CLASSIFICATION OF GASTRIC TUMORS USING SHAPE FEATURES OF GLAND

Toshiyuki Tanaka*, Yoshitaka Uchino* and Teruaki Oka**

* Department of Applied Physics and Physico-Informatics, Keio University,
3-14-1 Hiyoshi, Kohoku-ku, Yokohama, 223-8522 Japan
** Division of Pathology, Kanto Central Hosptal,
6-25-1 Kamiyouga, Setagaya-ku, Tokyo 158-8531, Japan

tanaka@appi.keio.ac.jp

Abstract: Recently in Japan the pathologists are shorthanded, and nevertheless each pathological diagnosis requires much time because each analyte has to be inspected by the plural pathologists for a adequate diagnosis. This paper deals with the classification method of gastric cancer and gastric adenoma, using image processing and pattern analysis. We first select R component, G component in RGB basis of digital image, and Y component in YIQ basis for our system. After pre-processing, we automatically extracted the shape of nucleus and cytoplasm. After many inspections, we selected 40 features about shape of nucleus and cytoplasm and 14 features about texture within the cytoplasm for the assessment of tumors. Man performed the principal component analysis (PCA), F test of homoscedasticity, t test of difference of average, stepwise method for selecting the smaller number of features, and discriminant method with Mahalanobis distance. The total ratio of diagnosis reached 96.9% by our proposed method. The results show the validness of our proposed method.

Introduction

In Japan the lack of doctors in many branches of medicine are reported at every occasion. The number of pathologists seems to be especially smaller than other branches of medicine. Since the pathological diagnosis is subjectively performed by each doctor, plural doctors have to diagnose the same specimen and they aggregate those results for accurate diagnosis, and each pathological diagnosis requires much time. Therefore the pathologist wishes the diagnosis supporting system with objective and quantitative approach. In medical image engineering, automatic and quantitative diagnosis systems[1]-[11] have previously been studied by many researchers. Most of those researches are performed for the development of diagnosis supporting system for dissolving the lack of pathologists. The prescreening system for uterus cytodiagnosis is in practical use as the representative system. Besides man performs many approaches of image processing[1]-[5] for the diagnosis systems. As researches for tumor images, man reported in the following; detection of the region of breast cancer[6], region extraction of glomerulus in kidney and so on. Although the sorts of intended tumor increase year by year, the objects of most researches are lung[7], colon[8][9], prostate[10], ovarian[11] and so on, and few researches are for gastric tumor.

The well-differentiated gastric cancer and gastric adenoma are different from each other, and the classification is recently discussed by the pathologists. However the automatic diagnosis system of gastric tumor is not enough to study as described in the above, and the pathologists push for the system for gastric tumor. A morphological classification of colorectal microscopic images[4] is reported as the recent diagnosis supporting system. This is the method that man obtained the features of tumors from the whole image, and has the demerits of low classification ratio for the images with more background region.

In this paper we propose the diagnosis supporting system for the gastric tumor based on the morphological features of cytoplasm and nucleus. In this system, we classify the gastric tumor into a gastric cancer and gastric adenoma, using the numeric features that are obtained from the morphology based on the observing point of pathologist. For the numeric conversion of morphological features, man performs in this study a manual extraction of region of interests (ROI), as pre-processing a selection of color component of image, contrast enhancement, binarization with Laplacian histogram method and discriminant method. Man classifies each tumor into the gland structure and nucleus by the pre-processing. From the divided regions man computes 40 shape features and 14 texture features, and classifies the gastric tumors by discriminant method with stepwise method using the obtained features. We select the small number of features for the diagnosis by stepwise method. Finally, man classifies the tumors into three categories such as not cancer, gastric adenoma, and gastric cancer by the obtained numeric features. As the results of our proposed method, the classification ratio of each case is shown in table. All the images for our research are stained at Kanto Central Hospital, since the results of dyeing are a little different at each hospital.

