Classification of Gastric Tumors using Shape Features of Gland

Toshiyuki Tanaka  Member  (Keio University, tanaka@appi.keio.ac.jp)
Yoshitaka Uchino  Non-member  (Keio University)
Teruaki Oka  Non-member  (Kanto Central Hospital)

Keywords : gastric tumor, gland, shape feature, classification, computer diagnosis

1. Introduction

In the field of imaging technology, studies of automation and quantification of tumor cell imaging diagnosis have been done for these three decades. The dominant application is development of diagnosis support system which solves the problem of pathologist shortage. For example, PAP smear screening systems for cervical cancer have already proposed to handle with problems such as nucleus segmentation and classification. But these studies mainly focus on prostate, colon and skin tumors, and few studies focus on gastric tumors.

Gastric cancer and gastric adenoma sometimes need to be distinguished. Recently, possibility of existence of spectrum between gastric cancer and gastric adenoma was reported and distinction criterion of these two tumors is often discussed. It is because pathologists have to depend on their own knowledge and experience when they diagnose. Therefore, it is necessary for pathologists to clarify natures of gastric tumors by objective and quantitative evaluation that are helpful when they diagnose gastric tumors.

In this paper, we propose the method for measuring objectively that helps pathologists diagnose gastric tumors by numerically expressing the same features such as atypism and pleomorphism that they pay attention to when they diagnose.

2. Materials and Methods

First, we extract nucleus region and cytoplasm region from each gland by binarization with implementation of Laplacian-histogram method and Discriminant analysis. Next, we computed 40 shape features from these extracted regions. Then, combining cytoplasm region with Y component gland image, we computed 14 texture features. And we did principal components analysis about 54 features so that principal components scores were efficient to use in discrimination.

3. Results

Table 1 shows the number of samples using our study. Figure 1(a) shows a sample of the original image, Fig. 1(b) shows the shape of nuclear after binarization, and Fig. 1(c) shows the shape of cytoplasm after binarization. As shown in Table 2 the case of not malignant is correctly classified at the ratio of 96% using 8 principal components selected by stepwise method, 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 91%. Although those ratios in Table 2 are obtained from each gland, total assessment is performed using the features of many glands in the actual diagnosis.

4. Discussion

Our method uses the texture features for diagnosis. The pathologists diagnose a gastric tumor with the shape and density of nucleus within gland. But the cases of gastric adenoma and gastric cancer have 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.

5. Conclusions

This paper deals with features of Gastric tumors for computer aided diagnosis. We performed diagnosis of gastric cancer by the selected seven features. The ratio of classification reached 91% for each gland. The results showed that our proposed method was efficient as the diagnosis supporting system.

References

Classification of Gastric Tumors using Shape Features of Gland

Toshiyuki Tanaka* Member
Yoshitaka Uchino* Non-member
Teruaki Oka** Non-member

Recently in Japan, pathologists have been in short supply, while each pathological diagnosis requires a substantial amount of time because each analyte must be inspected by multiple pathologists for adequate diagnosis. This paper deals with the classification method of gastric cancer and gastric adenoma, using image processing and pattern analysis. We first select the R component and G component from the RGB basis of the digital image, and the Y component from the YIQ basis for our system. After pre-processing, we automatically extracted the shape of the nucleus and cytoplasm. After many inspections, we selected 40 features for shape of the nucleus and cytoplasm and 14 features for texture within the cytoplasm for assessment of tumors. Principal component analysis, F test of homoscedasticity, t test of difference of average, stepwise method for selecting the smaller number of features, and discriminant method using Mahalanobis distance were all performed. Total ratio of diagnosis reached 96.9%, showing the validity of our proposed method.

Keywords: gastric tumor, gland, shape feature, classification, computer diagnosis

1. Introduction

In Japan, a lack of doctors has been often reported in many branches of medicine. The number of pathologists seems to be particularly small. Since pathological diagnosis represents a subjective assessment by each pathologist, multiple pathologists must examine the same specimen and collate those results for accurate diagnosis, and each pathological diagnosis requires a substantial amount of time. A diagnosis support system using objective and quantitative approaches would thus be extremely useful. In medical image engineering, automatic and quantitative diagnosis systems (1)–(11) have previously been studied by many researchers. Most such research has aimed at the development of diagnosis support systems addressing the lack of pathologists. As a representative system, a prescreening system for uterus cytodiagnosis is currently in use. The pathologist must still perform many steps in image processing (1)–(5) for these diagnosis systems. In research for tumor imaging, human input is still required for the following: detection of the region of breast cancer (6); region extraction of glomerulus in kidney; and so on. Although the types of tumor involved increase year by year, the objects of most research are the lung (7), colon (8) (9), prostate (10) and ovaries (11), with few studies examining gastric tumor.

