordinating the review of this paper and recommending its publication was N. Karssemeijer. *Asterisk indicates corresponding author.*

L. Wei is with the Department of Biomedical Engineering, Illinois Institute of Technology, Chicago, IL 60616 USA.

Y. Yang is with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, 3301 South Dearborn Street, Chicago, IL 60616 USA.

R. M. Nishikawa is with the Department of Radiology, The University of Chicago, Chicago, IL 60637 USA. He is a shareholder in R2 Technology, Inc. (Sunnyvale, CA).

M. W. Wernick is with the Department of Electrical and Computer Engineering, Illinois Institute of Technology, Chicago, IL 60616 USA.

A. Edwards is with the Department of Radiology, The University of Chicago, Chicago, IL 60637 USA..

Digital Object Identifier 10.1109/TMI.2005.855435

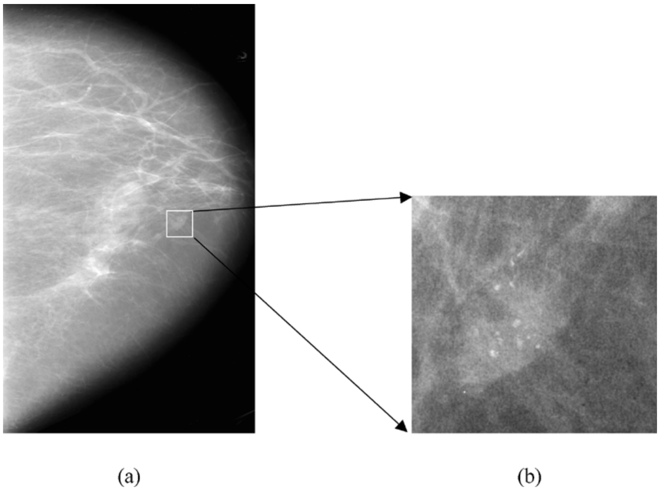

Fig. 1. (a) Mammogram in craniocaudal view. (b) Expanded view showing MCs.

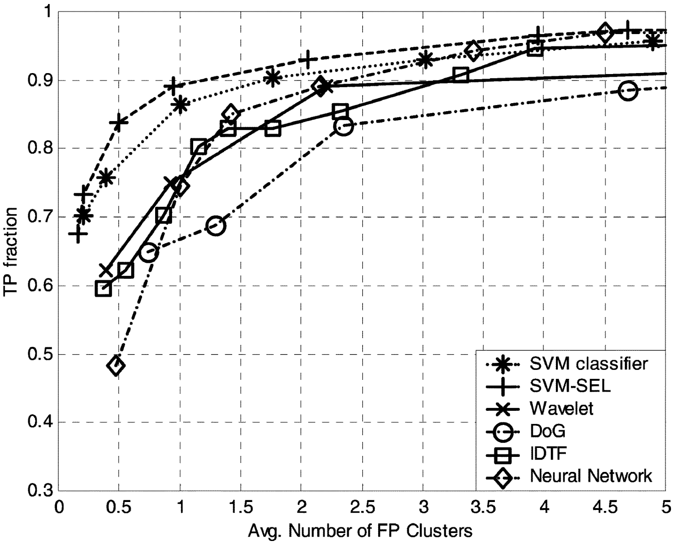

Fig. 2. FROC curves of the different tested methods (cited from [3]).

relevance vector machine (RVM) [8] – for MC detection. The advantage of this approach is that the RVM classifier can yield a decision function that is much sparser than the SVM while maintaining its detection accuracy. This can lead to significant reduction in the computational complexity of the decision function, thereby making it more suitable for real-time applications.

Pioneered by Tipping [8], the RVM learning approach is based on Bayesian estimation theory, which can be applied for both classification and regression problems. It is reported to yield nearly identical performance to, if not better than, that of SVM in several benchmark studies [8]. A key feature of RVM is that it can yield a solution function that depends on only a very small number of kernel functions [called *relevance vectors* (RVs)] than SVM.

In addition to development of the RVM classifier for MC detection, we also propose a two-stage RVM classification approach, which can further reduce greatly the processing time of a mammogram image. This new approach is motivated by the fact that the overwhelming majority (over 99%) of the image pixels in a mammogram, if not all, are background pixels that do not exhibit MC features. In the proposed two-stage classification approach, we first apply a liner RVM classifier (much simpler computationally, which amounts to a template matching detector) to quickly eliminate those (background) pixels that can be easily differentiated from MCs; we next apply a more sophisticated RVM classifier (computationally more expensive) in the second stage to further classify only those pixels not eliminated by the first-stage classifier. As will be demonstrated in our experiments (Section VI), this approach can significantly reduce the overall processing time for a mammogram image.

The rest of the paper is organized as follows: A brief review of the theory of RVM learning for classification is provided in Section II, which is used to facilitate the development of the rest of the paper. Then, the formulation of RVM learning for the problem of MC detection is given in Section III. The two-stage RVM classification network is presented in Section IV. Issues related to our evaluation study (including mammogram data set, RVM training and testing, and performance evaluation) are described in Section V. Experiment results and discussions are furnished in Section VI. Finally, conclusions are drawn in Section VII.

II. Review of RVM Learning for Classification

In this paper, MC detection is formulated as a binary classification problem. Specifically, at each location of a mammogram image, we apply a classifier to determine whether an MC object is present or not. To begin, let vector $\mathbf{x} \in R^n$ denote a pattern to be classified, and let scalar d denote its class label (i.e., $d \in \{\pm 1\}$). In addition, let $\{(\mathbf{x}_i, d_i),\ i = 1, 2, \ldots, N\}$ denote a given set of N training examples, where each sample $\mathbf{x}_i$ has a known class label d_i. The problem is how to determine a classifier $f(\mathbf{x})$ (i.e., a decision function) that can correctly classify an input pattern (not necessarily from the training set).

