Phys. Med. Biol. **49** (2004) 961–975 PII: S0031-9155(04)69324-9

A novel featureless approach to mass detection in digital mammograms based on support vector machines

Renato Campanini[1], **Danilo Dongiovanni**[1], **Emiro Iampieri**[1], **Nico Lanconelli**[1], **Matteo Masotti**[1], **Giuseppe Palermo**[1], **Alessandro Riccardi**[1] **and Matteo Roffilli**[2]

[1] Department of Physics, University of Bologna, and INFN, Bologna, Italy
[2] Department of Computer Science, University of Bologna, Bologna, Italy

E-mail: nico.lanconelli@bo.infn.it

Received 17 September 2003
Published 24 February 2004
Online at stacks.iop.org/PMB/49/961 (DOI: 10.1088/0031-9155/49/6/007)

Abstract

In this work, we present a novel approach to mass detection in digital mammograms. The great variability of the appearance of masses is the main obstacle to building a mass detection method. It is indeed demanding to characterize all the varieties of masses with a reduced set of features. Hence, in our approach we have chosen not to extract any feature, for the detection of the region of interest; in contrast, we exploit all the information available on the image. A multiresolution overcomplete wavelet representation is performed, in order to codify the image with redundancy of information. The vectors of the very-large space obtained are then provided to a first support vector machine (SVM) classifier. The detection task is considered here as a two-class pattern recognition problem: crops are classified as suspect or not, by using this SVM classifier. False candidates are eliminated with a second cascaded SVM. To further reduce the number of false positives, an ensemble of experts is applied: the final suspect regions are achieved by using a voting strategy. The sensitivity of the presented system is nearly 80% with a false-positive rate of 1.1 marks per image, estimated on images coming from the USF DDSM database.

1. Introduction

Breast cancer remains a leading cause of cancer deaths among women in many parts of the world. Mammography is considered to be the most reliable method for early detection of breast cancer. However, it could be difficult for radiologists to detect some lesions on mammograms. The missed detection may be due to the subtle nature of the radiographic findings, poor image quality, eye fatigue or oversight by the radiologists. Clinical trials and

0031-9155/04/060961+15$30.00

retrospective studies (Burhenne *et al* 2000, Birdwell *et al* 2001, Malich *et al* 2001) indicate that the detection rate can be increased with computer aided detection (CAD) systems, without any significant decrease of specificity. Masses and clustered microcalcifications are the most common lesions associated with the presence of breast carcinomas. The automatic detection of masses can be hampered by the wide diversity of their shape, size and subtlety. As is known, the tumoral masses present as thickenings, appear on images as lesions with a size ranging from 3 mm to 20–30 mm. These lesions can vary considerably in optical density, shape, position, size and characteristics at the edge. In addition, the visual manifestation in the mammogram of the shape and edge of a lesion does not only depend upon the physical properties of the lesion, but also is affected by the image acquisition technique and by the projection considered. A mass may appear round or oval, according to the projection, because other normal architectural structures of the breast could be superimposed on the lesion (in that perspective). From what has been said, it is difficult to identify morphological, directional or structural quantities that can characterize the lesions sought at any scales and any modalities of occurrence. Therefore, for a CAD system, it is very demanding to detect lesions of various types. The reason is that detection methods often rely on a feature extraction step: here, the masses are isolated by means of a set of characteristics which describe the opacities. Due to the great variety of the masses, it is extremely difficult to get a common set of features effective for every kind of mass. For this reason, many of the algorithms for detecting masses so far developed have concentrated on the detection of a particular type of mass or on masses of a specific size. Furthermore, the algorithms up to now used necessitated external information on the characteristics of the masses.

In this paper, we present a mass detection system which does not rely on any feature extraction step. Considering the complexity of the class of objects to be detected, considering that said objects frequently present characteristics that are very similar to the environment which surrounds them, and considering the objective difficulty of modelling this class of objects with few measurable quantities, in the approach proposed herein no modelling has been used. In contrast, the algorithm automatically learns to detect the masses by the examples presented to it. In this way, there is no *a priori* knowledge provided by the trainer: the only thing the system needs is a set of positive examples (masses) and a set of negative examples (non-masses). Basically, we consider mass detection as a two-class pattern recognition problem. The detection scheme codifies the image with a wavelet overcomplete representation; the great amount of information handled by the algorithm is classified by means of a support vector machine (SVM) classifier, a learning machine based on a well-founded statistical theory (Vapnik 1995, 1998). Given the ability of SVMs to handle multidimensional spaces, at the same time maintaining a good generalization capacity, the possibility of eliminating or limiting the feature extraction step for a classification task has emerged. SVMs have already been applied to breast cancer detection methods, giving very good results. In a couple of cases the SVM was used for reducing false-positive signals, in the detection of microcalcifications in mammograms (Bazzani *et al* 2001), and in the diagnosis of breast ultrasonography images (Chang *et al* 2003): in both cases SVM classified signals by means of extracted image features. A featureless approach based on SVM for the detection of lesions in mammograms has been investigated for the first time by our group (Campanini *et al* 2002). In another study, an approach similar to ours was used, but the class of object to be detected (microcalcifications) is much less heterogeneous in terms of size, shape and contrast (El-Naqa *et al* 2002). The advantages of SVM over other classifiers are that its setting is easier, it usually performs better on novel data and it is able to compress the useful information of high-dimensional spaces into a small number of elements named *support vectors*. SVMs are therefore capable of learning in sparse, high-dimensional spaces, by using very few training examples. To improve SVM