ISSN: 1727-1983

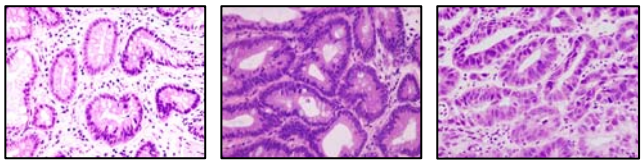
Materials and Methods

(1) Selection of used image

The pathologists classify a gastric tumor into five categories as shown in Table 1, based on the shape of nucleus and cytoplasm as shown in Figure1. Group 1 includes normal tissue and obvious benign lesion, group 5 corresponds to complete gastric cancer. Group 2 to 4 are boundary cases between benign tissue and malignant cancer. Classification of group 2 to 4 is difficult even if the pathologist diagnose. We classify the gastric tumor into group 1, group 3 and group 5 as shown in Figure 1. The original images are selected into the correct category by the pathologist. In this research we do not use the image in group2 and group 4.

Table1: Group Classification of Gastric Biopsy specimens

Group1	Normal tissue and benign lesion
Group2	Benign lesion with aberrant tissue
Group3	Boundary case between benign and malignant tumor
Group4	Tissue at increased risk for cancer
Group5	Complete cancer



(a) Not Malignant (b) Gastric Adenoma (c) Gastric Cancer

Figure 1: Original Images

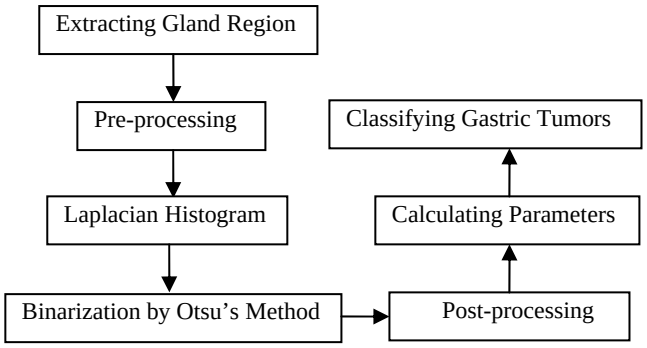


Figure 2: Scheme of Processing

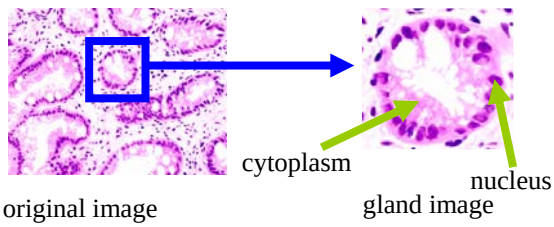


Figure 3: Nucleus and cytoplasm within the gland area

Figure 2 shows the process of our proposed method. First, we manually extract gland region as the rectangle that has sufficient size including the whole shape of gland as shown in Figure 3. The glands for our research are selected only when the whole shape exist within the original image. We do not select the defective glands that are clipped at the photographing. Therefore the number of usable gland becomes small. Next, we investigated the color information for our classification method. Figure 4 shows the gland image manually clipped from the original image.

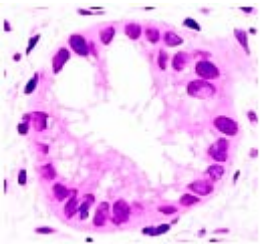
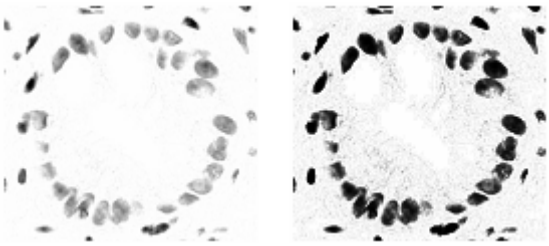
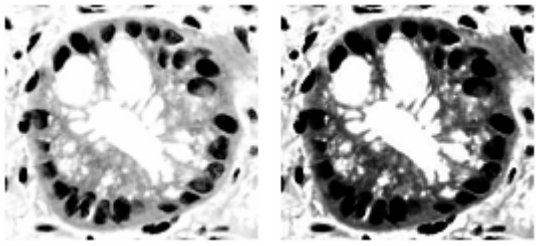