Well-differentiated gastric cancer and gastric adenoma differ from each other, and the classification has recently been discussed by pathologists. However, no adequate automatic diagnosis systems for gastric tumor are available for such studies as described above, and pathologists have been eagerly awaiting a system for gastric tumor. Morphological classification of colorectal microscopic images (4) has been reported as a recent diagnosis support system. This method allows features of tumors to be obtained from the whole image, but displays a low classification ratio for images with more background region.

In this paper we propose a diagnosis support system for gastric tumor based on morphological features of cytoplasm and nucleus. In this system, gastric tumor is classified into gastric cancer and gastric adenoma, using numerical features obtained from the morphology based on the points of observation of pathologists. For numerical conversion of morphological features, manual extraction of region of interests (ROIs) was performed, with pre-processing selection of color components comprising image, contrast enhancement, binarization using Laplacian histogram and discriminant methods. Each tumor was classified into gland structure and nucleus in pre-processing. From the divided regions, 40 shape features and 14 texture features were classified, and gastric tumors were classified by the discriminant method with the stepwise method using the obtained features. A small number of features were selected for diagnosis by the stepwise method. Finally, tumors were classified into three categories using the obtained numerical features: not cancer; gastric adenoma; and gastric cancer. All images for our research were stained at Kanto Central Hospital, since the results of dyeing differ slightly between hospitals.

2. Materials and Methods

2.1 Selection of used Image Pathologists clas-
Table 1. Group classification of gastric biopsy specimens

<table>
<thead>
<tr>
<th>Group</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Normal tissue and benign lesion</td>
</tr>
<tr>
<td>2</td>
<td>Benign lesion with aberrant tissue</td>
</tr>
<tr>
<td>3</td>
<td>Boundary case between benign and malignant tumor</td>
</tr>
<tr>
<td>4</td>
<td>Tissue at increased risk for cancer</td>
</tr>
<tr>
<td>5</td>
<td>Complete cancer</td>
</tr>
</tbody>
</table>

Classification of gastric tumors using 5 categories as shown in Table 1, based on the shape of the nucleus and cytoplasm as shown in Figure 1. Group 1 includes normal tissue and obvious benign lesions, while group 5 corresponds to complete gastric cancer. Groups 2 to 4 represent boundary cases between benign tissue and malignant cancer. Classification of groups 2 to 4 is difficult even when diagnosis is manually performed by the pathologist. We classified gastric tumor into group 1, group 3 and group 5 as shown in Figure 1. Original images were sorted into the correct category by the pathologist. In this study, images for groups 2 and 4 were not used.

Figure 2 shows the process of our proposed method. First, we manually extracted gland regions within rectangles with sufficient size to include the whole shape of the gland, as shown in Figure 3. Glands for our research were selected only when the whole shape existed within the original image. We did not select defective glands that were clipped on imaging. The number of usable glands was thus relatively small. Next, we investigated color information for our classification method. Figure 4 shows a gland image manually clipped from the original image.

2.2 Pre-processing and Binarization

Figure 5 and 6 show R component and G component images from the RGB color basis and binarization images. We decided after much discussion that R component would be used for extraction of the nucleus within the gland region, while G component would be used for extraction of cytoplasm from the image. Figure 7 shows the Y component image from YIQ color basis.
Fig. 8. Regions of cytoplasm and cavity

te image from the YIQ color basis (12) after conversion from the RGB color basis using equation (1). Y component images were used for computing texture features, since the Y component corresponds to the brightness of each pixel.

\[
\begin{bmatrix}
Y \\
I \\
Q
\end{bmatrix} =
\begin{bmatrix}
0.299 & 0.587 & 0.114 \\
0.596 & -0.274 & -0.322 \\
0.211 & -0.522 & 0.311
\end{bmatrix}
\begin{bmatrix}
R \\
G \\
B
\end{bmatrix}
\]

We performed binarization of R component and G component images using the Laplacian histogram method and a discriminant method called the Ohtsu method (13). Figure 5(b) and figure 6(b) show the results of binarization. Figure 5(b) clearly shows nucleus shape and figure 6(b) shows cytoplasm shape. Filtering and closing processing were performed as post-processing (12). Filtering was performed 3 × 3 median filter to eliminate the small noise of binary imaging. Closing processing was undertaken to eliminate the relative large noise. Next, the labeling process was performed to extract the large connected elements in the image. The largest area of dark color was selected as cytoplasm, while the largest area of light color was selected as a cavity within the cytoplasm. Other connected components were eliminated as noise.