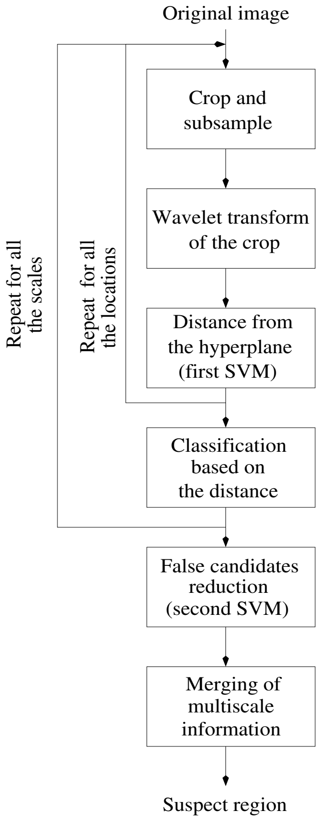
Original image
Crop and subsample
Wavelet transform of the crop
Distance from the hyperplane (first SVM)
Classification based on the distance
False candidates reduction (second SVM)
Merging of multiscale information
Suspect region
Repeat for all the scales
Repeat for all the locations

Figure 1. Chart of the detection method.

performance, a bootstrap learning technique is performed (Efron and Tibshirani 1993). We validated the detection scheme with images coming from the USF DDSM database (Heat *et al* 2000): images have a spatial resolution ranging from 43 to 50 μm and 12 bit grey-level resolution.

2. Methods

2.1. Detection scheme

Our algorithm encodes all the regions of the image in the form of vectors, these vectors being then classified as suspect or not by means of an SVM classifier. The system is virtually able to detect lesions whatever position these may occupy and at different scales in the input mammographic image; this is realized by scanning and classifying all the possible locations of the image with the passage of a window (crop). By combining the scanning pass with an iterated resizing of the window, multiscale detection is so achieved. Each crop classified as positive identifies an area judged as suspect by the CAD system.

Figure 1 shows a chart of the CAD system presented in this paper. One of the main problems encountered in mass detection is that the lesions we are searching for occur at different scales in the mammogram, typically in a range of dimensions from 3 mm to 20–30 mm. There thus emerges the problem of scanning the image at different scales. On the other hand, our system needs a fixed size crop, since the SVM classifier needs dimensionally homogeneous vectors. Consequently, the solution implemented is that of using scanning masks of different dimensions and subsampling the crops of the image extracted from that mask to a prefixed size of 64 × 64 pixels. For example, let us consider an input image of 4000 × 3000 pixels with a 50 μm pixel size and three scale targets of 32 mm (640 pixels), 16 mm (320 pixels) and 10 mm (213 pixels). The desired dimension (64 × 64 pixels) of the crop to which to apply the wavelet transform is obtained by subsampling, with a bilinear interpolation algorithm, the windows of $640 \times 640, 320 \times 320$ and 213×213 pixels to 10%, 20% and 30%, respectively. The analysis of the entire image is obtained by shifting the mask by a scanning step fixed to approximately 10% of the linear dimensions of the mask. In this way, there is a certain degree of superposition between contiguous squares. Without superposition, many lesions could fail to be detected because they are not centred on the scanning crop. This is consistent with the fact that during the training phase the positive examples are shown as crops centred on a mass. The number of analysed scales is strictly related to the range size of the masses we are interested in. A multiresolution analysis is then performed on each scaled crop, by transforming it with the Haar wavelet basis function. To exploit all the information available in the image, a redundant representation is obtained, by means of an overcomplete dictionary (Simoncelli *et al* 1992), as described in the next subsection. The number of coefficients so obtained is extremely high; these data represent the horizontal, vertical and diagonal coefficients of the considered levels in the multiresolution analysis. For each crop, the vector of coefficients is used as input for the first SVM classifier: a more complete description of SVMs is shown in a following subsection. Once trained, the SVM classifies each crop: the classification in the detection step is based upon the model created, starting from the set of examples presented during the training. For each crop, SVM gives the distance from the separating hyperplane for positive (suspect) regions. This distance is an index of confidence on the correctness of the classification: a vector classified as positive with a large distance form the hyperplane will have a higher likelihood of being a true positive as compared to a vector very close to the hyperplane, and hence close to the boundary area between the edges of the two classes. Research has been done, in order to extract *a posterior* probability from SVM outputs (Platt 1999). The scanning of all possible locations at all analysed scales provides a list of suspect candidates, each candidate consisting of a crop with a distance from the hyperplane greater than a prefixed threshold.

All the candidates are then passed to a second cascaded SVM classifier. The aim of this second SVM is to eliminate the false candidates selected by the first classifier. Some typical false signals are macrocalcifications or signals close to the pectoral muscle border. Those candidates usually have a high distance from the hyperplane, hence they survive the first SVM. However, by training a second cascaded classifier, it is possible to remove those signals. The training set for this second SVM is composed of the same positive examples (masses) used in the first training, augmented by the positive patterns of the validation set, and by the false-positive candidates obtained by the first SVM. The tasks of the two classifiers are quite different. The first SVM must have a very small error, since it has to discover true masses among a huge number of normal regions (it analyses about 100 000 crops on each image). As a consequence, even with a very small error (nearly 0.05%), it gives typically some dozens of false candidates per image. The second SVM could have a worse error, compared to the first one, since it analyses only about 50 candidates per image. However, almost all the candidates are now similar to the true lesions. Nevertheless, unlike the initial classification problem, now the classifier can focus on discarding only some particular classes of signals, according to the category of false candidates given by the first SVM. That allows the rejection of some