Figure 4: The image of gland



(a) R component (b) Binarization of R component

Figure 5: R component image of RGB color basis of gland and the binarization image



(a) G component (b) Binarization of G component

Figure 6: G component image of RGB color basis of gland and the binarization image

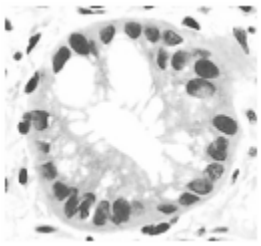


Figure 7: Y component image of YIQ color basis

ISSN: 1727-1983

(2) Pre-processing and binarization

Figure 5 and 6 show R component image and G component image of RGB color basis and the binarization images. We decided from many discussions that man used R component for extraction of nucleus within the gland region, and man used G component for extraction of cytoplasm from the image. Figure 7 shows the Y component image of YIQ color basis that is converted from RGB color basis using equation (1). We use Y component images for computing the texture features, since Y component corresponds to the intensity of each pixel.

$$\begin{bmatrix} Y \\ I \\ Q \end{bmatrix} = \begin{bmatrix} 0.299 & 0.587 & 0.114 \\ 0.596 & -0.274 & -0.322 \\ 0.211 & -0.522 & 0.311 \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \quad (1)$$

We perform binarization of R component image and G component image with Laplacian histogram method and discriminant method which is called Ohtsu method. Figure 5(b) and figure 6(b) shows the results of binarization. Figure 5(b) has clear shapes of nucleus and figure 6(b) has clear shape of cytoplasm. Man performs filtering and closing processing as post-processing. Filtering is carried out by 3×3 median filter for eliminating the small noise of the binary image. Closing process is carried out for eliminating the relative large noise. Next, man performs labeling process for extracting the large connected elements in the image. We select the largest area of dark color as the cytoplasm, and select the largest area of light color as cavity within the cytoplasm. Other connected components are eliminated as the noise.

(3) Selection of shape and texture features

Here we first explain the shape features. We select 40 shape features based on the area, the length, the chord and axis, the equivalent shape, and other shape features.

(3-1) features based on the area

Area of nucleus, area of cytoplasm within the gland, area of cavity, and the area of gland are used. The cytoplasm area corresponds to region A in Figure 8, cavity area is region B and C, gland area is region A+B+C, respectively.

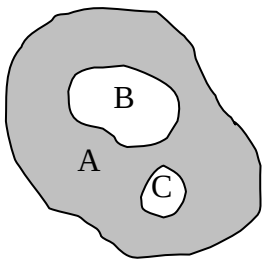


Figure 8: Regions of cytoplasm and cavity

(3-2) features based on the length

Length of gland and cavity, ferret diameter are used.

(3-3) features based on chord and axis

Maximal segment, maximal section, average of horizontal chord and average of vertical chord, average of vertical section are used. Maximal segment corresponds to line A in Figure 9, maximal section is line B, and average of vertical section is line C, respectively.

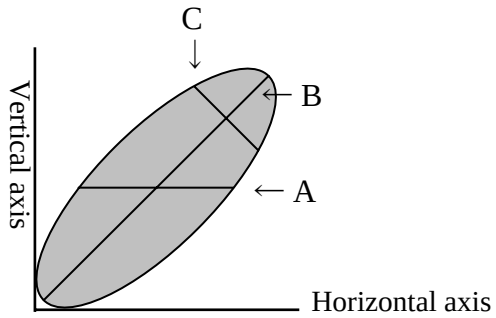


Figure 9: Definition of chord and axis

(3-4) features based on equivalent shape

Equivalent short axis of ellipse, short axis of ellipse, long axis of ellipse, ellipse ratio, narrow side of equivalent rectangle, wide side of equivalent rectangle, edge ratio of rectangle are used.