A. RVM Classifier Model

For an input vector $\mathbf{x}$, an RVM classifier models the probability distribution of its class label $d \in \{-1, +1\}$ using logistic regression as

$$p(d = 1 \mid \mathbf{x}) = \frac{1}{1 + \exp\left(-f_{\mathrm{RVM}}(\mathbf{x})\right)} \tag{1}$$

where $f_{\mathrm{RVM}}(\mathbf{x})$, the classifier function, is given by

$$f_{\mathrm{RVM}}(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i K(\mathbf{x}, \mathbf{x}_i) \tag{2}$$

where $K(\cdot, \cdot)$ is a kernel function, and $\mathbf{x}_i$, $i = 1, 2, \ldots, N$, are the training samples.

According to Tipping [8], the parameters $\alpha_i, i = 1, 2, \ldots, N$, in $f_{\mathrm{RVM}}(\mathbf{x})$ are determined using Bayesian estimation. Toward this end, a sparse prior is introduced on α_i. To be specific, these parameters are assumed to be statistically independent and each obeys a zero-mean, Gaussian distribution with variance λ_i^{-1}; furthermore, a so-called *hyper-prior* (based on the Gamma distribution) is assumed on the variance λ_i^{-1}, which is used to force the parameters α_i to be highly concentrated around 0, thereby leading to very few nonzero terms in $f_{\mathrm{RVM}}(\mathbf{x})$.

The parameters α_i in (2) are then obtained by maximizing the posterior distribution of the class labels given the input vectors.

This is equivalent to maximizing the following objective function:

$$J(\alpha_1, \alpha_2, \ldots, \alpha_N) = \sum_{i=1}^{N} \log p(d_i \mid \mathbf{x}_i) + \sum_{i=1}^{N} \log p(\alpha_i \mid \lambda_i^*) \tag{3}$$

where the first summation term corresponds to the likelihood of the class labels, and the second term corresponds to the prior on the parameters α_i, in which λ_i^* denotes the maximum *a posteriori* estimate of the hyper-parameter λ_i. In the resulting solution, only those samples associated with nonzero coefficients α_i, called *relevance vectors*, will contribute to the decision function $f_{\mathrm{RVM}}(\mathbf{x})$.

In (2), the kernel function $K(\cdot,\cdot)$ is used to form expansion basis functions for $f_{\mathrm{RVM}}(\mathbf{x})$, and in theory, is not limited by the Mercer's condition [10], unlike the case of SVM. The Mercer's condition states that $K(\cdot,\cdot)$ must be a positive integral operator, that is, for every square-integratable function $g(\cdot)$ defined on R^n the kernel $K(\cdot,\cdot)$ satisfies the following condition:

$$\int\!\!\int K(\mathbf{x},\mathbf{y}) g(\mathbf{x}) g(\mathbf{y})\, d\mathbf{x} d\mathbf{y} \geq 0. \tag{4}$$

In this paper, the popular polynomial kernels and the RBF kernels are used, which are known to satisfy the Mercer's condition. They are defined as follows.

1) Polynomial kernel:

$$K(\mathbf{x},\mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1)^p \tag{5}$$

where $p > 0$ is a constant that defines as the kernel order.

2) RBF kernel:

$$K(\mathbf{x},\mathbf{y}) = \exp\left(-\frac{\|\mathbf{x}-\mathbf{y}\|^2}{2\sigma^2}\right) \tag{6}$$

where $\sigma > 0$ is a constant that defines the kernel width.

B. Comparison to SVM Learning

SVM is a constructive learning procedure rooted in statistical learning theory [11]. It is based on the principle of structural risk minimization, which aims at minimizing the bound on the generalization error (i.e., error made by the learning machine on data unseen during training) rather than minimizing the mean square error over the data set [11]. As a result, this leads to good generalization and an SVM tends to perform well when applied to data outside the training set.

An SVM classifier in concept first maps the input data vector $\mathbf{x}$ into a higher dimensional space $\boldsymbol{\mathcal{H}}$ through an underlying nonlinear mapping $\Phi(\mathbf{x})$, then applies linear classification in this mapped space. Introducing a kernel function $K(\mathbf{x},\mathbf{y}) \equiv \Phi(\mathbf{x})^T\Phi(\mathbf{y})$, we can write an SVM classifier $f_{\mathrm{SVM}}(\mathbf{x})$ as follows:

$$f_{\mathrm{SVM}}(\mathbf{x}) = \sum_{i=1}^{N_s} \alpha_i K(\mathbf{x}, \mathbf{s}_i) + b \tag{7}$$

where $\mathbf{s}_i$, $i = 1, 2, \ldots, N_s$ are a subset of the training samples $\{\mathbf{x}_i, i = 1, 2, \ldots, N\}$ (which are called *support vectors*). For brevity, we refer the reader to [3] for details on SVM for MC detection.

The SVM classifier in (7) is noted to resemble in form the RVM classifier in (2), yet the two classifiers are derived from different principles. As will be illuminated later by the experimental results (Section VI), for SVM the SVs are typically formed by "borderline," difficult-to-classify samples in the training set, which are located near the decision boundary of the classifier; for RVM the RVs are formed by samples appearing to be more representative of the two classes, which are located away from the decision boundary of the classifier.

Compared to SVM, RVM is found to be advantageous on several aspects, including: 1) The RVM decision function can be much sparser than the SVM classifier, i.e., the number of RVs can be much smaller than that of SVs, and 2) RVM does not need the tuning of a regularization parameter (C) as in SVM during the training phase. As a drawback, however, RVM typically involves a highly nonlinear optimization process during the training phase. Fortunately, this affects the complexity of only the training phase; for our MC detection problem, the run time of a trained RVM will be much less than that of SVM.