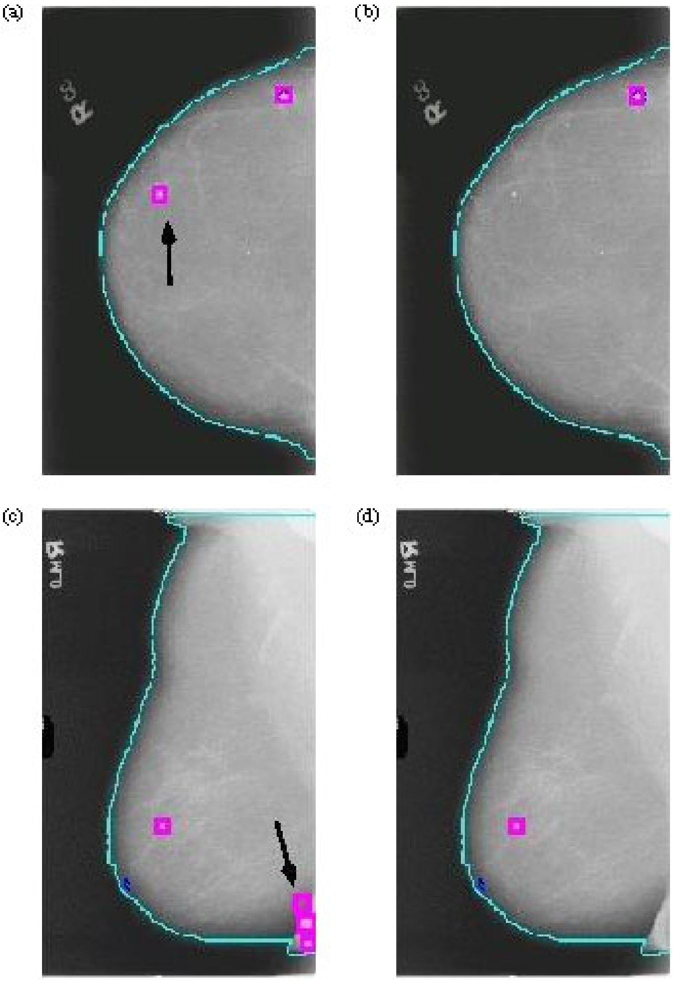

Figure 2. Suspect candidates selected by the first SVM: (a) and (c), and the candidates surviving after the second SVM classifier: (b) and (d). Here, a macrocalcification (pointed out by an arrow in (a)) and some signals near the pectoral muscle (pointed out by an arrow in (c)) are eliminated by the second classifier.

typical false signals, as mentioned above. Figure 2 displays a couple of examples of some false candidates eliminated by the second SVM classifier.

The last step of the detection scheme consists of the merging of the multiscale information. The output of the second SVM classifier is a set of candidates detected at any one of the scales. However, the same suspect region can be detected at several scales. In this case, the centres of the various candidates, representing that region at different scales, may not be the same, since the scanning step at one particular scale is diverse from the others. We fuse all the candidates within a specified neighbourhood into a single candidate. Therefore, the output of the detection method (named *expert* in the following discussion) is a list of suspect regions, each one detected at least at one scale.

To further reduce the number of false positives, we decided to apply multiple experts, and combine their output to produce the final detection. Each expert is a detector as illustrated

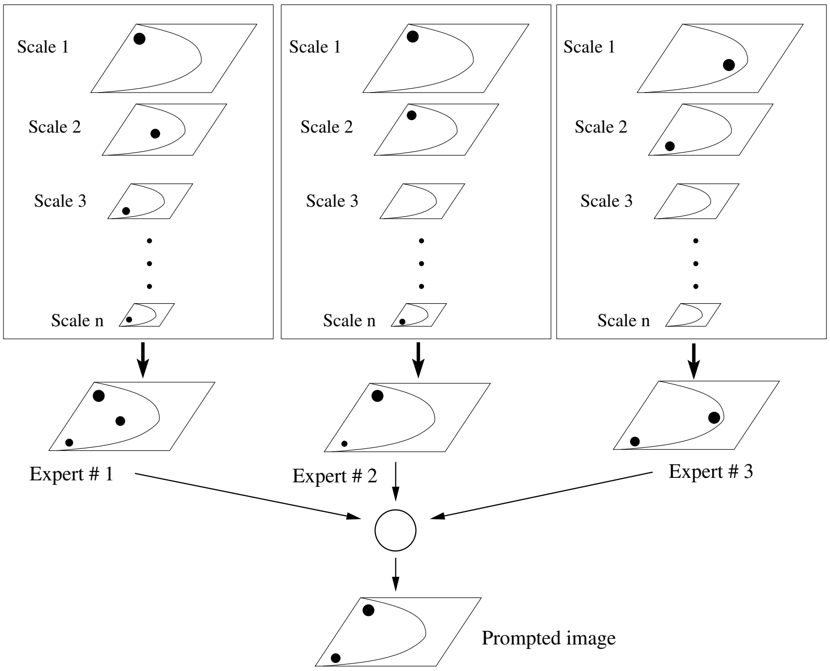

Figure 3. Committee of three experts: the prompted image consists of any overlapped suspect regions 'voted' by at least two (of three) experts. Each expert corresponds to a detection system as illustrated in figure 1, with the merging of multiscale information as described in the text.

in figure 1. The basic idea is that an ensemble of experts improves the overall performance of individual experts, if the individual experts are independent, or negatively dependent, i.e. they commit mistakes on different objects (Kuncheva *et al* 2000). Each expert differs from the others for the training sets and/or for the kernel used in the SVM classifiers (see the subsection *support vector machines*). The detection performance of the different experts can be quite close. However, because of the different kernels and because of the different training conditions, the experts will often make different errors. Hence, one way to reduce false positives could be to combine the output of the experts by ANDing them. Unfortunately, the detection rate can decrease, because a suspect region missed by only one network will be thrown out. Therefore, we chose a 'softer' combination heuristic, based on a voting strategy. Basically, a region is considered suspect only if at least two (of three) experts detect that region. For each suspect region discovered by each expert, it is checked whether there is another candidate in a neighbourhood surrounding that location. Hence, the final prompted image consists of suspect lesions detected by at least two experts, as illustrated in figure 3.