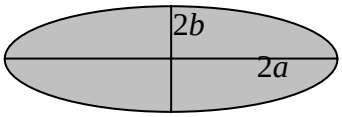


Figure 10: Definition of equivalent ellipse

(3-5) other shape features

Inertia moment, tension factor, degree of dispersion, Heywood roundness, hydraulic depth, Waddle disk radius, and ratio of nucleus to cytoplasm are used.

Next, we select 14 texture features based on the density histogram, difference statistic value, co-occurrence matrix. There are many sorts of texture features that are familiar in the field of image processing. We select 14 features among several dozen texture features, by discussing the characteristics of each feature. The selected 14 features are adequate for our proposed method.

(3-6) density histogram

Skewness, kurtosis, energy and entropy are used. In the following equation, MEM means average, and VAR means variance. The letter l means the number of bright level, and $p(l)$ means the distribution of normalized histogram.

$$\mathrm{MEN} = \sum_{l=0}^{L-1} l p(l) \quad (2)$$

$$\mathrm{VAR} = \sum_{l=0}^{L-1} (l - \mathrm{MEN})^2 \, p(l) \quad (3)$$

ISSN: 1727-1983

$$SKW = \frac{1}{VAR^3}\left\{\sum_{l=0}^{L-1}(l-MEN)^3 p(l)\right\}^2 \quad (4)$$

$$KRT = \frac{1}{VAR^2}\sum_{l=0}^{L-1}(l-MEN)^4 p(l) \quad (5)$$

$$EGY = \sum_{l=0}^{L-1} p^2(l) \quad (6)$$

$$EPY = -\sum_{l=0}^{L-1} p(l)\log p(l) \quad (7)$$

(3-7) difference statistic value

Mean, contrast, energy, entropy, and variance are used. The difference statistic value is essentially the same as co-occurrence matrix method in the following.

(3-8) co-occurrence matrix

Angular second moment, entropy, correlation, variance, and inverse difference moment are used. In the following equations, when a pixel is j at δ far from the noticed pixel i, we assume that the probability that brightness of both pixels are l_i and l_j is expressed by $P_\delta(l_i, l_j)$. It is difficult to physically explain the features obtained by co-occurrence matrix.

$$P_x(l_i) = \sum_{l_j=0}^{L-1} P_\delta(l_i, l_j) \quad (8)$$

$$P_y(l_j) = \sum_{l_i=0}^{L-1} P_\delta(l_i, l_j) \quad (9)$$

$$\mu_x = \sum_{l_i=0}^{L-1} l_i P_x(l_i) \quad (10)$$

$$\mu_y = \sum_{l_j=0}^{L-1} l_j P_y(l_j) \quad (11)$$

$$\sigma_x^2 = \sum_{l_i=0}^{L-1}(l_i-\mu_x)^2 P_x(l_i) \quad (12)$$

$$\sigma_y^2 = \sum_{l_j=0}^{L-1}(l_j-\mu_y)^2 P_y(l_j) \quad (13)$$

$$ASM = \sum_{l_i=0}^{L-1}\sum_{l_j=0}^{L-1}\{P_\delta(l_i, l_j)\}^2 \quad (14)$$

$$EPY = -\sum_{l_i=0}^{L-1}\sum_{l_j=0}^{L-1} P_\delta(l_i, l_j)\log\{P_\delta(l_i, l_j)\} \quad (15)$$

$$CRR = \frac{\sum_{l_i=0}^{L-1}\sum_{l_j=0}^{L-1} l_i l_j P_\delta(l_i, l_j) - \mu_x\mu_y}{\sigma_x\sigma_y} \quad (16)$$

$$VAR = \sum_{l_i=0}^{L-1}\sum_{l_j=0}^{L-1}(l_i-\mu_x)^2 P_\delta(l_i, l_j) \quad (17)$$

$$IDM = \sum_{l_i=0}^{L-1}\sum_{l_j=0}^{L-1}\frac{1}{1+(l_i-l_j)^2} P_\delta(l_i, l_j) \quad (18)$$

