# A Novel Method of Extracting and Classifying the Features of Masses in Mammograms

Han Zhen-zhong[1], Liu Pei-guo[1], Mao Jian[1,2]
1.National University of Defense Technology, Changsha, China.
2.Computer Engineering College, Jimei University, Xiamen, China

***Abstract*—Some improvements in the classification of masses in the breast are proposed in this paper. First, for the purpose of enriching the information concerning the shape of the mass, a new morphological feature is extracted. Then, the textural features of the region of interest (ROI) are extracted by combining the undecimated wavelet transform (UWT) and the gray level co-occurrence matrix (GLCM). Finally, based on the geometrical and textural features, the feature-weighted support-vector machine (FWSVM) is used to distinguish between malignant and benign masses. The experiments implemented on the public digital database for screening mammography (DDSM) indicated that the proposed improvements can achieve better results than the existing methods.**

***Keywords—mass classification; undecimated wavelet transform geometrical features; gray level co-occurrence matrix; textural features; feature weighted support vector machine***

## I. INTRODUCTION

Breast cancer is one of the major causes of cancer-related deaths in women [1-3]. According to statistics published by the World Health Organization (WHO), 458,000 women died from breast cancer, more deaths than were caused by any other cancer in women [4]. At the present time, the cause of breast cancer is unknown, so there is no way to prevent it. Early detection and treatment are considered as the most promising approaches for reducing the mortality associated with breast cancer [5]. Mammography is the main imaging methodology used to diagnose breast cancer due to its high sensitivity and resolution, which aid early detection [6]. When implementing a computer-aided detection system (CAD), one of the main points that should be considered very seriously is the selection of appropriate features that could be used to distinguish between benignancy and malignancy.

Modern scanners have much higher resolution than earlier devices, which has improved the ability to describe some obvious differences between the benignancy and malignancy of a tumor. Benign lesions form a pseudocapsule that prevents the tumor from invading the surrounding normal tissues. Therefore, benign masses always have well-defined contours with rounder, smoother shapes. By comparison, the contours of malignant tumors are poorly defined and irregular [7]. For these reasons, most mass classification methods are based on the shape and boundary of the mass. Shape-based classification methods have good ability for differentiating between typical benign and malignant tumors. However, these methods require high accuracy of mass segmentation. Unfortunately, medical images, including mammograms can have significant noise, and the boundary of the mass often is ambiguous (especially for a malignant mass). These two factors always result in inaccurate segmentation of the mass. Under this condition, the credibility of these classification methods drops dramatically. Furthermore, mistakes often appear when these methods are used to process malignant masses that have regular boundaries and the benign masses that do not have smooth boundaries.

Using only the contour of a mass to describe its features is not a comprehensive approach because no information is available to determine the inside characteristics of the mass. The structures of the cells in malignant and benign masses are different. For benign tumors, the gray level obviously changes between the internal and external parts of the mass. In contrast, malignant tumors have no obvious boundaries, and the edge gray changes gradually because of infiltrating growth [8]. Hence, the textural features of masses also can be used to classify them. Most systems that are designed to extract textural features are based on statistical analysis, mathematical modeling, and signal processing [5]. The characterization obtained from the traditional gray level co-occurrence matrix (GLCM) is not sufficiently detailed to make a judgment, so its classification performance is not good enough to meet the requirements. Likewise, since the conventional wavelet decomposition methods only use the statistical characteristics of the sub-bands, no description of the sub-bands' structure information is available. Therefore, these methods also cannot be used to make definitive judgments with poor quality images.

For these reasons, a novel method that can extract and classify the essential features of the mass appropriately is proposed in this paper. First, the proposed method extracts a new morphological feature to enrich the information about the shape of the mass. Then, the textural features of the region of interest (ROI) are obtained by combining the undecimated wavelet transform (UWT) and GLCM. Finally, by using these characteristics, including geometrical and textural features, the feature weighted support vector machine (FWSVM) is used to distinguish between malignant and benign masses.

The rest of this paper is organized as follows. Section 2 gives a brief introduction of the features that are extracted for classification, and the FWSVM classifier is discussed in section 3. Results and discussion are presented in section 4, and section 5 presents our conclusions and recommendations for future work.

## II. EXTRACTION OF THE FEATURES OF MASS

For a classifier, one of the most important things that should be taken into consideration is obtaining appropriate features. In this paper, both geometrical and textural features are used. It should be pointed out that the ROI and the results of segmentation of the mass were extracted based on previous work [9, 10].