2.2. *Wavelet overcomplete*

One of the most important issues in the development of an object detection system is the representation of the object class. Wavelets offer a representation of the image that is particularly suitable for highlighting structural, geometrical and directional characteristics of the objects within the image. The wavelet coefficients encode the differences in grey levels corresponding to different regions and in different directions in the image; this encoding is performed at different scales. The idea is to provide the classifier with a complete

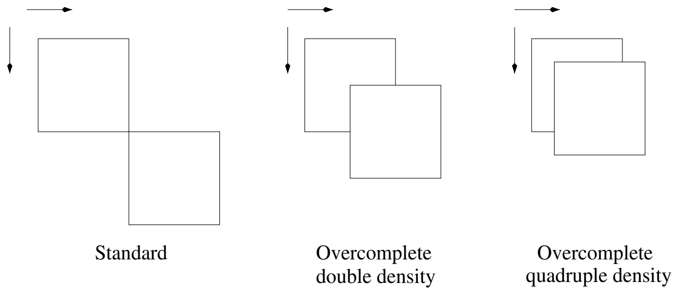

Figure 4. Difference between the standard wavelet shift, double and quadruple density overcomplete transform.

representation of the image, without guiding the generalization of the class with assumptions deriving from our modelling of the pattern. With this aim, we use an overcomplete dictionary of Haar wavelets. Haar transform is preferred, because it is the simplest and fastest one to be calculated and because it does not present problems of interpolation at the edge. The overcomplete transform provides a redundant encoding of the data with spatially superposed scale and wavelet basis functions. In this way, the information for each portion of the image is distributed over a greater number of coefficients and a richer set of characteristics. In the traditional wavelet transform, the base functions do not present any spatial superposition (they are shifted by amounts corresponding to the extent of their support). In an overcomplete scheme, according to the degree of superposition, there could be more or less redundancy in the encoding. Expressing the translation as a fraction of the support's extent, we will have single, double or quadruple density, according to whether the translation factor is equal to 1, $\frac{1}{2}$ or $\frac{1}{4}$ the extent of the support, respectively (figure 4). The wavelet transform is calculated for each of the crops produced by scanning at the various scales. In this way, for each level of decomposition, three types of coefficients are obtained, namely horizontal, vertical and diagonal. If it is assumed that the levels 4 and 6 of double-density overcomplete Haar wavelet transform are used, given an initial crop of 64×64 pixels, the total number of coefficients is nearly 3000. Therefore, for each crop, the classifier takes as input a 3000-dimensional vector. Before the classification step, this vector must be normalized, in order to ensure rapid convergence of the learning model and to balance the weights of the various characteristics. The normalization coefficients are computed during the training phase.

2.3. *Support vector machines*

In our approach, the detection of a lesion is treated as a two-class pattern recognition problem: the goal is to classify a crop window as a suspect region or not. The purpose of a classification task is to find a rule which assigns an object to one of the two classes. SVMs construct a binary classifier from a set of l training examples, consisting of labelled patterns $(\mathbf{x}_i, y_i) \in \mathbf{R}^N \rightarrow \pm 1, i = 1, \ldots, l$. The classifier aims at estimating a function $f : \mathbf{R}^N \rightarrow \pm 1$, from a given class of functions, such that f will correctly classify unseen test examples $(\mathbf{x}, y)$. An example is assigned to the class +1 if $f(x) \geqslant 0$ and to the class -1 otherwise. The test examples are assumed to be generated from the same unknown probability distribution as the training data. If no restriction is placed on the class of functions when choosing the estimate f, it could happen that even a function that performs well with training data may not generalize well to unseen examples. Thus, just the minimization of the training error (empirical risk) does not imply itself a good generalization on test examples (expected risk). In order to get

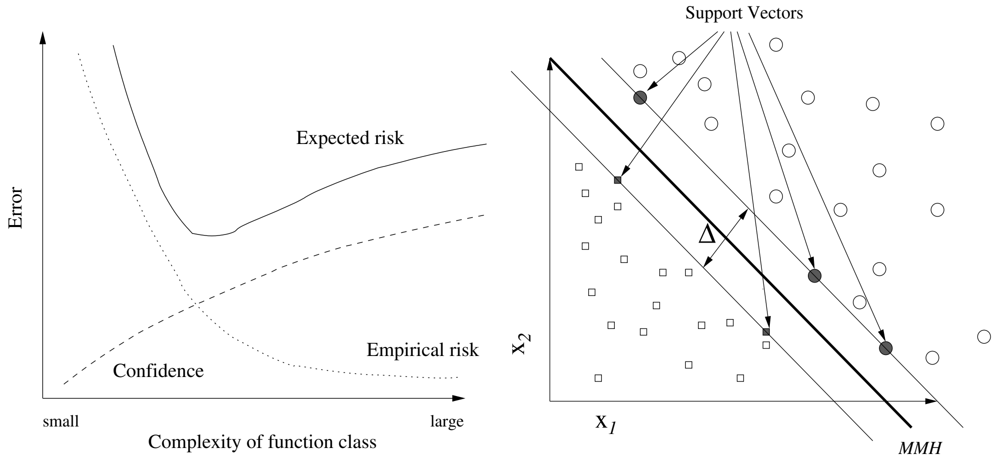

Figure 5. Left: general trend of expected risk (solid line), empirical risk (dotted line) and confidence term (dashed line). Empirical risk is related to the training error, whilst expected risk gives a measure of the generalization on test examples. The confidence term represents the upper bound on the complexity of the class function: with higher complexity the empirical error decreases, but the upper bound on the risk confidence becomes worse. In practice, the goal is to find the best tradeoff between empirical error and complexity. Right: SVM classification with the MMH that maximizes the separating margin Δ between the two classes (squares and circles). In the separable case, the support vectors are elements of the training set that lie on the boundary hyperplanes of the two classes.