(4) Experimental procedure

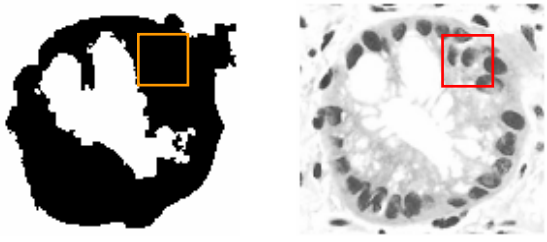
All the images for our study are afforded by the division of pathology, Kanto Central Hospital, Japan. As shown in Table2, there are 6 images of not malignant, 19 images of gastric tumors, and 8 images of gastric cancer. The original images have the size of 2240×1680 pixels and 24 bit color images (8 bit for each color component). First, we downsize the original image to 40% image and 20% image of the original one. Man uses 40% images for extracting shape features and 20% images for extracting texture features. In 40% image, we manually extract the gland that the whole shape exists within the images. As shown in Table2, we extract 28 glands from all the not malignant images, 84 glands from all the gastric adenoma, and 46 glands from all the gastric cancer.

Table2: Sample Number of Gastric Tumors

Lesion	Number of case	Number of Gland
Not Malignant	6	28
Gastric Adenoma	19	84
Gastric Cancer	8	46

Next, man extracts the shape of nucleus and cytoplasm by binarization with Laplacian histogram method and discriminant method, after pre-processing and post-processing for the gland images downsized to 40%. We compute numerical features as described in the previous section, based on the obtained shapes of nucleus and cytoplasm.

Similarly for the images downsized to 20%, we perform binarization and extract the shape of cytoplasm. We decide the region for texture analysis. In the obtained cytoplasm region, we search the small region with size of 16×16 pixels, which is comprised within the cytoplasm region as shown in Figure 11(a). In the same position of downsized grey scale image shown in Figure 11(b), we compute 14 texture features as described in the previous section.



(a)Cytoplasm region (b)Y image

Figure 11: Computation of texture features

After computation of shape and texture features, the principal component analysis (PCA) is performed for all of 54 features. The principal component analysis can reduce the dimension of data. We perform the method in consideration of contribution ratio and factor loadings of each component. We require the principal component scores obtained by PCA in our method.

ISSN: 1727-1983

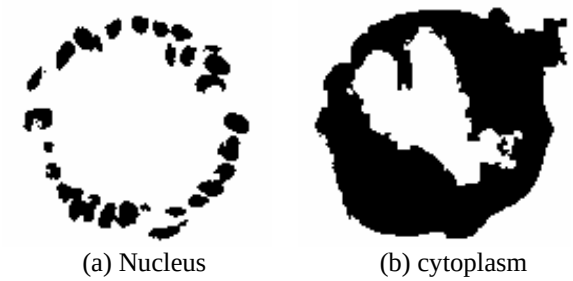
After computation of principal component scores, man performs the discriminant analysis. In discriminant method, we performs F test for homoscedasticity, t test for difference of average, stepwise method for selecting the smaller number of features, and discriminant method with Mahalanobis distance.

Results

We first confirmed the shape extraction of our method for nucleus and cytoplasm of grand. Figure 12 shows the result of binarization of the original image. Figure 12(a) is the obtained image of nucleus, and figure 12(b) is that of cytoplasm. Both of the images seem to have much noise such as other parts of images which have the same intensity of nucleus and cytoplasm. Figure 13 shows the obtained image after elimination of noise included in Figure 12. For elimination of the small noise, we performed filtering with 3×3 median filter, closing operation, and next for the large noise we eliminated the large connected regions in image except the cavity of grand. Figure 13(a) shows the extraction of nucleus, and figure 13(b) shows the extraction of cytoplasm. The results are confirmed by a pathologist that they are sufficient results as an automatic extraction.