III. RVM Classifier for Microcalcification Detection

For a given mammogram image, the MC detection process consists of the following two steps: 1) at each pixel location in the image, extract an input vector $\mathbf{x}$ to describe its surrounding image feature, and 2) apply the RVM classifier $f_{\mathrm{RVM}}(\mathbf{x})$ to decide whether $\mathbf{x}$ belongs to "MC present" class or "MC absent" class.

A. Input Feature Vector

As in the SVM classifier approach in [3], we define the input vector $\mathbf{x}$ to the RVM classifier to be formed by a small window of $M \times M$ pixels centered at the location of interest in a mammogram image. The choice of M should be large enough to cover an MC and yet small enough to avoid any interference from neighboring MCs. In the dataset used in this study, the mammograms were digitized at 0.05 mm/pixel, and $M = 15$ was chosen empirically in our experiments.

To suppress the background and thereby restrict the intra-class variations among the input samples, a high-pass filter with a narrow stop-band was applied to each mammogram image. The high-pass filter was designed to be a finite impulse response filter with cutoff frequency $w_c = 0.05$ cycles/pixel and length 10. To demonstrate its effect, we show in Fig. 3 the mammogram in Fig. 1 after filtering. As can be seen, the filtering has effectively reduced the inhomogeneity of the background.

In summary, the input vector $\mathbf{x}$ is obtained at each pixel location as follows:

$$\mathbf{x} = W[H\mathbf{f}] \tag{8}$$

where $\mathbf{f}$ denotes the entire mammogram image, H denotes the filtering operator, and W is the windowing operator. Note that for $M = 15$, the dimension of $\mathbf{x}$ is 225.

B. Determination of RVM Classifier

Before the RVM classifier $f_{\mathrm{RVM}}(\mathbf{x})$ in (2) can be applied, there are a few model parameters that need to be determined,

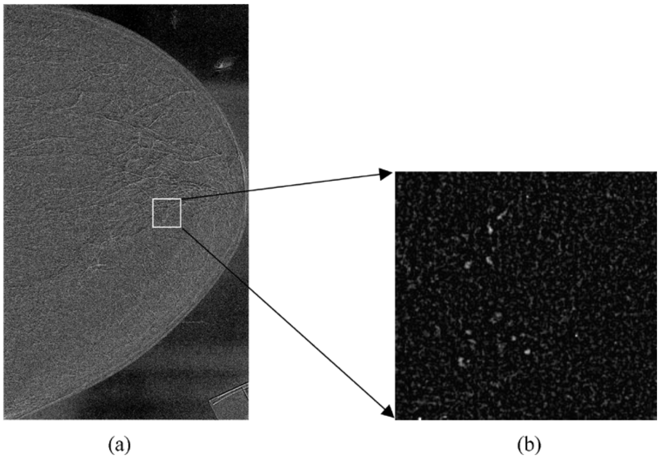

Fig. 3. The mammogram in Fig. 1 after background removal by a highpass filter.

including the type of kernel function to be used (i.e., polynomial vs RBF) and its associated parameter (i.e., the order p for the polynomial, or the kernel width σ for RBF). This is accomplished through a supervised learning procedure, in which the RVM classifier is trained using existing mammograms containing MCs (whose ground truth is known). It consists of the following two steps: 1) collect training samples $\{(x_i, d_i), i = 1, 2, \ldots, N\}$ from the existing mammograms, 2) optimize the model parameters of the RVM classifier for best performance. The details of these steps are relegated to the evaluation study (Section V).

As will be seen later (Section V), the resulting RVM classifier $f_{\mathrm{RVM}}(\mathbf{x})$ is a nonlinear function of the input vector $\mathbf{x}$. Consequently, the computational complexity of evaluating $f_{\mathrm{RVM}}(\mathbf{x})$ is directly proportional to the number of RVs. Thanks to the sparsity of $f_{\mathrm{RVM}}(\mathbf{x})$, the RVM classifier is computationally much more advantageous over the SVM classifier for MC detection (Section VI).

IV. Two-Stage RVM Classification Network

Observe that in a mammogram image the overwhelming majority of the pixels, if not all, are background (i.e., non-MC) pixels. Indeed, even in a mammogram containing clustered MCs, the total area occupied by all the MCs is typically much less than 1% of the whole mammogram. Motivated by this, we describe below a two-stage classification approach to speed up the RVM classifier further for MC detection. As illustrated in Fig. 4, we first employ a linear classifier, which is computationally much simpler, to quickly eliminate any trivial, non-MC pixels in a mammogram from further consideration; only those pixels surviving the first stage are further classified by the nonlinear RVM classifier in the second stage. Consequently, the computation time for MC detection on a mammogram can be greatly reduced.

A. First-Stage Linear Classifier

The first-stage classifier is designed to be an RVM with a linear kernel, i.e., $p = 1$ in (5)

$$K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}^T\mathbf{y} + 1). \tag{9}$$

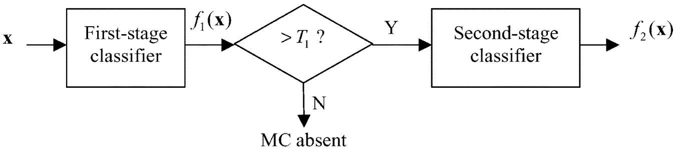

Fig. 4. Functional diagram of the two-stage RVM classification network.