good generalization performance, it is necessary to restrict the class of functions, so that f is chosen from a class with a capacity that is suitable for the amount of available training data, as determined by statistical learning theory (Vapnik 1995). This theory gives bounds for the test error; the minimization of these bounds depends on both the empirical risk and the capacity of the function class, as illustrated in the left-hand side of figure 5. One way to avoid the overfitting problem is to restrict the complexity of the chosen function class. The SVM selects hyperplanes as the class of separating functions. The optimal hyperplane is the one which maximizes the margin of the nearest examples; this is equivalent to minimizing the complexity function bound. Among all the separating hyperplanes, SVM finds the one that causes the largest separation between the decision function values for the borderline examples from the two classes. The maximal margin hyperplane (MMH) is computed as a decision surface of the form:

$$f(\mathbf{x}) = \operatorname{sgn}\left(\sum_{i=1}^{l} y_i \alpha_i (\mathbf{x} \cdot \mathbf{x}_i) + b\right) \tag{1}$$

where the coefficients α_i and b are calculated by solving the following quadratic programming problem:

$$\begin{cases} \text{maximize} \quad \displaystyle\sum_{i=1}^{l} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{l} \alpha_i \alpha_j (\mathbf{x}_i \cdot \mathbf{x}_j) y_i y_j \\ \text{with} \quad \displaystyle\sum_{i=1}^{l} \alpha_i y_i = 0 \qquad 0 \leqslant \alpha_i \leqslant C. \end{cases} \tag{2}$$

C is a regularization parameter selected by the user; it determines the tradeoff between the empirical error and the complexity term. Therefore, the classification of a pattern $\mathbf{x}$ is achieved according to the values of $f(\mathbf{x})$ in (1).

It is worth mentioning that in a typical classification problem the hyperplane (1) is determined by only a small fraction of training examples. These vectors, named *support vectors*, are those with a distance from the MMH equal to half the margin (figure 5, right). In the more general case in which the data are not linearly separable in the input space, a nonlinear transformation is used to map the input vectors into a high-dimensional space. In this space, the MMH will be determined as mentioned above. The kernel function guides the nonlinear mapping: admissible and typical functions are the polynomial and the Gaussian kernels. In our work, we utilized both the usual second degree polynomial and a *sparse* polynomial kernel (Schölkopf *et al* 1998). By using a common polynomial kernel $k(\mathbf{x}, \mathbf{y}) = (\mathbf{x} \cdot \mathbf{y})^d$, one implicitly constructs a decision boundary in the space of all possible products of d pixels. This may not be desirable, since in natural images, correlations over short distances are much more reliable as features than long-range correlations are. The general form of the sparse polynomial is the following:

$$k(\mathbf{x}, \mathbf{y}) = \left(\sum_{\text{patches}} \left(\sum_{i \in \text{patch}} x_i \cdot y_i + 1 \right)^{d_1} \right)^{d_2} \tag{3}$$

where the patches are a sequence of overlapping small crops within the crop to be classified, d_1 and d_2 set the degree of the *intra*-patch and of the *inter*-patch combinations, respectively. The *intra*-patch combinations consist of all the products of the pixels within the patch up to degree d_1, whereas the *inter*-patch ones consist of all the products of the *intra*-patch combinations up to degree d_2. Therefore, with the sparse kernel we can approximately set the range of local and global correlations, allowing the rejection of the majority of useless pixel combinations.

2.4. Training

The training of the system is obtained by presenting a set of 64 × 64 pixel windows containing masses (positive examples) and a set of crops without lesions (negative examples): this combined set forms the initial training database. Each positive example is a portion of a digital mammographic image containing a mass; the mass is contained completely within the square. The size of the positive crops is chosen as follows: the ratio between the crop area and the area of the mass core should be nearly 1.3. In this way, all the positive examples are characterized by having about 30% of background and 70% of the area taken by the mass. As a consequence, the real size of the masses is smaller than the size of the searching scale (e.g. a scale with a 40 mm crop is appropriate for searching for masses of 35 mm). The classifier is then trained to recognize as positive a vector corresponding to a square centred on a lesion. Each negative example has no superposition with any positive crop, since negative crops are extracted from normal cases, whilst positive patterns come from malignant cases. Figure 6 shows some patterns used in the training of the first classifier. Training for a mass detection task is challenging because of the difficulty in characterizing the ‘non-masses’ examples: indeed, whilst the positive examples are quite well defined, there are no typical negative examples. To overcome the problem of defining this extremely large negative class, a bootstrap technique is used: after the initial training, the system is retrained, by using a new set containing some misclassified false-positive examples. Those examples are obtained from the detection of images not present in the initial training set. This procedure is iterated until an acceptable performance is achieved. In this way, the system is forced to learn by its own errors.

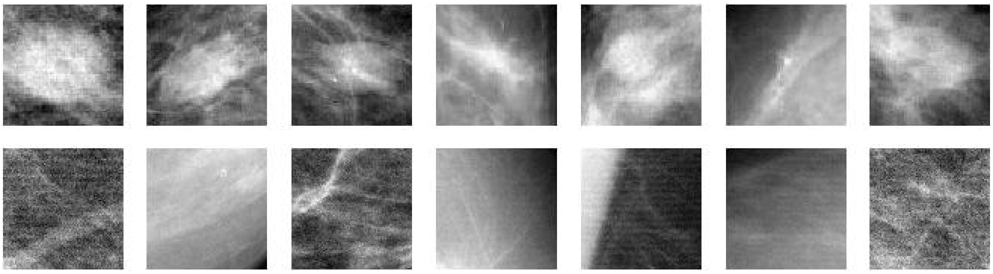

Figure 6. Examples of training crops: 'masses' (top row) and 'non-masses' (bottom row).

The second SVM classifier has been trained on the same positive examples used for the first SVM, augmented by the positive patterns of the validation set, and by the false-positive signals detected by the first SVM. With this aim, we performed the detection both on the training images and on unseen images (training and validation set). The validation step aims at choosing the best SVM architectures and the best wavelet representation. The system is then tested on new unseen mammograms (test set).

