Hybrid Classification Approach of Malignant and Benign Pulmonary Nodules Based on Topological and Histogram Features

Y. Kawata, N. Niki, H. Ohmatsu[a], M. Kusumoto[b], R. Kakinuma[a], K. Mori[c], H. Nishiyama[d], K. Eguchi[e], M. Kaneko[b], N. Moriyama[b]

Dept. of Optical Science, Univ. of Tokushima, Tokushima ,
[a]National Cancer Center Hospital East, [b]National Cancer Center Hospital,
[c]Tochigi Cancer Center, [d]The Social Health Insurance Medical Center,
[e]National Shikoku Cancer Center Hospital

Abstract. This paper focuses on an approach for characterizing the internal structure which is one of important clues for differentiating between malignant and benign nodules in three-dimensional (3-D) thoracic images. In this approach, each voxel was described in terms of shape index derived from curvatures on the voxel. The voxels inside the nodule were aggregated via shape histogram to quantify how much shape category was present in the nodule. Topological features were introduced to characterize the morphology of the cluster constructed from a set of voxels with the same shape category. The properties such as curvedness and CT density were also built into the representation. In the classification step, a hybrid unsupervised/supervised structure was performed to improve the classifier performance. It combined the k-means clustering procedure and the linear discriminate (LD) classifier. The performance of the hybrid classifier was compared to that of LD classifier alone. Receiver operating characteristics (ROC) analysis was used to evaluate the accuracy of the classifiers. We also compared the performance of the hybrid classifier with those of physicians. The classification performance reached the performance of physicians. Our results demonstrate the feasibility of the hybrid classifier based on the topological and histogram features to assist physicians in making diagnostic decisions.

1. Introduction

Lung cancer is the leading cause of cancer death among Japanese men and its incidence continues to increase [1]. In order to improve the recovery rate for lung cancer, detection and treatment at an early stage of growth is necessary. Radiography of the chest is ordinary used for the screening of lung cancer. At present, after screening via chest radiograph to detect suspicious areas, differential diagnosis is ordinarily concluded by histology from biopsy. There are, however, a significant number of malignant cases which should be discriminated as early as possible from benign lesions. Particularity when the peripheral lung for suspicious areas in early development are diagnosed, it is often the case that the differential diagnosis by means of transbronchial or percutaneous biopsies becomes difficult.

S.L. Delp, A.M. DiGioia, and B. Jaramaz (Eds.): MICCAI 2000, LNCS 1935, pp. 297-306, 2000.

There has been a considerable amount of interest in the use of thin-section CT images to observe small pulmonary nodules for differential diagnosis without invasive operation [1]-[3]. In assessing the malignant potential of small pulmonary nodules in thin-section CT images, it is important to examine the condition of nodule interface, the nodule internal intensity, and the relationships between nodules and surrounding structures such as vessels, bronchi, and spiculation [1]-[3]. A number of investigators have developed a feature extraction and a classification methods for characterizing pulmonary nodules. Siegelman et al. investigated CT density in the center of a nodule on two-dimensional (2-D) CT images [4]. Other groups also presented nodule density analysis with a special reference phantom to improve measurement accuracy [5]. Cavouras demonstrated that multiple features including nodule density and texture were useful to classify malignancies from other lesions [6]. Following his work, McNitt-Gray proposed pattern classification approach incorporating multiple features, including measures of density, size measures, and texture of nodules on CT slice images [7]. One promising area of recent researches has been the analysis of three-dimensional (3-D) pulmonary nodule images. We quantified the concave and convex surfaces by using surface curvatures to characterize surface condition of malignant and benign nodules [8],[9]. Hirano presented an index to quantify how a nodule evolved the surrounding vessels [10]. Tozaki proposed a classification approach between pulmonary artery and vein to characterize the relationships between nodules and surrounding structures [11]. Kitaoka developed mathematical models of bronchial displacements caused by nodules to discriminate cancers from inflammatory pulmonary nodules [12]. Although the performances of the computer algorithms are expected to depend strongly on data set, they indicate the potential of using computer aided diagnosis techniques to improve the diagnostic accuracy of differentiating malignant and benign nodules.