2.3 Selection of Shape Features Here we first explain the shape features (12). We select 40 shape features based on area, length, chord and axis, equivalent shape, and other shape features.

(1) Features based on area
Area of nucleus, area of cytoplasm within the gland, area of cavity, and the area of gland were used. The area of cytoplasm corresponds to region A in Figure 8, area of cavity is seen as regions B and C, and gland area is region A+B+C.

(2) Features based on length
Length of gland and cavity and ferret diameter were used.

(3) Features based on chord and axis
Maximal segment, maximal section, mean horizontal chord, mean vertical chord and mean vertical section were used. Maximal segment corresponds to line A in Figure 9, maximal section is line B, and mean vertical section is line C.

(4) Features based on equivalent shape
The equivalent shape is the ellipse as shown in Fig.10, with area equivalent to the gland and half the maximum length of the gland. Length 2a in Fig.10 is the maximum length of the gland, and length 2b was obtained using area of gland and length 2a. 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, and edge ratio of rectangle were used.

(5) Other shape features
Inertia moment, tension factor, degree of dispersion, Heywood roundness, hydraulic depth, Waddle disk radius, and ratio of nucleus to cytoplasm were used.

2.4 Selection of Texture Features Next, we selected 14 texture features based on density histograms, differences in statistical values and co-occurrence matrix (12). Many sorts of texture features are familiar in the field of image processing. We selected 14 features from among several dozen texture features, after discussing the characteristics of each feature. The selected 14 features were adequate for our proposed method.

(1) Density histogram
Skewness, kurtosis, energy and entropy were used. In the following equation, μ means average, and σ means variance. The letter l means the numerical brightness level, and p(l) means the distribution in a normalized histogram.

\[
\mu = \sum_{l=0}^{L-1} lp(l) 
\]

\[
\sigma^2 = \sum_{l=0}^{L-1} (l - \mu)^2 p(l)
\]
In the above equations, SKW means skewness of 3rd moment around the centroid, and shows how skewed the density histogram is from a symmetrical shape such as normal distribution.

\[
\text{SKW} = \frac{1}{\sigma^3} \sum_{l=0}^{L-1} (l - \mu)^3 p(l) \quad \cdots \cdots \cdots \cdots (4)
\]

KRT means kurtosis of the 4th moment around the centroid, and shows how converged the density histogram is to the mean.

\[
\text{KRT} = \frac{1}{\sigma^4} \sum_{l=0}^{L-1} (l - \mu)^4 p(l) \quad \cdots \cdots \cdots \cdots (5)
\]

EG means energy, and angular second moment. This value becomes large when pixels with a specific brightness value are present in clumps.

\[
\text{EG} = \sum_{l=0}^{L-1} p(l)^2 \quad \cdots \cdots \cdots \cdots (6)
\]

EP means entropy of the image, and this value becomes large when many pixels with differing brightness exist in the same image. In this paper, SKW, KRT, EG and EP were used as texture features obtained from density histograms.

(2) Difference statistic values

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

(3) Co-occurrence matrix

We defined probability \(P_b(i, j)\) that the pixel with the brightness \(j\) exists at the pixel at the distance of \(\delta = (r, \theta)\) as shown in Fig. 11 from the pixel with brightness \(i\). Furthermore, \(\delta = (r, \theta)\) means distance \(r\) in the direction of \(\theta\). We computed the following features by probability \(P_b(i, j)\). After discussing the parameters \(r\) and \(\theta\), we used \(r = 4\) and \(\theta = 0^\circ\) for our system.

\[
P_x(i) = \sum_{j=0}^{L-1} P_b(i, j) \quad \cdots \cdots \cdots \cdots (8)
\]

\[
P_y(j) = \sum_{i=0}^{L-1} P_b(i, j) \quad \cdots \cdots \cdots \cdots (9)
\]

\(P_x(i)\) means the summation of components in each row, and \(P_y(j)\) means the summation of components in each column in the co-occurrence matrix.

\[
\mu_x = \sum_{i=0}^{L-1} i P_x(i) \quad \cdots \cdots \cdots \cdots (10)
\]

\[
\mu_y = \sum_{j=0}^{L-1} j P_y(j) \quad \cdots \cdots \cdots \cdots (11)
\]

\(\mu_x\) means the average of each row, and \(\mu_y\) means the average of each column in the matrix.

\[
\sigma^2_x = \sum_{i=0}^{L-1} (i - \mu_x)^2 P_x(i) \quad \cdots \cdots \cdots \cdots (12)
\]

\[
\sigma^2_y = \sum_{j=0}^{L-1} (j - \mu_y)^2 P_y(j) \quad \cdots \cdots \cdots \cdots (13)
\]

\(\sigma^2_x\) means the variance of each row, and \(\sigma^2_y\) means the variance of each column in the matrix.

