

A novel approach to mass detection in digital mammography based on Support Vector Machines (SVM).

Renato Campanini¹, Armando Bazzani¹, Alessandro Bevilacqua², Dante Bollini¹,
Danilo Dongiovanni¹, Emiro Iampieri¹, Nico Lanconelli¹,
Alessandro Riccardi¹, Matteo Roffilli³ and Roberto Tazzoli¹

¹ Physics Dpt., University of Bologna, Viale Berti-Pichat 6/2, 40127 Bologna, Italy

² DEIS, University of Bologna, Viale Risorgimento 2, 40136 Bologna, Italy

³ Computer Science, University of Bologna, Via Sacchi 3, 47023 Cesena, Italy
nico.lanconelli@bo.infn.it

Abstract. In this paper we present a novel approach to mass detection in digital mammograms. The great variability of the masses appearance is the main obstacle of 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 decide not to extract any feature, for the detection of the region of interest; on the contrary we exploit all the information available on the image. No a priori knowledge and no appearance model are used. A multiresolution overcomplete wavelet representation is achieved, in order to codify the image with redundancy of information. The vectors of the very-large space obtained are classified by means of an SVM classifier. Training, validation and test are accomplished on images coming from USF DDSM database. The sensitivity of the presented system is 84% with a false-positive rate of 3.1 marks per image.

1. Introduction

The incidence of breast cancer among women has been increasing during last years and this cancer is one of the leading cause of death in civilized countries. Mammographic screening programs result in earlier detection of breast lesions, thus permitting earlier treatment. 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. It is indeed very difficult for a CAD system to discover all types of lesion; 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 helpful for every kind of masses.

In this paper we present a mass detection system which does not rely on any feature extraction task; on the contrary, the algorithm automatically learns to detect the masses by the examples presented to it. In this way there is no *a priori* knowledge

needed by the trainer: the only thing the system wants is a set of positive examples (masses) and a set of negative examples (non-masses). 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 [1]. SVMs have already been applied to microcalcifications detection methods [2], giving rise to very good results. 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 extract useful information in high-dimensional spaces. To improve SVM performance a bootstrap learning technique is performed. We validated the detection scheme with images coming from the USF DDSM database: images have a spatial resolution ranging from 43 to 50 μm and 12 bit gray-level resolution.

2. Methods

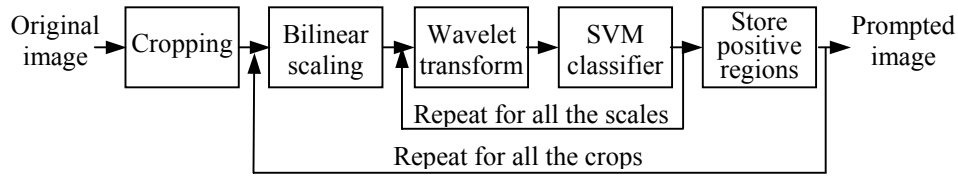


Fig. 1. Scheme of the detection method.

Figure 1 shows an overview of the CAD system presented in this paper. Since the system needs a fixed size crop (in our case 64×64 pixels), in order to detect masses of different size we first chose a window from the original image and then scale this window to a 64×64 crop, by means of a bilinear interpolation. The number of analyzed scales is strictly related to the range size of the masses we are interested in. A multiresolution analysis is then performed on the 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 [3]. In the traditional wavelet transform, the wavelets do not overlap; they are shifted by the size of the wavelet support. An overcomplete transform allows the achievement of a richer set of features, with a shift equal to $\frac{1}{2}$ (or $\frac{1}{4}$) of the size of the support of each wavelet. Unfortunately the number of coefficients obtained is extremely high: in our case more than 14000. These data represent the horizontal, vertical and diagonal coefficients of the considered levels in the multiresolution analysis. For each crop, we have a 14000 dimensional vector, which is used as input for the SVM classifier. SVMs are capable of learning in sparse, high-dimensional spaces, by using very few training examples. Once trained, the SVM classifies each crop and positive (suspect) regions are then stored and prompted. The system is able to detect lesions of different size; this is achieved by the shifting of a 64×64 window over the entire image combined with the scaling of the image, to get a multi-scale detection.

The training of the system is obtained by presenting a set of 64×64 pixels windows containing masses (positive examples) and a set of crops without lesions (negative examples): this combined set forms the initial training database. While the positive examples are 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 with false-positive examples added to the negative set. 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.

3. Results and conclusion

The CAD system has been tested on images coming from USF database; about 600 images have been used: 300 for training, 150 for validation and 150 for test. In figure 2 there is shown the performance of the detection algorithm on the test images. Results are promising, especially if we consider that those images contain masses of different types: oval, circumscribed, spiculated. It is also 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.

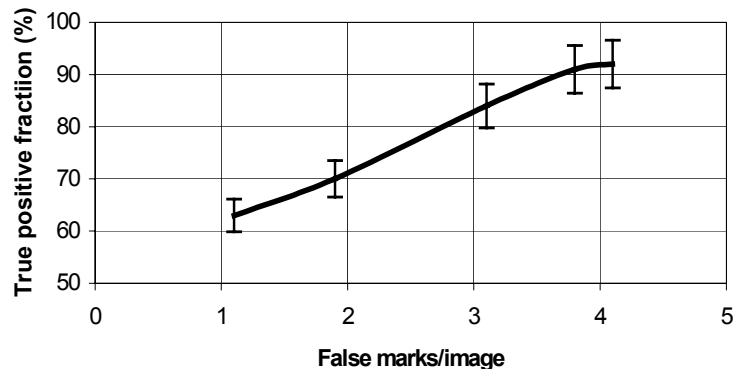


Fig. 2. FROC of the CAD system.

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

References

1. Vapnik, V.: The nature of Statistical Learning Theory. Springer-Verlag, 1995
2. Bazzani, A., Bevilacqua, A., Bollini, D., Brancaccio, R., Campanini, R., Lanconelli, N., Riccardi, A., Romani, D.: An SVM classifier to separate false signals from microcalcifications in digital mammograms. *Physics in Medicine and Biology*, 46 (6), (2001) 1651-1663
3. Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. *Computer Vision and Pattern Recognition*, 1999, 193-199