This paper focuses on the analysis of the internal structure in the 3-D pulmonary nodules. In previous study [13],[14] we found that curvature indexes such as shape index and curvedness were promising quantities for characterizing the internal structure of nodules. However, there were several distribution patterns of CT density inside the nodule, such as solid or infiltrative types. Therefore, it might be desirable to decompose input samples into classes with different properties to improve classification performance. In this present study, we combine an unsupervised and a supervised model and apply them to classification of malignant and benign nodules. The unsupervised model is based on k-means clustering (KMC) procedure [21] which clustered the nodules into a number of classes by using CT density distribution. A supervised linear discriminate (LD) classifier [21] is designed for each classes by using topological and histogram measures based on curvature indexes. By improving the homogeneity of the nodules, the LD classifier designed may be more robust. The performance of the hybrid classifier will be compared with those of the LD classifier alone and physicians. We will demonstrate that the proposed hybrid structure can improve the accuracy of classification in computer-aided diagnosis applications.

This paper is organized as follows. Section 2 introduces a hybrid classifier and describes the feature extraction schemes of the nodule internal structure. In Section 3, our data set is described and experimental results are presented. Finally, Section 4 concludes this investigation.

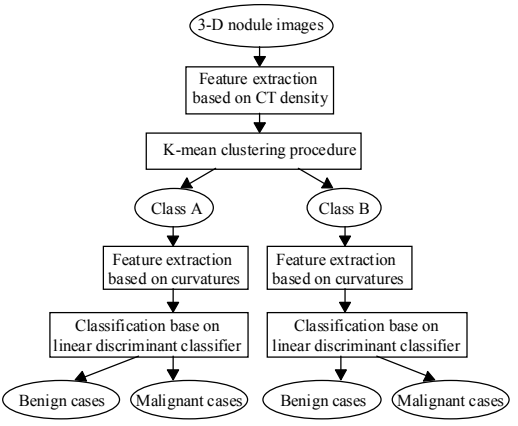


Fig. 1. Block diagram of the hybrid classification of malignant and benign pulmonary nodules.

2. Method

2.1 Overview

We propose to design a hybrid classifier that combines the unsupervised KMC procedure with a supervised LD classifier. The classification procedure of the pulmonary nodules is shown in Fig.1. Since there are several distribution patterns of CT density value inside nodules, the KMC procedure improve the homogeneity of the sample distributions by classifying classes with different properties regarding the distribution patterns of CT density value. In this study the KMC separates the sample data into two classes denoted as class A and B. The class A is the class in which the mean CT density value inside the nodule is high and the class B is the class in which the mean CT density value inside the nodule is low. For each class a LD classifier is designed by using the CT density and the curvature based features inside nodules and then discriminate malignant and benign nodules. The first-stage KMC procedure may improve the performance of the LD classifier if the subclass causes the sample data to deviate from multivariate normal distributions for which the LD classifier is an optimal classifier.

2.2 Nodule Segmentation

The segmentation of the 3D pulmonary nodule image consists of three steps[9]; 1) extraction of lung area, 2) selection of the region of interest (ROI) including the nodule region, 3) nodule segmentation based on a geometric approach. This lung area extraction step plays an essential role when the part of a nodule in the peripheral lung

area touches the chest wall. The ROI including the nodule was selected interactively. A pulmonary nodule was segmented from the selected ROI image by the geometric approach proposed by Caselles [15]. The deformation process of the 3-D deformable surface model can automatically stop when the deforming surfaces reach the object's boundary to be detected. In our application we added a stopping condition to exclude vessels and bronchi which were in contact with the nodule [9].

2.3 Curvature Based Representation

Each voxel in the region of interest (ROI) including the pulmonary nodule was locally represented by a vector description which relied on the CT density value and two curvature indexes that represented the shape attribute and the curvature magnitude [17], [18]. By assuming that each voxel in the ROI lies on the surface which has the normal corresponding to the 3-D gradient at the voxel, we computed directly the principal curvatures κ_1 and κ_2 ($\kappa_1 \geq \kappa_2$)on each voxel from the first and the second derivatives of the gray level image of the ROI [16]. To compute the partial derivatives of the ROI images, the ROI images were blurred by convolving with a 3-D Gaussian function of width σ . In order to take only nonnegative values, we used the shape index defined as

$$S(\mathbf{x};\sigma) = \frac{1}{2} + \frac{1}{\pi} \arctan \frac{\kappa_1(\mathbf{x};\sigma) + \kappa_2(\mathbf{x};\sigma)}{\kappa_1(\mathbf{x};\sigma) - \kappa_2(\mathbf{x};\sigma)} \tag{1}$$

The curvedness of a surface at the voxel **x** is defined as

$$R(\mathbf{x};\sigma) = \sqrt{\frac{\kappa_1(\mathbf{x};\sigma)^2 + \kappa_2(\mathbf{x};\sigma)^2}{2}} \tag{2}$$

It is a measure of how highly curved a surface and its dimension is that of the reciprocal of length.