### *A. Geometrical features*

When the segmentation of the mass is completed, the boundary of the mass can be identified. Therefore, a set of features related to the geometrical features of the boundary of

the mass can be calculated. The first eight parameters that were detailed in [8, 11] are listed below, and the ninth parameter is described below. (Since some of them are utilized to calculate other features, not all of them are applied directly.)

$f_{g1}$: The circularity of the segmented mass

$f_{g2}$: Normalized mean radius of the segmented mass

$f_{g3}$: Normalized standard deviation of the radii

$f_{g4}$: Boundary roughness

$f_{g5}$ : The aspect ratio of the segmented mass

$f_{g6}$: Eccentricity of the ellipse for which its second order central moment is the same as the segmented result

$f_{g7}$: Compactness of the segmented mass

$f_{g8}$: Zero crossing of the boundary

$f_{g9}$: number of bifurcation points in the skeleton drawing of the segmented region, which indicates the skeleton complexity of the region. Compared to benign cases, the contours of malignant masses are poorly defined and irregular [7]. Thus, malignant masses always have more bifurcation points than benign masses. A graphical description is provided in Fig. 1.

(a) (b) (c)

Fig. 1. Graphical description of skeleton complexity: (a) the ROI, (b) the segmented mass, and (c) the skeleton. There were 31 bifurcation points in this skeleton.

## B. Textural features

From section 2.1, it is obvious that the geometrical features are based mostly on the segmented results. Hence, if the segmentation is not accurate, mistakes may appear in any system that uses these features as the discriminant basis. As noted in the Introduction, segmentation is often inaccurate since the geometrical boundary is always indefinable.

Textural features deal with the internal structure of the mass, and there are some problems, more or less, in the traditional methods used to extract textural features. Therefore, by combining UWT and GLCM, an appropriate method for the extraction of features is proposed in this paper, this method inherits the advantages of both of them. First, by the help of UWT, the ROI is decomposed into $3L+1$ sub-images, where $L$ is the parameter of decomposition. Then, for every sub-image, GLCM is used to obtain the textural features. Finally, these characteristics are screened by analysis.

### 1.1.1. Undecimated wavelet transform

By analyzing the Mallat algorithm [12], it is known that the decimation filter is used in each level of decomposition in the traditional wavelet transform, which mean that the dimensions of the sub-images become smaller and smaller. Therefore, with the level increased, the dimensions and resolution of the sub-images become smaller and smaller. For example, if the Haar wavelet is used to process an image that has the size of $M*N$, the dimension of the $l^{th}$ sub-image would be $(M/2^l)$ $*(N/2^l)$. However, the size of ROI is not always large, so this is obviously unreasonable.

Taking this into account, the UWT algorithm, which modifies each filter instead of decimating, is selected to decompose the ROI, which guarantees that the dimensions of the sub-image are the same as the original image. For the filter hl on level $l$, $(2^l-1)$ zeros are inserted after each coefficient, which is shown as follows:

$$\overline{h^l} = h^l(1), \underbrace{0\cdots0}_{2^l-1}, h^l(2), \ldots, h^l(i), \underbrace{0\cdots0}_{2^l-1}, , h^l(i+1) \quad (1)$$

where h(i) is the coefficient of the traditional wavelet transform filter.

### 1.1.2. Extraction of the features of sub-images

To some extent, the textural features of an image can be reflected by GLCM, which is the joint probability occurrence of pixel pairs with a defined spatial relationship (distance $d$ and direction $\theta$) having gray level values $i$ and $j$ in the image. Thus, for each sub-image, we use a different distance parameter $d$, a different direction parameter $\theta$, and the corresponding quantization rule to extract three co-occurrence features, i.e., entropy *Ent*, contrast *Con*, and Correlation *Cor*.

**Direction parameter $\theta$:** Different from the common GLCM, a different direction parameter, which considers the physical significance of the sub-band and is based on the principal direction of the sub-image, is set for each sub-image. Hence, for the horizontal component $W^H$, choose $\theta = 0°$; set $\theta = 90°$for the vertical component $W^V$; Select $\theta = 45°$ and $\theta = 135°$ for the diagonal component $W^D$; Calculate all the four directions for the approximation component W.

**Distance parameter $d$:** Traditional GLCM analysis usually sets $d = 1$, which makes it possible to use the grayscale co-occurrence features instead of high frequency components of the wavelet [13]. Considering that interpolation is used to modify filters, different distance parameters are chosen for each level. Namely, set $d = 2^l$ for the $l^{th}$ sub-image. (According to this method, when $l = 0$, $d = 2^0 = 1$, which is consistent with the traditional situation.)