2.5. Materials

The dataset used is part of the digital database for screening mammography (DDSM) database collected by the University of South Florida, and freely available on the net at http://marathon.csee.usf.edu/Mammography/Database.html. The entire database consists of more than 2500 cases, divided among benign, malignant and normal cases. Images containing suspect areas have associated ground truth information about the locations and types of those regions. We selected images digitized with a Lumisys laser film scanner at 50 μm and a Howtek scanner at 43.5 μm pixel size. All the images have a 12-bit grey-level resolution. The normal cases have been used both for estimating the false-positive rate and for providing the system with the negative examples during the training phase. Positive examples were extracted from malignant cases, among about 800 images containing masses. The different training sets used for the various experts are subsets of these images. A total of 512 images have been used for test: 312 malignant cancers from volume 'cancer_02', 'cancer_07' and 'cancer_12', and 200 normal images from volume 'normal_08' and 'normal_10'. For each volume, all cases are included, except those cases where the microcalcifications were the only visible sign. For each case, four mammograms are present: the cranio-caudal and the medio-lateral projections of left and right breast. Nevertheless, for malignant cases we used only images containing masses, excluding their contralateral views, if no masses were present there. In a few cases, the lesion was visible in only one view. In addition, some cases present more than one mass per view. Table 1 summarizes the composition of the database used. It is worth noting that our definition of *case*, as described in table 1, is significant only for the estimation of the *per-case* performance, as we will see in the *Results* section. Figure 7 shows the distribution of lesion subtlety for training and test images, as ranked by radiologists who evaluated each individual mass.

3. Results

With our CAD system we are searching for masses with a size smaller than 35 mm, therefore we performed a multiscale detection, by using the following eight scales: 8, 10, 13, 17, 22, 27, 33 and 40 mm. The three experts, named A, B and C, consist of classifiers based on a

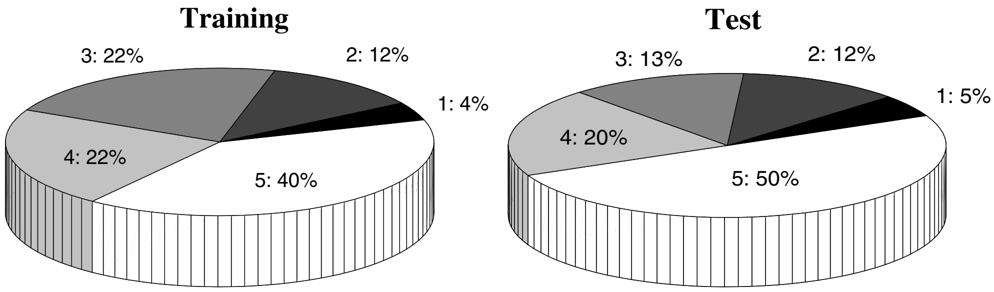

Figure 7. Graphs representing the subtlety distribution of the lesions used in training (left) and test (right). Subtlety data are considered according to BI-RADS description (1: subtle lesion, 5: obvious lesion).

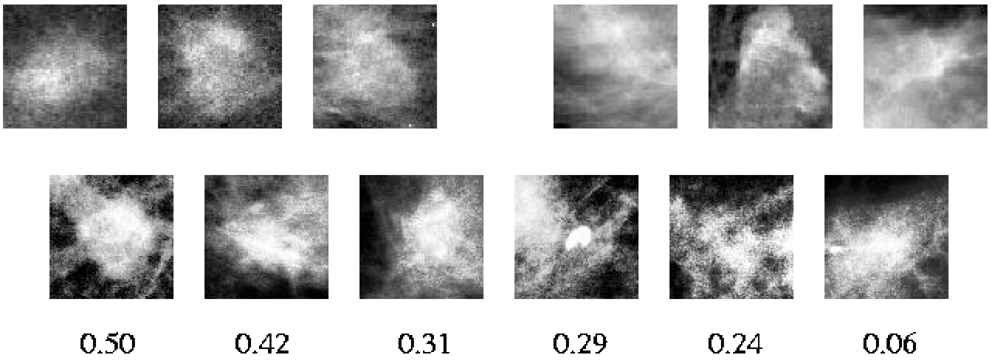

Figure 8. Top row: examples of positive (left) and negative (right) support vectors. Bottom row: positive classified crops with their distance from MMH; the three crops on the left are true masses, whilst the three crops on the right are false-positives.

Table 1. Summary of the composition of the database used: numbers of masses, images and cases in training and test for cancer and normal patients. Some images contain more than one visible mass.

	Training		Test	
	Malignant	Normal	Malignant	Normal
Masses	900	/	327	/
Images	800	600	312	200
Cases	420	150	144[a]	50

[a] For the estimation of the *per-case* system performance, a case is defined as a patient with visible masses in at least two views, only for malignant lesions.

polynomial kernel of second degree (A and C) and a sparse polynomial kernel of second degree (B). Experts A and C differ in the patterns presented during the training. Within each expert, the same kernel function has been used, both for the first and the second SVM classifiers. The committee ensemble prompts regions, when at least two experts have detected a signal in a 7 mm neighbourhood.