Substituting (9) into (2), we obtain

$$\begin{aligned} f_{\mathrm{RVM}}(\mathbf{x}) &= \sum_{i=1}^{N} \alpha_i(\mathbf{x}^T\mathbf{x}_i + 1) \\ &= \mathbf{x}^T\left(\sum_{i=1}^{N}\alpha_i\mathbf{x}_i\right) + \left(\sum_{i=1}^{N}\alpha_i\right) \triangleq \mathbf{x}^T\mathbf{x}^* + b' \end{aligned} \tag{10}$$

where $\mathbf{x}^* \triangleq \sum_{i=1}^{N}\alpha_i\mathbf{x}_i$, and $b' \triangleq \sum_{i=1}^{N}\alpha_i$.

As can be seen from (10), the RVM classifier in this case amounts to a template matching detector, where the template $\mathbf{x}^*$ is formed by the weighted average of the RVs. In a practical implementation, the template $\mathbf{x}^*$ can be precalculated only once, and the classifier in (10) is then computed efficiently through linear convolution. The computational complexity in such a case is no longer dependent on the number of RVs.

B. Second-Stage RVM Classifier

As described above, when the feature vector $\mathbf{x}$ at a pixel location is not classified by the first-stage classifier as non-MC (i.e., passing a prescribed operating threshold T_1), it will be further classified by the nonlinear RVM classifier in the second stage.

We note that when the kernel function $K(\mathbf{x}, \mathbf{y})$ satisfies the Mercer's condition, the corresponding RVM decision function can also be written in a form of template matching in a mapped feature space. In such a case, there exists an underlying nonlinear mapping $\Phi(\mathbf{x})$ from vector $\mathbf{x}$ into a higher dimensional space $\boldsymbol{\mathcal{H}}$ such that $K(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x})^T\Phi(\mathbf{y})$. The RVM classifier in (2) can then be written as

$$f_{\mathrm{RVM}}(\mathbf{x}) = \Phi^T(\mathbf{x})\left[\sum_{i=1}^{N}\alpha_i\Phi(\mathbf{x}_i)\right] \triangleq \Phi^T(\mathbf{x})\Phi^* \tag{11}$$

where $\Phi^* \triangleq \sum_{i=1}^{N}\alpha_i\Phi(\mathbf{x}_i)$. Indeed, $f_{\mathrm{RVM}}(\mathbf{x})$ functions as a template matching detector in the mapped space $\boldsymbol{\mathcal{H}}$, in which the template is formed by the weighted average of the RVs mapped in $\boldsymbol{\mathcal{H}}$.

Based on the expression in (11), it seems that one could also precalculate the template Φ^*, then compute $f_{\mathrm{RVM}}(\mathbf{x})$ as an inner product in $\boldsymbol{\mathcal{H}}$. As in the case of the linear RVM above, the complexity in evaluating $f_{\mathrm{RVM}}(\mathbf{x})$ would no longer be dependent on the number of RVs. While seemingly plausible, such an approach unfortunately does not yield any computational saving. This is because the dimension of $\boldsymbol{\mathcal{H}}$ is typically very high, the computation of the mapping $\Phi(\mathbf{x})$ easily outweighs that of directly evaluating $f_{\mathrm{RVM}}(\mathbf{x})$ in its kernel form in (2). To demonstrate this, we consider the case of order-2 polynomial kernel as an example, i.e., $K(\mathbf{x}, \mathbf{x}_i) = (\mathbf{x}^T\mathbf{x}_i + 1)^2$. The dimension of its corresponding mapping $\Phi(\mathbf{x})$ is $N(N + 1)/2$, where N is the dimension of the input vector $\mathbf{x}$. In our MC detection problem, $N = 225$. As will be seen later (Section VI), the dimension of

$\Phi(\mathbf{x})$ is far larger than the number of RVs, making its computation even more expensive than the direct evaluation of $f_{\mathrm{RVM}}(\mathbf{x})$ in its kernel form.

V. Performance Evaluation Study

A. Mammogram Data Set

In this study, we used a set of 141 mammograms from 66 clinical cases collected by the Department of Radiology at the University of Chicago. Each mammogram had one or more clusters of MCs which were histologically proven. These mammograms were digitized with a spatial resolution of 0.05 mm/pixel and 10-bit grayscale with a dimension of 3000×5000 pixels. The MCs in each mammogram were manually identified by a group of experienced radiologists. To save computation time, a section of 900×1000 pixels, containing all the identified MCs, was cropped from each mammogram such that it was free of nontissue areas. These section images were used in our subsequent experiments. Compared with the dataset used in [3], the new dataset contains a larger number of mammograms, and the spatial resolution of the mammograms is higher. The mammogram earlier shown in Fig. 1 is an example from this dataset. For clarity, we show in Fig. 1(b) a region of interest (ROI) from this mammogram containing a cluster of MCs.

In our study, we divided the dataset in a random fashion into two separate subsets, each containing 33 cases; moreover, mammograms from the same case but of different views were always in the same subset. Subsequently, mammograms in one subset were used for training the classifiers, and mammograms in the other subset were used exclusively for testing the classifiers. Thus, mammograms from the same case were used either for training or testing, but never for both. For clarity, the mammograms in the training subset are referred to as training mammograms from this point on, and those in the testing subset are referred to as test mammograms.

B. Preparation of Training Data Set

As mentioned above, the clustered MCs in the mammograms were manually identified. To avoid any possible ambiguity associated with visual localization, we applied a template matching procedure to locate the centers of all the manually identified MCs: 1) A circularly-shaped Gaussian function with a standard deviation of nine pixels was convolved with the mammogram image; 2) The center of an MC was located to be the detection peak within a search window of 7×7 pixels centered at the manual location of the MC.

The mammograms in the training subset were found to have a total of 1291 individual MCs. For each of these MCs, a window of $M \times M$ image pixels centered at its center was extracted; the vector formed by this window of pixels, denoted by $\mathbf{x}_i$, was then treated as an input pattern to the classifier for the "MC present" class ($d_i = +1$). This yielded a total of 1291 samples for the "MC present" class. Similarly, nearly twice as many (2232, to be exact) "MC absent" samples were collected ($d_i = -1$), except that their locations were selected randomly from the set of all "MC absent" locations in the training mammograms. In this procedure no sample window was allowed to overlap with any other sample windows.