**Quantization rules:** Because the dimension of the GLCM is $N_g$ x $N_g$, where $N_g$ is the gray level of the image, extensive calculations are needed, and, as a result, corresponding quantization rules should be proposed. Considering that the high-frequency sub-band for which the grayscale distribution is similar to a normal distribution is different from the approximation component, uniform quantization is provided for the approximation component, while non-uniform quantization is applied to sub-bands. All of this is shown in Fig. 2. In this paper, the quantization level is set as $N_q$, and divided stage quantization is used to achieve the purpose of non-uniform quantization. The specific non-uniform quantization rules are shown in Formula 2.

$$\begin{cases} N_{q1} = a, |i| \le \mu + \sigma \\ N_{q2} = b/2, others \end{cases} \quad (2)$$

Where $i$ is the gray level of a pixel; $\mu$ is the grayscale mean; $\sigma$ is the grayscale variance; $a$ is the quantization parameter; and $b$ is the other quantization parameter

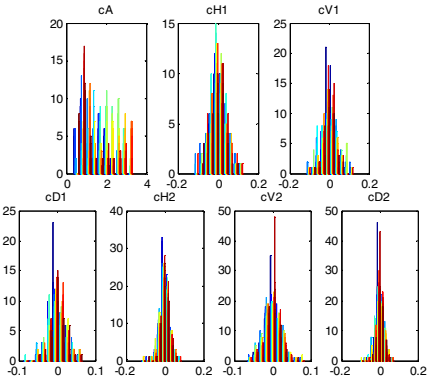


Fig. 2. Grayscale distribution of sub-bands (In this paper, $N_q$ = 16, $a$ = 12, and $b$ = 4.)

*1.1.3. Synthetic rules of features*

As noted in Section 2.2.2, only one direction is selected for the horizontal component $W^H$ and the vertical component $W^V$, thus, only one GLCM is generated for $W^H$ and $W^V$ on each level. However, in the calculation of the diagonal components $W^D$ and approximation sub-band W, two and four directions are chosen, respectively, which means more than one GLCM must be analyzed. In this situation, the co-occurrence features must be synthesized for $W^D$ and W.

Previous analysis clearly indicates that entropy and contrast are scalar and that correlation is a vector related to the direction of the main texture. Therefore, the arithmetic average is provided for synthesizing entropy and contrast, and the modulus of the vector sum is used for correlation. All are shown in Formulas 3 and 4:

$$S = \frac{1}{n}\sum_{i=1}^{n} S_i, \quad S \in \{Ent, Con\} \tag{3}$$

$$Cor = \left| \sum_{i=1}^{n} Cor_i * \overline{V}_i \right| \tag{4}$$

Where: $n$ is the synthetic parameter ($n$ = 2 for diagonal components $W^D$, $n$ = 4 for approximation sub-band W),

$\overline{V}_i$ is the direction vector ( $\overline{V}_i$ = (0, 1) for $\theta$ = 0°; $\overline{V}_i$ = ( $-\sqrt{2}/2$ , $\sqrt{2}/2$ ) for $\theta$ = 45°; $\overline{V}_i$ = (-1, 0) for $\theta$ = 90°; $\overline{V}_i$ = ( $-\sqrt{2}/2$ , $-\sqrt{2}/2$ ) for $\theta$ = 135°)

*1.1.4. Selection and analysis of features*

From the previous analysis, it is known that there are $3L$+1 sub-images after decomposition, and three co-occurrence features are extracted for each sub-band. As a result, there will be a total of $9L$+3 co-occurrence features, which will lead to extensive calculations and low efficiency. Taking into account the large number of redundancies that exist in these features, a screening operation must be used.

Based on the experimental results, the textural features that were selected are as follows: wavelet energy of approximation sub-band $f_{t1}$, entropy of approximation sub-band $f_{t2}$, correlation of approximation sub-band $f_{t3}$, contrast of approximation sub-band $f_{t4}$, wavelet energy of the second horizontal sub-band $f_{t5}$, contrast of the second horizontal sub-band $f_{t6}$, correlation of the second horizontal sub-band $f_{t7}$, correlation of the second vertical sub-band $f_{t8}$, correlation of the second diagonal sub-band $f_{t9}$, wavelet energy of the first horizontal sub-band $f_{t10}$, contrast of the first horizontal sub-band $f_{t11}$, correlation of the first horizontal sub-band $f_{t12}$, correlation of the first vertical sub-band $f_{t13}$, and correlation of the first diagonal sub-band $f_{t14}$.