2.4 Curvature Based Representation

In order to characterize the pulmonary nodule through the local description, we used the shape spectrum which was introduced for object recognition by Dorai and Jain [18]. Using the shape spectrum, we measured the amount of the voxel which had a particular shape index value *h*. The augment shape spectrum with scale σ is given by

$$H(h;\sigma) = \frac{1}{V} \iiint_O \delta(S(x;\sigma) - h) dO \tag{3}$$

where , *V* is the total volume of the specified region *O*, *dO* is a small region around ***x*** and$\delta$is the Dirac delta function. The discrete version of Eq.(1) is derived by dividing the shape index range into *B* bins and counting the number of voxels falling in each bin *k* and normalizing it by the total number *N* of discrete voxels in the specified region. The discrete version is expressed by

$$H(h=\frac{k}{B})=\frac{1}{N}\sum_{i=1}^{N}\chi_k(S(x_i;\sigma)) \quad (4)$$

with

$$\chi_k(t)=\begin{cases}1 & \frac{k-1}{B}\le t<\frac{k}{B}\\ 0 & otherwise\end{cases} \quad (5)$$

Here, the segmented 3-D pulmonary nodule image is utilized as the specified region *O*. The shape index value one is included in the *B*-th bin. The discrete version of the shape spectrum was called shape histogram. The number of voxel falling in each bin represented the value of the histogram feature. For computational purposes, such as comparing spectra of different nodules, the shape histogram was normalized with respect to the volume of nodule. The normalized number of voxel falling in each bin represents the value of the shape histogram feature. In this study the number of bin *B* was given the value 100. The similar equations for the curvedness and CT density are obtained in the same manner. The domains of curvedness and CT density were specified to [0,1] and [-1500, 500], respectively. A voxel in which the curvedness value was larger than one was considered as a voxel with curvedness value one. For the CT density the similar process was performed. To classify malignant and benign nodules, we combined a set of histogram features, such as shape, curvedenss, and CT density histogram.

2.5 Topological Features

The distribution morphology of the shape category can characterize the internal structure of the nodule. Therefore, we divided the inside of the nodule into four shape categories by the shape index value and then, computed the topological features of each 3-D cluster which constructed from a set of voxels with the same shape category. The four shape categories were peak, saddle ridge, saddle valley, and pit surface types and the interval of the shape index for each shape categories were set [0, 0.25], [0.25, 0.5], [0.5, 0.75], [0.75, 1], respectively. The topological features used here were the Euler number, the number of connected components, cavities, and holes of a 26-connected object [19]. The Euler number of a 3-D digital figure is defined as the following equation,

$$E=b_0+b_1+b_2 \quad (6)$$

where E is the Euler number and b_0, b_1, and b_2 respectively represent the number of connected components, holes, and cavities. These valuables, b_0, b_1, and b_2 are called the first, second, an third Betti-number, respectively. Yonekura et al. [19] provided computation schemes for the basic topological properties such as the connected component and the Euler number of 3-D digital object. The Euler number, the number of connected components, cavities, and holes were obtained by their schemes. In addition, we quantified how each shape category distributes inside the nodule by using a technique computing exact Euclidean distance transform [20]. Using the Euclidean distance in the nodule, we measured an amount of voxel which had a

particular distance value d in the i-th shape category. The distance spectrum with shape category is given by

$$DH(d;i) = \frac{1}{V}\iiint_O \delta(D(x;i) - d)dO \tag{7}$$

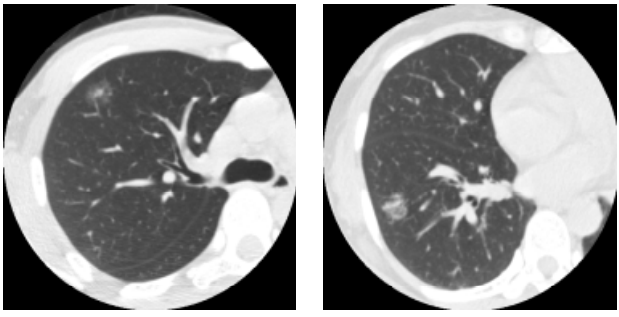
where $D(\boldsymbol{x}; i)$ is the distance value at the voxel **x** with i-th shape category. The discrete version was derived by the similar manner as the histogram features. In this study the number of bin was given the value 25. The normalized number of voxel falling in each bin represents the value of the shape distribution feature. This feature was included in the topological features.