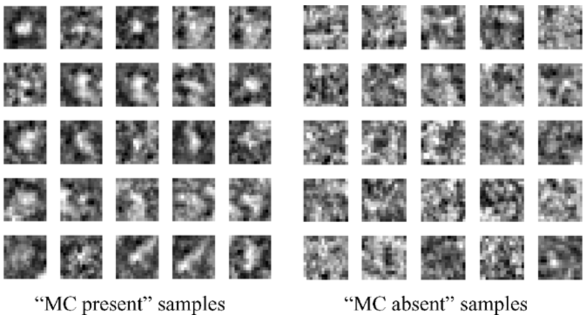

Fig. 5. Examples of 15×15 image windows of training samples from the "MC present" and "MC absent" classes. These are randomly selected from the training set.

For demonstration purpose, we show in Fig. 5 some examples of sample image windows for "MC present" and "MC absent" classes in the resulting training data set.

C. Machine Training for RVM Parameter Selection

To determine the fine-tuning parameters of the RVM classifier model for optimal performance, we apply a tenfold cross validation [14] in the training set: first, randomly divide all the training samples into ten equal-sized subsets. Second, for each parameter setting, train the classifier model ten times; during each time one of the ten subsets is held out in turn while the remaining nine subsets are used to train the classifier; the trained classifier is then used to classify the held-out subset, and the classification result is recorded. In the end, the classification results for the ten subsets are averaged to obtain an estimate of the generalization error of the classifier model. Generalization error is defined as the number of misclassified samples divided by the total number of samples classified. The parameter setting with the smallest generalization error is chosen.

Once the best parametric setting (i.e., the type of the kernel function and its associated parameter) is determined, the RVM classifier is retrained using all the available samples in the training set to obtain the final form of the decision function [i.e., the RVs and its associated coefficients α_i in (2)]. The resulting classifier is then tested using the test mammograms for performance.

D. Training of the Two-Stage RVM Network

In our experiments, the first-stage classifier in the two-stage network was trained first. This was obtained by training the RVM classifier with a linear kernel with all the samples in the training set described above.

The second-stage classifier was trained using the following procedure: First, the first-stage linear classifier was applied to classify all the training mammograms. Those background pixels misclassified by this classifier as the "MC present" class were then randomly sampled to form "MC absent" samples for the second-stage classifier. In a sense, these samples are "more difficult" to classify. In total, there were 2232 such samples collected (same as the number of "MC absent" samples in the original training set). They were used together with the 1291 "MC present" samples in the original training set to train the second-stage RVM classifier.

The rationale behind this training procedure is as follows: the second-stage classifier is meant to classify only those samples that were not eliminated by the linear RVM. Thus, the linear RVM was applied to select training samples from the training mammograms for the second-stage.

One issue associated with the two-stage network is the choice of decision threshold T_1 for the first-stage classifier. In our experiments, we used the following strategy: T_1 was selected such that on average approximately 95% of the image pixels in a mammogram were classified as "MC absent" by the first-stage classifier. This means that on average there will be approximately only 5% of the image pixels in a mammogram will be classified by the second-stage. Of course, one could use a larger value for T_1, which could lead to an even lower fraction and hence more computational saving. In our experiments, we simply used 5%, which should provide a safeguard against that a too large value for T_1 might lead to misclassification of the MC pixels by the first-stage.

E. Performance Evaluation

The performance of the RVM classifier for detection of clustered MCs is summarized using FROC curves. An FROC curve [12] plots the correct detection rate (i.e., true positive fraction) versus the average number of false-positives (FPs) per image varied over the continuum of the decision threshold. It provides a comprehensive summary of the tradeoff between detection sensitivity and specificity.

For ease of comparison, the same procedure as in [3] was used to construct the FROC curves in our experiments. For convenience, we redescribe this procedure below. First, the trained RVM classifier was applied with different thresholds to classify the pixels in turn as "MC present" or "MC absent" in each mammogram in the test set; next, the detected "MC present" pixels were grouped into potential MC objects by a morphological processing procedure described in [4] (which is run-length labeling with 8-connections), during which isolated spurious pixels were removed; and finally, MC clusters were identified by grouping the detected potential MC objects using a criterion recommended by Kallergi *et al.* [13], which was reported to yield more realistic performance than several other alternatives. Specifically, a group of objects classified as MCs is considered as a true positive (TP) cluster only if the following two conditions are met: 1) the objects are connected with nearest-neighbor distances (D_{nn}) less than 0.4 cm; and 2) at least three true MCs should be detected by the algorithm within an area of 1 cm^2. Likewise, a group of objects is labeled as an FP cluster if the objects satisfy the above conditions but contain no true MCs.

With the above clustering criterion, there were a total of 100 MC clusters identified (53 malignant, 47 benign) in the training subset of mammograms; likewise, there were a total of 110 clusters identified in the test subset of mammograms (51 malignant, 59 benign).

We emphasize that for realistic evaluation the FROC curves in this study were all computed using only the test mammograms. As mentioned before, these 74 mammograms (from 33 cases) were held aside from the beginning of the study, and were never used in any means by any of the training algorithms. This was to ensure that no positive bias was introduced in the evaluation results.