## III. Classification of Masses

### A. Feature-weighted SVM

The traditional SVM considers the problem based on the importance of samples, rather than based on the importance of features. Thus, implicitly, it is assumed that the contribution of each feature used for classification is uniform. Adequate consideration of the importance of features is provided by the FWSVM algorithm, which is not the case for the traditional SVM [14].

The core of the SVM is the kernel function, which is commonly defined by the inner product or the distance of the vectors that are used to describe the similarity of the samples, and the inner product, or the distance, is calculated according to all of the features, which include some features that have weak correlations. Therefore, the calculation may be dominated by the weak-correlation features, affecting the classification performance of the classifier. For the above reasons, the feature-weighted kernel is defined.

The original motivation for introducing the kernel was to determine the linear function in the feature space, which is established by non-linear feature mapping for a non-linear model. And the weighted matrix can scale the geometrical shape of the input space as well as zoom in or out of the feature space to change the weight of the linear function.

Geometrical features are excessively dependent on the segmentation results, but this is not the case for the textural characteristics; however, the classification performance of the textural characteristics is generally worse than that of the geometrical features. Thus, the weights of the features should be allocated reasonably in order to achieve the best classification results.

### B. Analysis of feature weighting

Since the weights of features are related closely to the classification result, it is important to allocate reasonable weights to the features. In my opinion, two factors affect the weights that are set, i.e., 1) the distinguishing abilities of some features are innately stronger than others (the so-called internal factor) and 2) inaccurate segmentation (the so-called external factor). The final weight is the product of the two types of weights. In this paper, only the internal factor was considered and analyzed. While, the external weight for each feature would be equal to 1 if we assume that the segmentation result is accurate.

Between-class distance, a measure of the ability to distinguish between two types of samples, is calculated to denote the importance of the features. For a certain feature, the between-class distance between two categories is calculated.

It is apparent that the distinguishing ability becomes stronger as this value increases. Between-class distances of all the features used herein are shown in Table 1.

Table 1. Between-class distance of each feature

| | | | | |
|---|---|---|---|---|
| Geometrical features | $f_{g1}$ 1.0959 | $f_{g3}$ 0.1009 | $f_{g4}$ 0.8183 | $f_{g5}$ 0.3467 |
| | $f_{g6}$ 0.7182 | $f_{g7}$ 1.6594 | $f_{g8}$ 0.5913 | $f_{g9}$ 1.2001 |

| | | | | |
|---|---|---|---|---|
| Textural features | $f_{t1}$ 0.2887 | $f_{t2}$ 0.1997 | $f_{t3}$ 0.2222 | $f_{t4}$ 0.2502 |
| | $f_{t5}$ 0.2988 | $f_{t6}$ 0.2945 | $f_{t7}$ 0.1444 | $f_{t8}$ 0.2572 |
| | $f_{t9}$ 0.0641 | $f_{t10}$ 0.2295 | $f_{t11}$ 0.1845 | $f_{t12}$ 0.2146 |
| | $f_{t13}$ 0.0454 | $f_{t14}$ 0.0271 | | |

The distinguishing ability of textural features is generally weaker than that of geometrical features. Maybe the reason is that the extraction of textural features is conducted for the entire ROI. This may result in counting a significant amount of background area (non-mass region), which would weaken the distinguishing ability of the textural features.

So, F-score [15], a simple technique that measures the discrimination of two sets of real numbers, was used to set weights.

## IV. Results and Discussion

### A. Database and method

In order to verify the effectiveness of the proposed method, the Library for Support Vector Machines (LIBSVM) [16] and the images that were provided by the Digital Database for Screening Mammography (DDSM) [17] were used to implement the experiments. Based on discretionary decision making, 200 images were chosen from these mammograms. The images were divided into two groups, i.e., 1) a subset that contained 100 images, which was used to train the model and for some preliminary texture-selection experiments and 2) a subset to be used in the tests. In this paper, we address two experiments that were performed in this study.

### B. Improvement of the extraction of textural features

In this experiment, three schemes for the extraction of textural features were used to obtain classification criteria. For 100 test images (50 benign and 50 malignant masses), the statistic of the number of misjudgments (false positives and false negatives) was obtained, and the accuracy of the classification was calculated. The results are shown in Table 2. In order to make the results of this experiment immediately obvious, only the textural features extracted by each method are shown.