2.6 Classification

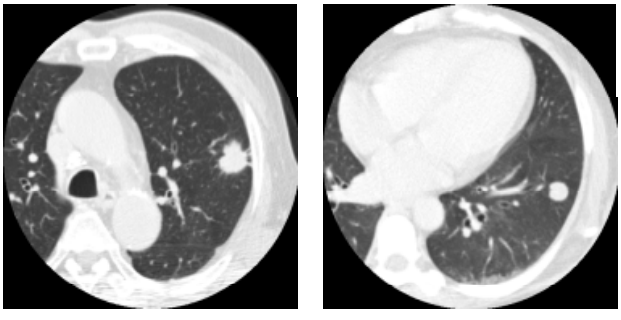
In order to improve the accuracy of a classifier, we combined the unsupervised KMC procedure with the supervised LD classifier. The KMC classified the sample nodules into two classes by using the mean CT density value for three different regions. These different considered regions are as follows: (i)core region shrinking to T_1% of the maximum distance value of the 3-D nodule image, denoted as R1, (ii) complement of the core region in the 3-D nodule image, denoted as R2, (iii) marginal region extended to T_2% of the maximum distance value of the 3-D nodule image, denoted as R3. The maximum distance value was obtained by applying the Euclidean distance transformation technique [20] to the segmented 3-D nodule image. In this study T_1 and T_2 values were assigned to 60% and 26%, respectively. For each class a LD classifier was designed by using the topological and histogram features. In order to reduce the number of the features and to obtain the best feature set to design a good classifier, feature selection with forward stepwise selection procedure was applied [22]. In this study the minimization of Wilks' lambda was used as an optimization criterion to select the effective features. A leave-one-out procedure was performed to provide a less biased estimation of the linear discriminate classifier's performance. The discriminant scores were analyzed using receiver operating characteristic (ROC) method [23]. The discriminant scores of the malignant and the benign nodules were used as the decision variable in the ROCKIT program developed by Metz which fits the ROC curve based on maximum likelihood estimation. This program was also used to test the statistical significance of the difference between pairs of ROC curves. The two-tailed p values were reported in the comparison procedure described in the next section.

3. Experimental Results

Thin-section CT images were obtained by the helical CT scanner (Toshiba TCT900S Superhelix and Xvigor). Per patient, thin-section CT slices at 1mm intervals were obtained to observe whole nodule region and its surroundings. The range of pixel size in each square slice of 512 pixels was between 0.3x0.3 mm^2 and 0.4x0.4 mm^2. The 3-D thoracic image was reconstructed by a linear interpolation technique to make each voxel isotropic. The data set in this study included 210 3-D thoracic images provided



(a) (b)



(c) (d)

Fig. 2. Slice images including pulmonary nodules in class A and class B. (a) Malignant nodule in class A. (b) Benign nodule in class A. (c) Malignant nodule in class B. (d) Benign nodule in Class B.

by National Cancer Center Hospital East and Tochigi Cancer Center. Among the 210 cases,141 contained malignant nodules, and 69 contained benign nodules. Whole malignant nodules were cytologically or histologically diagnosed. In benign cases, lesions showed no change or decreased in size over a 2-year period were considered as the benign nodules. The size of nodules was less than 20 mm in diameter.

The selection of the width σ of the Gaussian function is a critical issue. In previous work [13],[14], we selected the value of the width which provided high accuracy of classification between malignant and benign nodules for discrete width values. From the result, we assigned the value 2.0 to the width. We compared the hybrid classifier with the LD classifier alone. In the hybrid classifier, the KMC procedure classified input samples into two classes denoted as class A and class B. The class A contained 50 cases (13 benign cases and 37 malignant cases) and the class B contained 160 cases (56 benign cases and 104 malignant cases). Fig.2 presents slice images including pulmonary nodules in class A and class B. In comparison with the class B, the nodule of the class A had ill-defined surface and the region with lower CT density value occupied the inside of the nodule. The histogram feature used here was combined with three histogram features such as shape, curvedness, and CT density histograms. The topological features were yield for four clusters. The LD