(a) Nucleus (b) cytoplasm

Figure 12: Binarization by discriminant method with Laplacian histogram



(a) Nucleus (b) cytoplasm

Figure 13: Shape extraction after elimination of noise by filtering and closing

Next we performed the classification of each gland into the three categories such as not malignant, gastric adenoma and gastric cancer. The results of classification are shown in Table 3. Three cases (not malignant, gastric adenoma, and gastric cancer) of left side mean the results diagnosed by a pathologist, and those of the upper side mean the results of our method. As shown in Table 3 the case of not malignant is correctly classified at the ratio of 96%, that of gastric adenoma is classified at the ratio of 93%, and that of gastric cancer is classified at the ratio of 82%. The total ratio of classification reached 90.5%. Although those ratios in Table 3 are obtained from each gland, total assessment is performed using the features of many glands in the actual diagnosis. Table 4 shows the results of classification for each case by linear combination of several glands in the same case. Six cases of not malignant are correctly classified at the ratio of 100%, nineteen cases of gastric adenoma are classified at the ratio of 100%, and eight cases of gastric cancer are classified at the ratio of 88%. The total ratio of diagnosis by our proposed method reached 96.9%.

Table 3: Discriminant Analysis (8 Feature Parameters)

		Assigned Lesion			
		Not Malignant	Gastric Adenoma	Gastric Cancer	
Actual Lesion	Not Malignant	27 (96%)	0 (0%)	1 (4%)	28
	Gastric Adenoma	0 (0%)	78 (93%)	6 (7%)	84
	Gastric Cancer	4 (9%)	4 (9%)	38 (82%)	46

Table 4: Classification result of gastric tumor

		Assigned Lesion			
		Not Malignant	Gastric Adenoma	Gastric Cancer	
Actual Lesion	Not Malignant	6 (100%)	0 (0%)	0 (0%)	6
	Gastric Adenoma	0 (0%)	19 (100%)	0 (0%)	19
	Gastric Cancer	0 (0%)	1 (12%)	7 (88%)	8

Discussion

There are some problems for making our proposed method as the diagnosis supporting system. We manually extracted each gland from the original images, since we require the gland that has the whole shape. Although there are some glands in the original images, we can only use several glands among them. Many of glands are eliminated before analysis of our method. We need take digital images with the wider range of raw tissue. Moreover automatic extractions of gland have to be performed for the diagnosis system. As further research, more samples should be necessary, since we think that the number of samples is not enough for assessment of our method.

In this study the dyeing of tissue is performed at Kanto Central Hospital, Tokyo. Since the dyeing results such as color and density are different every hospital, we must investigate the results of our method according to the difference of dyeing. We used R component of

ISSN: 1727-1983

RGB color basis for extraction of nucleus, G component for cytoplasm, and Y component of YIQ color basis for texture analysis. If the dyeing result is considerably different, man has to perform compensation of each selected color component. Binarization seems to have good performance of our method for the samples dyed at Kanto Central Hospital.

As post-processing we performed closing processing such as expanding and shrinking. Small noises of binary image are eliminated by the closing processing. The number of closing process was selected through the try and error process. The number of closing process needs be automatically obtained in our method. After closing process man performed labeling processing, the largest area among dark region is selected as gland, and the largest area among light region is selected as cavity. Although there are a few cavities in one gland region in some cases, the cavities is neglected except the largest area. The influence by neglected cavity needs be also investigated.

Our method uses the texture features for diagnosis. The pathologists diagnose a gastric tumor with the shape and density of nucleus within the gland. But the cases of gastric adenoma and gastric cancer have the nucleus with large and atypical shape, and the nucleus of those cases agglutinate each other. Since we could not extract each nucleus from the image, texture features are computed. Other features of nuclear shape are necessary for the system with further performance.