Table 2 Classification results using different methods for extracting textural features

| | Misjudgment | | Accuracy (%) |
|---|---|---|---|
| | False Negative | False Positive | |
| Mencattini's method | 38 | 38 | 24 |
| Ibrahima's method | 39 | 23 | 38 |
| Proposed method | 38 | 13 | 49 |

Since only textural features were used as classification criteria in order to make the contrast more evident, the accuracy rates of these methods were not high. But it still demonstrates that the proposed method has better classification results than the existing schemes. This is mainly attributed to the improvement of the false-positive rate. By the use of UWT, parts of the noise (such as the Gaussian noise) were averaged and assigned into each sub-band. Coupled with some follow-up operations in feature selection, the effects of these noises were inhibited to a certain extent. For example, the features of the detailed sub-bands were reduced from 24 to 10, which is equivalent to decreasing the influence of high-frequency noise. By selecting the appropriate parameters and using appropriate synthesis rules, it is obvious that the entire process is much more detailed and legitimate for the extraction of textural features.

The misjudgments may be attributed to the following two reasons:1. The extraction of textural features was conducted for the entire ROI. Most of the malignant cases have irregular boundaries. And the presence of long spicule results in many non-mass regions that have uniform, gray distributions being included, which causes the grayscale of the mass region to become weaker (Fig. 1(a)). Under these circumstances, this kind of malignant tumors is misjudged to be benign. This is also the reason that the three methods do not reduce the number of the false negatives. Even though this situation is inevitable for texture features, it is easy to resolve this kind of problem for geometric feature. 2. In order to prevent the image processing method from damaging the textural features, the extraction of the features is implemented directly for original ROI, which has a lot of noise. Therefore, some benign masses are misjudged, producing false positive results.

### C. Improvement of classification

In this experiment, two classification methods were compared to determine whether the FWSVM classifier provides more accurate results than the SVM classifier. The experimental results for 100 test images are shown in Table 3. (The same features were used for the two classifiers; both of them chose the proposed features. Optimal parameters were set through K-fold cross validation (K-CV) for each classifier). Fig. 3 shows the classification results of the two methods.

Table 3. Classification results using different methods

| Classifier | Accuracy (%) | Sensitivity (%) | Specificity (%) |
|---|---|---|---|
| SVM | 88 | 78 | 98 |
| FWSVM | 97 | 98 | 96 |

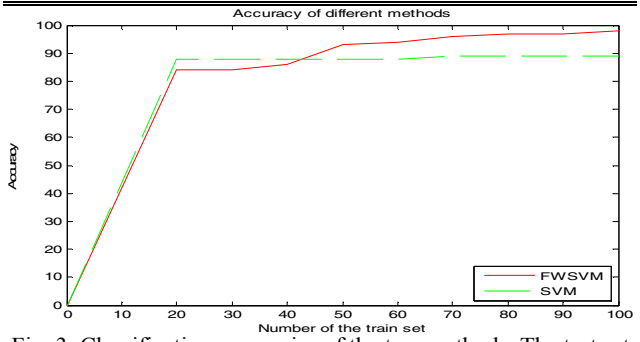


Fig. 3. Classification accuracies of the two methods. The test set contained 50 benign and 50 malignant masses, and the number of the training set ranged from 20 to 100 in increments of 10.

There were 50 benign and 50 malignant masses in the test subset. Forty-nine benignancies and 39 malignancies were classified correctly by the SVM classifier, while 48 benignancies and 49 malignancies were classified correctly by FWSVM (the number of the training set is 100). Therefore, it is obvious that FWSVM has greater sensitivity than SVM, while the specificity is almost the same. The false cases were analyzed further in the interest of determining the reasons for misjudgments. Compared with the SVM classifier, which had 11 false negatives, the FWSVM misjudged only one malignancy as a benign mass. The reason was that the

different features were assigned different weights. Features that have stronger distinguishing abilities are given greater weights. Also, the process decreases the influence of noise fluctuations, because the textural features, which are generally affected more seriously by noise than geometrical features, were given smaller weights. But with the improvement, the operation also created a new problem. Although 49 malignant masses were classified correctly by FWSVM classifier, which is better than the result of SVM classifier, the misjudged case is classified correctly by SVM classifier. The misjudged mass has a relatively regular boundary, but it was indeed a malignancy due to its larger textural fluctuations. However, it was misjudged by the FWSVM classifier because of the smaller weights of the textural features. Perhaps this problem occurred because the external factor was not considered in this study, resulting in smaller weights being assigned to the textural features. Future research will be focused on determining an appropriate approach for taking the external factor into account for the further improvement of the classification results.