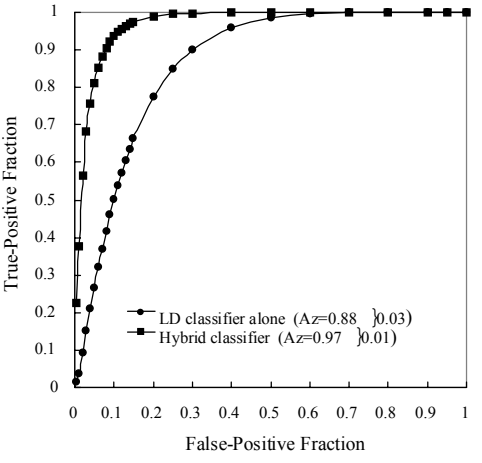


Fig. 3. ROC curves of the hybrid classifier and the LD classifier alone.

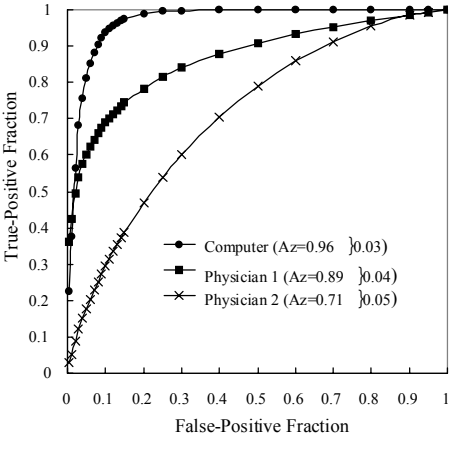


Fig. 4. Comparison of ROC curves obtained by using physicians malignancy rating and the discriminate score of hybrid classifier.

classifiers were designed from the combined topological and histogram features. For the class A and B, the number of the selected features was 9 and 21, respectively. In the LD classifier alone, the number of the selected features was 15. The ROC curves of the hybrid and the LD classifiers were plotted in Fig. 3. The classification accuracy in the hybrid classifier was significantly higher than those in the LD classifier alone ($p<0.01$).

In order to compare the performance of physicians with that of the hybrid classifier, the probability of malignancy of each pulmonary nodule in thin-section CT images which were printed on films, was ranked by eleven physicians on a scale of 1 to 10, where a ranking of 1 corresponded to the nodules with the most benign cases. The number of nodules used in this comparison was 119 cases provided by the

National Cancer Center East. Based on the ranking, the ROC curves using three physicians malignancy rating and the computer's discriminate score output in the combined feature space were plotted in Fig. 4. The physicians 1 and 2 respectively have 15 years, and one year of experience in the chest radiology. The difference between the ROC curves of the hybrid classifier and those of two physicians was statistical significance ($p < 0.05$). These results show that the classification performance of the hybrid classification approach achieved the experienced physician results.

4. Conclusion

In this study, the topological and histogram measures based on curvature information were introduced to characterize the internal structure of 3-D nodule images. A hybrid classifier combining an unsupervised k-means clustering algorithm with a supervised LD classifier has been designed and applied to the classification of malignant and benign nodules. The *Az* value under the ROC curve for our data set was higher for the hybrid classifier compared to that of the LD classifier alone. A greater improvement was obtained by introducing the k-means clustering procedure. The performance of the hybrid classifier was also compared with those of the experience physicians. The classification performance of the hybrid classifier reached the performance of the experienced physicians. These results indicate that the hybrid classifier is a promising approach for improving the accuracy of classifiers for CAD applications.

Acknowledgments

The authors are grateful to physicians cooperating to the reading test. The authors would like to thank Prof. Charles E. Metz for the ROCKIT program.

References

1. M.Kaneko, K.Eguchi, H.Ohmatsu, R.Kakinuma,T.Naruke, K.Suemasu, N. Moriyama, "Peripheral lung cancer: Screening and detection with low-dose spiral CT versus radiography," Radiology, Vol.201, pp.798-802 (1996)
2. K.Mori, Y.Saitou, K.Tominaga, K.Yokoi, N.Miyazawa, A.Okuyama, M.Sasagawa, "Small nodular legions in the lung periphery: new approach to diagnosis with CT," Radiology, Vol.177, pp.843-849 (1990)
3. C.V. Zwirewich, S.Veda, R.R.Miller, N.L.Muller, "Solitary pulmonary nodule: high-resolution CT and radiological pathologic correlation," Radiology, Vol.179, pp.469-476, (1991)
4. S.S.Siegelman, E.A.Zerhouni, F.P.Leo, N.F. Khouri, F.P. Stitik, "CT of the solitary pulmonary nodule," AJR, Vol.135, pp.1-13 (1980)
5. A.V.Proto and S.R.Thomas, "Pulmonary nodules studied by computed tomography," Radiology, Vol.156, pp.149-153 (1985)