TABLE I
THE GENERALIZATION ERRORS OBTAINED BY THE RVM CLASSIFIER UNDER DIFFERENT PARAMETRIC SETTINGS

Polynomial kernel	Order 1	Order 2	Order 3	Order 4
Error rate	0.0855	**0.0489**	0.0517	0.0522
RBF kernels	Sigma=1	Sigma=2.5	Sigma=5	Sigma=10
Error rate	0.0703	**0.0526**	0.0607	0.0623

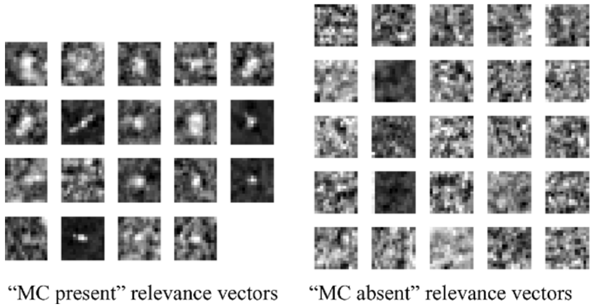

Fig. 6. Examples of 15 × 15 image windows of the RVs from the "MC present" and "MC absent" classes. All the 19 "MC present" RVs are shown and only 25 of the 46 "MC absent" RVs are shown. Note that the RVs from the two classes are distinctly different, i.e., the "MC present" RVs consist of MCs that are clearly visible, and the "MC absent" RVs consist of image windows that do not include MC-like features.

VI. EXPERIMENTAL RESULTS AND DISCUSSIONS

A. RVM Training and Model Selection

For model selection, we first trained the RVM classifier using the tenfold cross-validation procedure. We summarize the training results in Table I, where the generalization error obtained by the trained classifier is listed under different parametric settings. From these results, we see that the best error level (4.89%) was obtained by an order-2 polynomial kernel; a similar error level (5.26%) was obtained by the RBF kernel with $\sigma = 2.5$. Consequently, the order-2 polynomial kernel was chosen for the RVM classifier, which was subsequently retrained with all the samples in the training set. The resulting classifier was then used for performance evaluation.

For comparison, we also trained the SVM classifier in (7) using the same training procedure as for the RVM. The best error level (5.09%) was achieved when an order-3 polynomial kernel was used and the regularization parameter C was set to 10. As for RVM, the SVM classifier was retrained using these tuned parameters with all the samples in the training set.

For the RVM classifier, the number of RVs (produced during training) was found to be 65 (1.85% of the number of training samples); for SVM, the number of SVs was found to be 521 (14.79% of the number of training samples). Indeed, the RVM classifier is much sparser than the SVM.

To gain further insight on the RVM classifier, we show in Fig. 6 the corresponding image windows for some RVs from both "MC present" and "MC absent" classes; for comparison, we show in Fig. 7 the image windows for some SVs of the SVM classifier. As can be seen, for the RVM the RVs from the two classes are distinctly different. The "MC present" RVs consist of MCs that are clearly visible, and the "MC absent" RVs consist of image windows that do not show MC-like features at all. In a sense, the RVs are formed by "easy-to-classify" samples from both classes. In contrast, for the SVM the SVs from the two

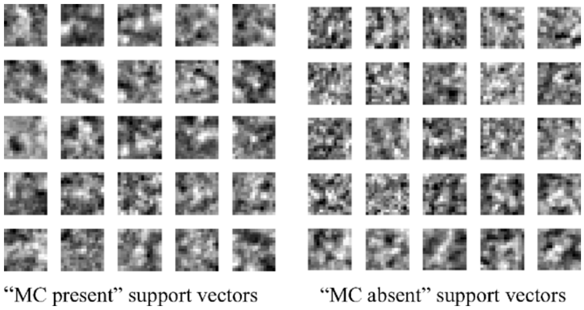

Fig. 7. Examples of 15×15 image windows of the SVs from the "MC present" and "MC absent" classes. Note that the SVs from the two classes are not distinctly different from each other, i.e., "MC present" SVs could be mistaken for "MC absent" image regions, and vice versa.

TABLE II
A. THE COEFFICIENTS α_i OF THE RVs OF THE MC-PRESENT CLASS, WITH EACH ENTRY IN THE TABLE CORRESPONDING TO THE RV AT THE SAME ROW AND COLUMN POSITION IN FIG. 6 (MC PRESENT). B. THE COEFFICIENTS α_i OF THE RVs OF THE MC-ABSENT CLASS, WITH EACH ENTRY IN THE TABLE CORRESPONDING TO THE RV AT THE SAME ROW AND COLUMN POSITION IN FIG. 6 (MC ABSENT)

-8.88×10^{-13}	1.22×10^{-6}	-2.02×10^{-6}	-5.92×10^{-8}	-4.75×10^{-13}
-4.75×10^{-13}	4.48	3.80×10^{-11}	-0.11	12.35
1.12×10^{-11}	1.94×10^{-9}	4.71	15.73	7.89×10^{-12}
11.82	6.90	0.04	3.30	

(a)

0.22	-9.49	-1.78×10^{-7}	-4.31×10^{-7}	-3.15×10^{-13}
-2.52	-5.20×10^{-4}	-6.63	-11.48	-0.48
-1.11×10^{-11}	-0.02	-0.06	-3.89	9.94×10^{-6}
-3.23×10^{-11}	-2.27×10^{-8}	-2.99×10^{-11}	-6.52×10^{-10}	-8.99×10^{-9}
8.21×10^{-7}	9.88×10^{-7}	-1.76	-1.13×10^{-8}	-2.08×10^{-4}

(b)

classes do not seem to be distinctly different, that is, the "MC present" SVs could be mistaken for "MC absent" image regions, and vice versa. These SVs are samples that appear to be "borderline," "difficult-to-classify". These results demonstrate that the two classifiers are quite different from each other, yet it is remarkable that they achieved nearly the same level of generalization error above.