One of the main advantages of SVMs is that they are able to compress the useful information gained in the training into a reduced set of patterns, named support vectors. Figure 8 shows an example of the support vectors obtained with one of the classifiers used in

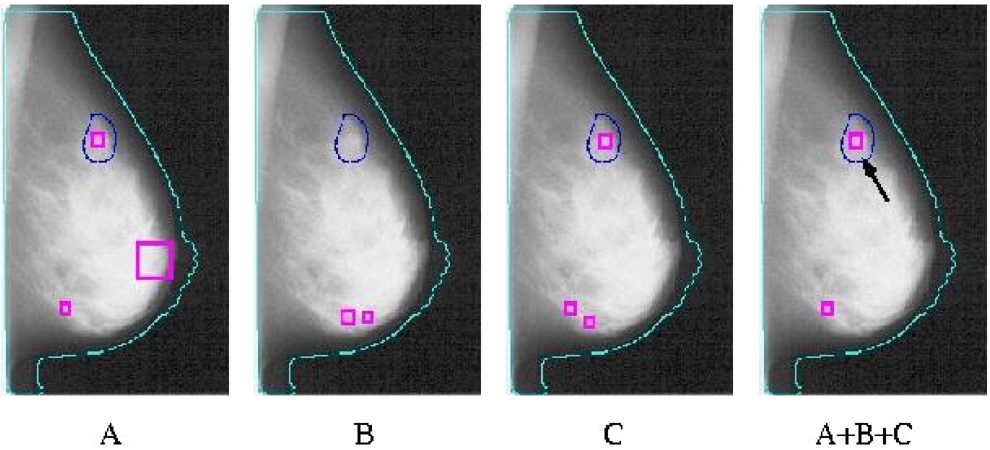

Figure 9. Example of reduction of false signals due to the combination of experts. Here, the true mass (pointed out by the arrow) survives, whereas all but one false-positive is rejected, thanks to the voting strategy. The different size of the prompts is related to the scale which detects that lesion.

the present study. It is worth remarking that the positive (masses) and negative (non-masses) support vectors are very similar to each other, by confirming the fact that the support vectors are those examples near the separation of the two classes. The same figure also illustrates some vectors classified as positive, with their distance from MMH. These crops can be both true masses and false-positive detections. We can note that higher distance from MMH means a more evident lesion. The combination of several experts aims at reducing the number of false positives. Figure 9 shows an example that explains this fact. Here, the true mass (pointed out by the arrow) is detected, since it is discovered by experts A and C. At the same time, one false positive survives, while all other false signals are rejected, thanks to the voting strategy.

We evaluated the performance of the detection method, by means of FROC curves. An FROC curve is a plot of the detection rate versus the average number of false-positive marks per image. An FROC curve provides a summary of the trade-off between sensitivity and specificity. We put a threshold for each expert on the maximum number of signals to be considered for the committee fusion. Basically, the suspect candidates are ranked, according to their distance from the MMH of the second SVM; then, we decided to keep only the best *n* signals. In fact, the distance from the hyperplane is an index of confidence of the likelihood of a candidate being malignant. The different points of the FROC curve are obtained by varying *n* for each expert. A region is defined as true if its centre falls within the ground-truth annotations, otherwise it is considered as a false-positive. The false-positives are computed using normal cases only. The performance results are presented on a *per-mammogram* and a *per-case* basis. In the former, the cranio-caudal and medio-lateral oblique views are considered independently. In the latter, a mass is considered discovered if it is detected in either one of the views. Here, we considered only cases where at least two views per patient contain masses. The *per-case* evaluation takes into consideration that, in clinical practice, once the CAD alerts the radiologist to a cancer on one view, it is unlikely that the radiologist will miss the cancer. Our scoring method considers all the malignant masses on a mammogram (or in a case) as a single true-positive finding. The rationale is that a radiologist may not need to be alerted to all malignant lesions in a mammogram or case before taking action. Anyway, the great majority of cases (more than 95%) present just one mass per view, so practically this method

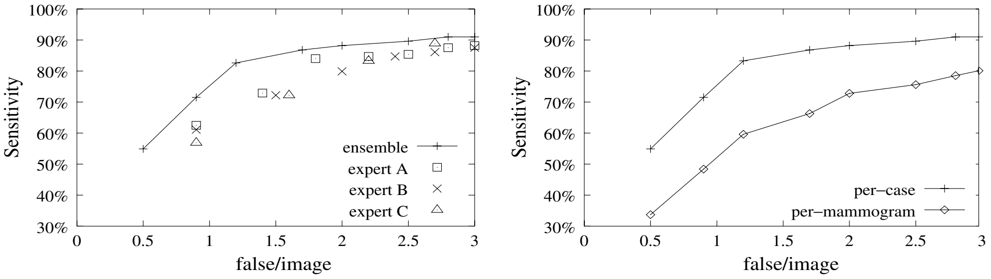

Figure 10. Performance of our detection scheme on the test images: FROC *per-case* of the single classifiers and of the committee ensemble (left) and FROC of the committee ensemble *per-mammogram* and *per-case* (right).

gives the same results as if we consider each mass on a mammogram or in a case as a different true-positive finding.

Figure 10 shows the performance of our CAD system on the 512 test images. On the left-hand side of the figure, the *per-case* performance of the three separate experts and of the committee is depicted. The combination of several experts gives a clear improvement, especially in the most important range of less than 1.5 false-positives per image. Once again, that confirms the reduction of false alarms gained, thanks to the experts fusion. On the right-hand side of figure 10, the *per-mammogram* and *per-case* outcomes of the committee ensemble are shown. Results are promising, especially if we consider that those images contain lesions of different sizes and types: oval, circumscribed, speculated masses and architectural distortions. We recall here that the system has been trained on lesions characterized by a dense core, centred on the crop, and with a crop/core area ratio of about 1.3. The test set also contains some type of lesions very different from the training patterns, such as architectural distortions or masses close to the chest border. The performance on the test images clearly indicates the effectiveness of the presented system in detecting breast masses. Our results seem comparable with others obtained on the same database (Heat and Bowyer 2000, Petrick *et al* 2002), even if we are aware that care must be taken when comparing different results. Indeed, several factors affect the performance, such as the characteristics of test images (e.g. lesion subtlety, size, etc), and the strategy for the estimation of true and false-positive detection (Kallergi *et al* 1999). For instance, Petrick *et al* defined as true-positive a lesion with an overlap between the bounding box of the detected object and the bounding box of a true mass greater than 25%. Instead, Heat and Bowyer adopted the same policy as us for the computation of the true-positives. On the other hand, they counted false-positives on cancer images, whereas Petrick *et al* decided to count them on normal images. We have also investigated the characteristics of the masses missed by our system. At a false rate of 1.2 false-positive marks per image, we miss 24 cases: seven of them represent patients with an architectural distortion as the only visible sign, four of them have masses bigger than 3.5 cm and another three cases present masses very close to the chest border on both views. By analysing these outcomes, we can state that it is reasonable that the system overlooks those lesions for the following reasons: first of all, in the detection step we focus our attention on masses with a size smaller than 35 mm, since we performed a multiscale detection with searching scales up to 40 mm (recall the area ratio factor mentioned above). A CAD system must detect small lesions, to be helpful for an early diagnosis. Therefore, we reckon that missing very big masses is a little sin for a CAD software, since radiologists will not miss them for sure. On the other hand, our system has never seen examples of architectural distortion or of masses close to the chest border during the training. We make the decision to choose training patterns consisting of masses with a well-defined core and centred on the crop, in order to facilitate the task of the SVM classifier. A possible improvement of the present work could be the training of other experts, each one focusing on a particular kind of lesion neglected in this version. Thus, we could have experts for the detection of architectural distortion, experts for masses close to the chest border, and so on. We are confident that this will further improve our performance.