6. D. Cavouras, P. Prassopoulos and N. Pantelidis, “Image analysis methods for solitary pulmonary nodule characterization by computed tomography,” European Journal of Radiology, Vol.14, pp.169-172 (1992)
7. M.F. McNitt-Gray, E.M.Hart, J. G. Goldin, C.-W, Yao, and D.R. Aberle, " A pattern classification approach to characterizing solitary pulmonary nodules imaged on high resolution computed tomography," Proc. SPIE, Vol. 2710, pp.1024-1034 (1996).
8. Y.Kawata, N.Niki, H.Ohmatsu, R.Kakinuma, K.Eguchi, M.Kaneko, N.Moriyama, “Shape analysis of pulmonary nodules based on thin-section CT images”, Proc. SPIE, Vol. 3034, pp.967-974 (1997)
9. Y.Kawata, N.Niki, H.Ohmatsu, R.Kakinuma, K.Eguchi, M.Kaneko, N.Moriyama, "Quantitative surface characterization of pulmonary nodules based on thin-section CT images", IEEE Trans. Nuclear Science, Vol. 45, pp.2132-2138 (1998)
10. Y.Hirano, Y.Mekada, J.Hasegawa, J. Toriwaki, H.Ohmatsu, and K.Eguchi, “Quantification of vessels convergence in three-dimensional chest X-ray CT images with three-dimensional concentration index”, Medical Imaging Technology, Vol.15, pp.228-236 (1997)
11. T.Tozaki, Y.Kawata, N.Niki, H.Ohmatsu, R. Kakinuma, K.Eguchi, N. Moriyama, "Pulmonary organs analysis for differential diagnosis based on thoracic thin-section CT images", IEEE Trans. Nuclear Science, Vol.45, pp.3075-3082 (1998)
12. H. Kitaoka and R. Takaki, “Simulations of bronchial displacement owing to solitary pulmonary nodules,” Nippon Acta Radiologica, Vol.59, pp. 318-324 (1999)
13. Y. Kawata, N.Niki, H.Ohmatsu, M. Kusumoto, R. Kakinuma, K. Mori, K. Eguchi, M. Kaneko, N. Moriyama, “Classification of pulmonary nodules in thin-section CT images by using multi-scale curvature indexes” , IEEE Int. Conf. on Image Processing, Vol.2, pp.197-201 (1999)
14. Y. Kawata, N.Niki, H.Ohmatsu, “Curvature based internal structure analysis of pulmonary nodules using thoracic 3-D CT images”, IEICE Trans., Vol.J-83-D-II, pp.209-218 (2000).
15. V.Caselles, R.Kimmel, G.Sapiro, and C.Sbert, "Minimal surfaces based object segmentation," IEEE Trans. Pattern Analysis Machine Intelligence, Vol.19, pp.394-398 (1997)
16. J.-P, Thirion and A. Gourdon, "Computing the differential characteristics of isointensity surfaces," Computer Vision and Image Understanding, Vol.61, pp.190-202 (1995)
17. J. J. Koenderink and A.J.V. Doorn, "Surface shape and curvature scales", Image and Vision Computing, Vol.10, pp.557-565 (1992)
18. C.Dorai and A.K. Jain, " COSMOS-A representation scheme for 3D free-form objects", IEEE Trans. Pattern Analysis Machine Intelligence, Vol.19, pp.1115-1130 (1997)
19. T. Yonekura, S. Yokoi, J. Toriwaki, T. Fukumura, “Connectivity and Euler number of figures in the digitized three-dimensional space”, Trans. IECE, vol. J65-D, pp.80-87,1982.
20. T. Saito and J. Toriwaki, “Euclidean distance transformation for three-dimensional digital images”, Trans. IEICE, Vol.J76-D-II, pp.445-453 (1993)
21. R.O. Duda and P.E. Hart, “Pattern classification and scene analysis”, John Wiley, Sons, (1973)
22. M.C. Costanza and A.A. Afifi, “Comparison of stopping rules in forward stepwise discriminate analysis”, J. of the American Statistical Association, Vol. 74, pp.777-785, (1979)
23. C.E. Metz, "ROC methodology in radiologic imaging", Investigative Radiology, Vol.21, pp.720-733 (1986)