In our experiments, the iteratively-reweighted least-square algorithm method developed by Tipping [9] was used for optimizing the RVM objective function $J(\alpha_1, \alpha_2, \ldots, \alpha_N)$ in (3); In the final solution, the coefficients α_i were pruned to zero when their associated variances λ_i^{-1} were less than a preset threshold (1.0×10^{-12}). For comparison, we furnish the corresponding coefficients of the RVs in Fig. 6 in Table II.

It is noted from Table II that some of the coefficients α_i are extremely small. Conceivably, these RVs could be ignored without causing much impact on the final decision function, leading to an even sparser representation (hence faster classification). Indeed, the error rate of the RVM classifier (tested on all the training samples) was found to be increased only slightly to 4.00% from 3.97% when the coefficients α_i were further pruned with a threshold of 1.0×10^{-5} (which led to 49.2% reduction in the number of RVs). In our subsequent experiments, we simply kept all the RVs without any further pruning in order not to compromise the detection performance evaluation.

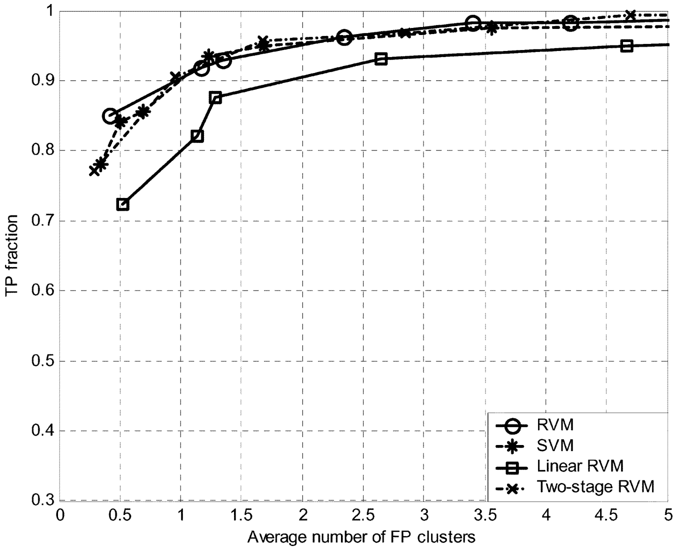

Fig. 8. FROC curves of the different methods.

B. Two-Stage RVM Network

From the results in Table I, the RVM classifier with a linear kernel achieved a generalization error rate of 8.55%. Upon retraining with all the samples in the training set, this linear RVM was used to form the first-stage classifier in the two-stage network.

For the second-stage RVM classifier the order-2 polynomial kernel was used, again based on the model selection results in Table I. This classifier is then retrained using the procedure described in Section V-D . The resulting network is referred to as "Two-stage RVM" below.

C. Performance Evaluation Results

The trained classifiers were evaluated using all the mammograms in the test subset. The test results are summarized using FROC curves in Fig. 8 for the following methods: the single RVM classifier (RVM), SVM, and Two-stage RVM. For comparison, the FROC curve is also shown for the linear RVM classifier. In our experiments, the output of the trained RVM classifier was found to be well within $[-2, 2]$. We selected a number of threshold values within this range to compute the operating points of the FROC curves.

As can be seen, the RVM classifier is nearly identical in detection accuracy to the SVM; the two-stage RVM network also achieved nearly identical performance to the RVM. However, as we will discuss later, the two-stage RVM is about 35 times faster than the SVM.

In particular, the RVM achieved a sensitivity of approximately 90% when the false positive rate is at one FP cluster on average per image. Interestingly, this sensitivity level is also similar to that achieved by the SVM in [3] (Fig. 2), where a different set of mammograms (lower spatial resolution) was used.

It is noted that the FROC curves are influenced by the criteria used in defining MC clusters. Therefore, caution should be exercised when comparing FROC results reported in the literature when they were derived using various criteria. In our experiments, we also plotted the FROC curves using different values of the nearest-neighbor-distance threshold (D_{nn}). It was

TABLE III
THE COMPUTATION TIME FOR RVM AND SVM CLASSIFIERS. THE TRAINING TIME IS THE AVERAGE AMOUNT OF TIME TAKEN DURING EACH ROUND OF TRAINING IN THE CROSS-VALIDATION PROCEDURE, AND THE TESTING TIME IS THE AVERAGE AMOUNT OF TIME TAKEN BY THE TRAINED CLASSIFIER FOR CLASSIFYING EACH IMAGE DURING THE TESTING PHASE

	RVM	SVM	Two-stage RVM
Training time (s)	2063.20	297.43	2353.67
Testing time (s)	30.04	249.33	**7.26**

observed that while the resulting FROC curves were different from those in Fig. 8, the relative ordering of the curves for the different methods was preserved; this is similar to results reported in our previous work [3]. For brevity, these results were not included here.

D. Execution Times

Finally, we show in Table III the computation times for the different methods (implemented with MATLAB on a Pentium IV 1400 MHz PC). The training time is the average amount of time taken during each round of training in the cross-validation procedure during the training phase; the testing time is the average amount of time taken by the trained classifier for classifying each image during the testing phase. Compared to SVM, the RVM classifier has reduced the detection time from nearly 250 s to about 30 s per image, nearly an order of magnitude reduction; with a two-stage network, the detection time is further reduced down to only 7.26 s, nearly 35-fold reduction compared to the SVM classifier. Note that such a dramatic saving in computation time is achieved without sacrificing the best detection performance.

In addition, for the TMNN approach (shown earlier in Fig. 2) the testing time was about 257 s per image (nearly the same as SVM); the testing times of the other methods, namely, the wavelet based method, DoG, and IDT, were in the same order as that of a linear classifier (less than 10 s). Of course, these results were dependent on the specific implementation.