4. Conclusion

The main goal of this paper is to show the feasibility of a novel featureless approach for the detection of masses in digital mammography. The use of SVM classifiers has allowed us to manage the great amount of information provided by the wavelet overcomplete representation. Cascade classifiers and combination of different experts have been adopted, in order to reduce the number of false-positive detections.

Results obtained on the difficult USF DDSM database are very promising: 80% of cancers detected with 1.1 false marks per image. It is worth remarking that our procedure automatically extracts the useful information directly from the images, without needing an external set of features for classifying the suspect regions. The only thing the system needs is a set of positive examples (masses) and a set of negative examples (non-masses).

Future research could be done, in order to improve the performance of the algorithm. First, we could train other experts, each one focusing on a particular kind of lesion. In addition, we are planning to test the CAD system on images coming from a full-field digital mammography apparatus.

Acknowledgment

This work is partially supported by 'Fondazione del Monte di Bologna e Ravenna'.

References

Bazzani A, Bevilacqua A, Bollini D, Brancaccio R, Campanini R, Lanconelli N, Riccardi A and Romani D 2001 An SVM classifier to separate false signals from microcalcifications in digital mammograms *Phys. Med. Biol.* **46** 1651–63

Birdwell L, Ikeda D M, O'Shaughnessy K F and Sickles E A 2001 Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection *Radiology* **219** 192–202

Burhenne L J W, Wood S A, D'Orsi C J, Feis S A, Kopana D B, O'Shaughnessy K F, Sickles E A, Tabar L, Vyborny C J and Castellino R A 2000 Potential contribution of computer-aided detection to the sensitivity of screening mammography *Radiology* **215** 554–62

Campanini R, Bazzani A, Bevilacqua A, Bollini D, Dongiovanni D N, Iampieri E, Lanconelli N, Riccardi A, Roffilli M and Tazzoli R 2002 A novel approach to mass detection in digital mammography based on support vector machines (SVM) *Digital Mammography: IWDM2002, 6th Int. Workshop on Digital Mammography* ed H O Peitgen (Berlin: Springer) pp 399–401

Chang R F, Wu W J, Moon W K, Chou Y H and Chen D R 2003 Support vector machines for diagnosis of breast tumors on US images *Acad. Radiol.* **10** 189–97

Efron B and Tibshirani R J 1993 *An Introduction to the Bootstrap* (London: Chapman and Hall)

El-Naqa I, Yang Y, Wernick M N, Galatsanos N P and Nishikawa R M 2002 A support vector machine approach for detection of microcalcifications *IEEE Trans. Med. Imaging* **21** 1552–63

Heat M and Bowyer K 2000 Mass detection by relative image intensity *Digital Mammography: IWDM2000 5th Int. Workshop on Digital Mammography* ed M J Yaffe (Madison, WI: Medical Physics Publishing) pp 219–25

Heat M, Bowyer K, Kopans D, Moore R and Kegelmeyer P 2000 The digital database for screening mammography *Digital Mammography: IWDM2000 5th Int. Workshop on Digital Mammography* ed M J Yaffe (Medical Physics Publishing) pp 212–8

Kallergi M, Carney G M and Gaviria J 1999 Evaluating the performance of detection algorithms in digital mammography *Med. Phys.* **26** 267–75

Kuncheva L I, Whitaker C J, Shipp C A and Duin R P W 2000 Is independence good for combining classifiers? *Proc. 15th Int. Conf. on Pattern Recognition (Barcelona, Spain)* vol 2 pp 168–71

Malich A, Marx C, Facius M, Bohem T, Fleck M and Kaiser W A 2001 Tumour detection rate of a new commercially available computer-aided detection system *Eur. Radiol.* **11** 2454–9

Petrick N, Sahiner B, Chan H P, Helvie M A, Paquerault S and Hadjiiski L M 2002 Breast cancer detection: evaluation of a mass-detection algorithm for computer aided diagnosis—experience in 263 patients *Radiology* **224** 217–24

Platt J C 1999 Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods *Advances in Large Margin Classifiers* A Smola, P Bartlett, B Schölkopf and D Schuurmans (Cambridge, MA: MIT Press) pp 61–74

Schölkopf B, Simard P, Smola A and Vapnik V 1998 Prior knowledge in support vector kernels *Advances in Neural Information Processing Systems* vol 10 ed M I Jordan, M J Kearns and S A Solla (Cambridge, MA: MIT Press) pp 640–6

Simoncelli E P, Freeman W T, Adelson E H and Heeger D J 1992 Shiftable multi-scale transforms *IEEE Trans. Inform. Theory* **38** 587–607

Vapnik V 1995 *The Nature of Statistical Learning Theory* (Berlin: Springer)

Vapnik V 1998 *Statistical Learning Theory* (New York: Wiley)