## V. Conclusions and Future Work

In this paper, we have presented some upgrades for the purpose of improving the results of mass classification. First, a new morphological feature was extracted to enrich the information concerning the shape of the mass. Then, by combining UWT and GLCM, the textural features are extracted in a much more detailed fashion, which made them more legitimate. Finally, the FWSVM classifier, which used both geometrical and textural features, was utilized to distinguish between malignant and benign masses. The experimental results indicated that the proposed method is efficient and effective.

In future studies, the features will be considered more meticulously, and the weights will be assigned more reasonable values. The preliminary idea is that the extraction of textural features should be targeted at the mass region and its border, rather than the entire ROI. Moreover, the accuracy and the reliability of the segmentation results will be used as a factor in setting weights.

## VI. Acknowledgment

This work was done based on the public DDSM database, which provides many standard digital mammograms and guidance information. The authors thank staff members at the Center of Breast Disease at Peking University People's Hospital for detailed guidance and assistance.

## VII. References

[1] American Cancer Society, Breast Cancer Facts and Figures 2009-2010 [R]. Atlanta: American Cancer Society. 2010

[2] Cancer Research U.K., Breast Cancer Mortality Statistics [EB/OL].

[3] Yao C, Chen H J, Yang Y Y, et al. Microcalcification clusters processing in mammograms based on relevance vector machine with adaptive kernel learning[J]. Acta Physica Sinica. 2013, 62(8):088702(1-11).

[4] International Agency for Research on Cancer, Most frequent cancers: women [EB/OL].

[5] Ibrahima F, Brahim B S, Mohamed M M E. Digital Mammograms Classification Using a Wavelet Based Feature Extraction Method [C]. 2009 Second International Conference on Computer and Electrical Engineering (ICCEE'09). 2009,2:318-322. [DOI:10.1109/ICCEE.2009.39]

[6] Tsui P H, Liao Y Y, Chang C C, et al. Classification of Benign and Malignant Breast Tumors by 2-D Analysis Based on Contour Description and Scattered Characterization [J]. IEEE Transactions on Medical Imaging. 2010, 29(2):513-522.

[7] Tang J S, Rangayyan R M, Xu J, et al. Computer-Aided Detection and Diagnosis of Breast Cancer with Mammography: Recent Advances [J]. IEEE Transactions on Information Technology in Biomedicine. 2009, 13(2):236-251.

[8] Shen J L, Mammary Ultrasound Image Analysis and Breast Tumors Classification [D]. Shanghai: Fudan University, 2006.

[9] Han Z Z, Chen H J, Li J P, et al. Mass Modeling and Segmentation in Mammogram [J]. Transactions of Beijing Institute of Technology. 2013, 33(5):495-499.

[10] Han Z Z, Chen H J, Li J P, et al. A Novel Method of Mass Segmentation in Mammogram [C]. 2012 International Conference on Systems and Informatics (ICSAI). Yantai: IEEE Press, 2012:1412-1416.

[11] Mencattini, A, Salmeri M, Rabottino G, et al. Metrological Characterization of a CADx System for the Classification of Breast Masses in Mammograms [J]. IEEE Transactions on Instrumentation and Measurement. 2010, 59(11):2792-2799.

[12] Gonzalez R C, Woods R E. Digital Image Processing (2nd Ed.) [M]. Trans. Ruan Q Q. Beijing: Publishing House of Electronics Industry. 2007,8: 300-306

[13] Clausi D A, Deng H. Design-based texture feature fusion using Gabor filters and co-occurrence probabilities [J]. IEEE Transactions on Image Processing. 2005, 14(7):925-936.

[14] Wang T H, Tian S F, Huang H K. Feature Weighted Support Vector Machine [J]. Journal of Electronics and Information Technology. 2009, 31(3):514-518.

[15] Chen Y W, Lin C J. Combining SVMs with Various Feature Selection Strategies [EB/OL].

[16] Chih-Chung Chang, Chih-Jen Lin. LIBSVM -- A Library for Support Vector Machines [EB/OL].

[17] Digital Database for Screening Mammography.2012. Available: http://marathon.csee.usf.edu/Mammography/Database.html