In this paper, we used the following features of the co-occurrence matrix.

\[
\text{ASM} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (P_b(i, j))^2 \quad \cdots \cdots \cdots \cdots (14)
\]

ASM means angular second moment, and this value becomes large when pixels with specific brightness exist in clusters.

\[
\text{EPY} = - \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} P_b(i, j) \log\{P_b(i, j)\} \quad \cdots \cdots \cdots \cdots (15)
\]

EPY means entropy, and this value becomes large when many pixels with differing brightness value exist in the image.

\[
\frac{\sum_{i=0}^{L-1} \sum_{j=0}^{L-1} ij P_b(i, j) - \mu_x \mu_y}{\sigma_x \sigma_y} \quad \cdots \cdots \cdots \cdots (16)
\]

CRR means the correlation, and this value shows the correlation between values of rows and columns in the matrix.

\[
\text{VAR} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} (i - \mu_x)^2 P_b(i, j) \quad \cdots \cdots \cdots \cdots (17)
\]

VAR means variance, and this value becomes large when many pixels with brightness distant from the average exist in the image.

\[
\text{IDM} = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} \frac{1}{1 + (i - j)^2} P_b(i, j) \quad \cdots \cdots \cdots \cdots (18)
\]

IDM means the inverse difference moment, and this value becomes large when many pixels with low brightness exist in the image.
Table 2. Sample number of gastric tumors

<table>
<thead>
<tr>
<th>Lesion</th>
<th>Case number</th>
<th>Number of glands</th>
</tr>
</thead>
<tbody>
<tr>
<td>Not Malignant</td>
<td>6</td>
<td>28</td>
</tr>
<tr>
<td>Gastric Adenoma</td>
<td>19</td>
<td>84</td>
</tr>
<tr>
<td>Gastric Cancer</td>
<td>8</td>
<td>46</td>
</tr>
</tbody>
</table>

2.5 Experimental Procedure  
All images for our study were provided by the Division of Pathology at Kanto Central Hospital, Japan. As shown in Table 2, images included not malignant (n=6), gastric tumor (n=19) and gastric cancer (n=8). The original images include the size of pixels and 24-bit color images (8 bit for each color component). First, we downsized original images to a 40% image and 20% images of the original. The 40% images were used for extracting shape features, while 20% images were used to extracting texture features. In the 40% image, we manually extracted glands for which the whole shape was present within the images. As shown in Table 2, we extract 28 glands from all not malignant images, 84 glands from all gastric adenomas, and 46 glands from all gastric cancers.

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

Similarly for images downsized to 20%, binarization was performed and cytoplasm shape was extracted. We decided the region for texture analysis. In the obtained cytoplasm region, we searched a small region with a size of $16 \times 16$ pixels within the cytoplasm region, as shown in Figure 12(a). In the same position of the downsized grey-scale image shown in Figure 12(b), we computed 14 texture features as described in the previous section.

After computation of shape and texture features, principal component analysis (PCA) was performed for all 54 features. PCA can reduce the dimension of data, and this method was utilized after considering the contribution ratio and factor loadings of each component. We required principal component scores obtained by PCA for our method. After computation of principal component scores, discriminant analysis was performed. In the discriminant method, F testing was performed for homoscedasticity, t testing for differences in means, stepwise method was used for selecting a smaller number of features, and the discriminant method using Mahalanobis distance.

3. Results  
We first confirmed shape extraction of our method for nucleus and cytoplasm of the gland. Figure 13 shows the result of binarization of the original image. Figure 13(a) shows the obtained image of the nucleus, and figure 13(b) shows the cytoplasm image. Both images seem to display high levels of noise, such as other parts of images that have the same brightness as the nucleus or cytoplasm. Figure 14 shows the obtained image after elimination of noise included in Figure 13. For elimination of small noise, we performed filtering with a $3 \times 3$ median filter, closing operation, while for large noise we eliminated large connected regions on images except the gland cavity. Figure 14(a) shows extraction of the nucleus, and figure 14(b) shows extraction of cytoplasm. A pathologist confirmed that the results represented sufficient automatic extraction.