Finally, we point out that the focus of this work has been on detecting individual MCs based on local image features. The detected MCs by the classifier are then grouped afterward to form clusters. Our evaluation results demonstrate that the RVM classifier is best suited for this task, thanks to its good detection accuracy and execution speed. Of course, the development of a complete system for MC cluster detection also involves consideration of several other factors, such as preprocessing the (signal dependent) noise in the mammogram images [15]–[18], optimization of the clustering criteria [13], and processing the detection results to reduce false positives [19]–[21]. For example, a curve detector was employed to remove line-like or curve-like irrelevant breast structures for further reduction of the FP rate in [21]. These factors were out of the scope of this study.

VII. CONCLUSION

We proposed an RVM technique for detection of MC clusters in digital mammograms. In this approach, an RVM classifier was trained through supervised learning to determine at each location in a mammogram whether an MC is present or not. The sparse property of the SVM classifier model yielded a decision function that was formed by a very small number of RVs. Consequently, the trained SVM classifier was not only accurate but also computationally efficient for MC detection. To further speed up the algorithm, we developed a two-stage RVM classification approach, in which the overwhelming majority, non-MC pixels were eliminated by a faster linear RVM classifier. Experimental results showed that the RVM technique could greatly reduce the computational complexity of the SVM while maintaining its best detection accuracy. This makes RVM more feasible for real-time processing of MC clusters in mammograms.

REFERENCES

[1] *Cancer Facts and Figures 1998*, American Cancer Society, Atlanta, GA, 1998.

[2] R. M. Nishikawa, "Detection of microcalcifications," in *Image-Processing Techniques for Tumor Detection*, R. N. Strickland, Ed. New York: Marcel Dekker, 2002.

[3] I. El-Naqa, Y. Yang, M. N. Wernick, N. P. Galatsanos, and R. M. Nishikawa, "A support vector machine approach for detection of microcalcifications," *IEEE Trans. Med. Imag.*, vol. 21, no. 12, pp. 1552–1563, Dec. 2002.

[4] R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schimidt, "Computer aided detection of clustered microcalcifications in digital mammpgrams," *Med. Biol. Eng. Comput.*, vol. 33, pp. 174–178, 1995.

[5] J. Dengler, S. Behrens, and J. F. Desaga, "Segmentation of microcalcifications in mammograms," *IEEE Trans. Med. Imag.*, vol. 12, no. 4, pp. 634–642, Dec 1993.

[6] R. N. Strickland and H. L. Hahn, "Wavelet transforms for detecting microcalcifications in mammograms," *IEEE Trans. Med. Imag.*, vol. 15, no. 2, pp. 218–229, Apr. 1996.

[7] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films," *IEEE Trans. Med. Imag.*, vol. 19, no. 2, pp. 115–126, Feb. 2000.

[8] M. E. Tipping, "Sparse Bayesian learning and the relevance vector machine," *J. Mach. Learn. Res.*, no. 1, pp. 211–244, 2001.

[9] M. Tipping. Sparse Bayesian Learning and the Relevance Vector Machine. Microsoft Research Tech. Rep., Cambridge, U.K.. [Online]. Available: http://www.research.microsoft.com/MLP/RVM

[10] B. Ripley, *Pattern Recognition and Neural Networks*. Cambridge, U.K.: Cambridge Univ. Press, 1996.

[11] V. Vapnik, *Statistical Learning Theory*. New York: Wiley, 1998.

[12] P. C. Bunch *et al.*, "A free-response approach to the measurement and characterization of radiographic-observer performance," *J. Appl. Eng.*, vol. 4, 1978.

[13] M. Kallergi, G. M. Carney, and J. Gaviria, "Evaluating the performance of detection algorithms in digital mammography," *Med. Phys.*, vol. 26, no. 2, pp. 267–275, 1999.

[14] K. R. Muller, S. Mika, G. Ratsch, K. Tsuda, and B. Scholkopf, "An introduction to kernel-based learning algorithms," *IEEE Trans. Neural Netw.*, vol. 12, no. 2, pp. 181–201, Mar. 2001.

[15] W. M. Morrow, R. B. Paranjape, R. M. Rangayyan, and J. E. L. Desautels, "Region-based contrast enhancement of mammograms," *IEEE Trans. Med. Imag.*, vol. 11, no. 3, pp. 392–406, Sep. 1992.

[16] W. Veldkamp and N. Karssemeijer, "Noise of local contrast in mammograms," *IEEE Trans. Med. Imag.*, vol. 19, no. 3, pp. 731–738, Sep. 2000.

[17] K. J. Mcloughlin, P. J. Bones, and N. Karssemeijer, "Noise equalization for detection of microcalcification clusters in direct digital mammogram images," *IEEE Trans. Med. Imag.*, vol. 23, no. 3, pp. 313–320, 2004.

[18] H. Li, K. J. Liu, and S. Lo, "Fractal modeling and segmentation for the enhancement of microcalcifications in digital mammograms," *IEEE Trans. Med. Imag.*, vol. 16, no. 6, pp. 785–798, Dec. 1997.

[19] A. Bazzani, A. Bevilacqua, D. Bollini, R. Brancaccio, R. Campanini, N. Lanconelli, A. Riccardi, and D. Romani, "An SVM classifier to separate false signals from microcalcifications in digital mammograms," *Phys. Med. Biol.*, vol. 46, pp. 1651–1663, 2001.

[20] H. D. Cheng, Y. M. Lui, and R. I. Freimanis, "A novel approach to microcalcification detection using fuzzy logic technique," *IEEE Trans. Med. Imag.*, vol. 17, no. 3, pp. 442–450, Jun. 1998.

[21] R. Zwiggelaar, S. M. Astley, and C. R. Taylor, "Linear structures in mammographic images: Detection and classification," *IEEE Trans. Med. Imag.*, vol. 23, no. 9, pp. 1077–1085, Sep. 2004.