From the result our method has good performance of diagnosis of gastric tumor. For some misclassification of diagnosis, we require more information of gastric tumor about shape of nucleus and gland. As the further research we make automatic system of diagnosis with more performance.

Conclusions

This study deals with features of gastric tumors for computer aided diagnosis. We extracted the nucleus and cytoplasm region by image processing. We performed selection of color component and gamma compensation for our method as preprocessing. Binarization with Laplacian histogram method and discriminant method are performed after the preprocessing. A post-processing of binarization is performed by median filtering and closing processing. We obtained image for diagnosis, which is confirmed by a pathologist. After the extraction of nucleus and cytoplasm 40 features were computed based on shape of gland, and 14 features were computed based on texture within the shape of cytoplasm. We performed PCA for 54 features of all the shape features and texture features. After PCA 7 features were selected for the diagnosis by stepwise method. Twenty eight gland were manually extracted from the 6 healthy tissue, eighty four gland were extracted from 19 gastric adenoma, and forty six gland were extracted from 8 gastric cancer. We performed diagnosis of gastric cancer by the selected seven features. The ratio of classification reached 90.5% for each gland. The ratio of diagnosis for 33 cases reached 96.9% by our proposed method. The results showed that our proposed method was efficient as the diagnosis supporting system.

References

[1] BAMFORD, P., LOVELL, B. (1998): 'Unsupervised cell nucleus segmentation with active contours', Signal Processing, **71**, pp.203-213

[2] O'GORMAN, L., SNDERSON, A.C, PRESTON JR., K. (1983): 'Image segmentation and nucleus classification for automated tissue section analysis', Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp.89-94

[3] CUISENAIRE, O., ROMEO, E., VERAART, C., MACQ, B. (1997): 'Automatic segmentation and measurement of axons in microscopic images', SPIE Proceedings Society of Photo-Optical Instrumentally Engineers, **3661-94**, pp.4211-4216

[4] CHAN, F. H. Y., LAM, F. K., ZHU, H. (1998): 'Adaptive Thresholding by Variational Method', IEEE Transactions on Image Processing, **7**, pp.468-473

[5] THIRAN, J. P., MACQ, B. (1996): 'Morphological Feature Extraction for the Classification of Digital Images of Cancerous Tissues', IEEE Transactions on Biomedical Engineering, **43**, pp.1011-1020

[6] OKII, H., UOZUMI, T. (2003): 'Automatic Feature Extraction from Breast Tumor Images Using Artificial Organisms', IEICE Trans. INF. & SYST., **E86-D**, pp.964-972

[7] WU, H. S., GIL, J. (2003): 'Linear Clustering for Segmentation of Color Microscopic Lung Cell Images', Journal of Imaging Science and Technology, **47**, pp.161-170

[8] MARGHANI, K. A., DLAY, S. S., SHARIF, B. S., SIMS, A. J. (2003): 'Automated morphological analysis approach for classifying colorectal microscopic images', Proc. of SPIE, **5267**, pp.240-249

[9] ESGAIR, A. N., NAGUIB, R. N. G. (1998): 'Microscopic Image Analysis for Quantitive Measurement and Feature Idenitification of Normal and Cancerous Colonic Mucosa', IEEE Transactions on Information Technology in Biomedicines, **2**, pp.197-203

[10] KHOUZANI, K. J., ZADEH, H. S. (2003), 'Multiwavelet Grading of Pathological Images of Prostate', IEEE Transactions on Biomedical Engineering, **50**, pp.697-704

[11] BAAK, J. P. A., DELEMARRE, J. F. M., LANGLEY, F. A. et al (1986): 'Grading ovarian tumors, Evaluation of decision making by different pathologists', Anal Quant. Cytol. Histol, **8**, pp.349-353

ISSN: 1727-1983