Next we performed classification of each gland into the 3 categories of 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) on the left side are results diagnosed by a pathologist, while the upper side are results from our method. As shown in Table 3, cases of not malignant were correctly classified at a ratio of 96%, compared to 93% for gastric adenoma and 82% for gastric cancer. Total ratio of classification reached 91%. Although those ratios in Table 3 were obtained from each gland, total assessment was 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 ma-
lignant were correctly classified at a ratio of 100%, 19 cases of gastric adenoma were classified at a ratio of 100%, and 8 cases of gastric cancer were classified at a ratio of 88%. Total ratio of diagnosis using our proposed method reached 97%.

4. Discussion

Some problems remain before our proposed method can be used as a diagnosis support system. We manually extracted each gland from the original images, since we required glands with a whole shape. Although original images sometimes included multiple glands, only limited numbers were able to be used. Many glands were eliminated before analysis using our method. We need to take digital images with a wider range of raw tissue. Moreover, automatic extractions of glands need to be performed for the diagnosis system. As further research, more samples should be examined, as we think that the number of samples is insufficient for assessment of our method.

In this study, tissue dyeing was performed at Kanto Central Hospital, Tokyo. Since dyeing results such as color and density differ between hospitals, we must investigate the results of our methods according to differences in dyeing. We used the R component of the RGB color basis for extraction of the nucleus, the G component for cytoplasm, and the Y component of the YIQ color basis for texture analysis. If dyeing results differ considerably, compensation of each selected color component would need to be performed. Binarization seems to offer good performance for our method using samples dyed at Kanto Central Hospital.

As post-processing, we performed closing processing such as expanding and shrinking. Small noise from binary images was eliminated by closing processing. The number of processes in the closing process were selected by trial and error, but needs be automatically obtained in our method. After closing process, labeling processing was performed, with the largest of the dark regions selected as gland, and the largest of the light regions selected as cavity. Although multiple cavities were present in a single gland region in some cases, all cavities other than the largest were neglected. The influence of neglecting cavities must also be investigated.

Our method uses texture features for diagnosis. Since features of statistic value differences display a particularly strong correlation to features of the co-occurrence matrix, one of them are not necessary for analysis. In consideration of the dynamic range of computed values with both methods, we selected the co-occurrence matrix method. Parameters $r = 1, 2, 3, 4, 5$ and $\theta = 0^\circ, 45^\circ, 90^\circ$ as candidate parameters. Each index shows large differences according to variation of $r$. Conversely, each index shows small difference according to variation of $\theta$. Direction of imaging is also not necessarily the same. From the results of feature computation, we selected $r = 4, \theta = 0^\circ$.

Pathologists diagnose gastric tumors from shape and density of the nucleus within the gland. However, cases of gastric adenoma and gastric cancer have nuclei with large and atypical shape, and the nucleus in those cases agglutinate each other. Since we could not extract each nucleus from the image, texture features were computed. Other features of nuclear shape are necessary to improve the system. The results indicate our method offers good performance for diagnosis of gastric tumor. For some misclassification of diagnosis, more information about gastric tumor is required for nucleus and gland shape. Further research is underway to improve performance of this automatic system.

5. Conclusions

This study dealt with features of gastric tumors by computer-aided diagnosis. We extracted nucleus and cytoplasm regions by image processing, and performed selection of color components and gamma compensation as preprocessing. Binarization with Laplacian histograms and discriminant methods was performed after preprocessing. Post-processing of binarization was performed by median filtering and closing processing. We obtained images for diagnosis that were confirmed by a pathologist. After extraction of the nucleus and cytoplasm, 40 features were computed based on shape of gland, and 14 features were computed based on texture within the shape of the cytoplasm. PCA was performed for all 54 shape and texture features. After PCA, 7 features were selected for diagnosis by stepwise method. Next, 28 glands were manually extracted from 6 healthy tissues, 84 glands were extracted from 19 gastric adenomas, and 46 glands were extracted from 8 gastric cancers. Diagnosis of gastric cancer was performed using the selected 7 features. Ratio of classification reached 91% for each gland. Ratio of diagnosis for the 33 cases reached 97% using our proposed methods. These results show that our proposed method is efficient as a diagnosis support system.

(Manuscript received Oct. 3, 2005, revised April 25, 2006)
References


Toshiyuki Tanaka (Member) received a B.S. in 1982 from the Department of Instrumentation Engineering, Faculty of Engineering, Keio University. He completed his doctoral program in 1989, receiving D. Eng. From 1989, he became an instructor at the Faculty of Science and Technology. He was a visiting Researcher in 1995-1996 at RWTH Aachen, Germany. Now he is an associate professor, Department of Applied Physics and Physico-Informatics, Faculty of Science and Technology, Keio University. He is engaged in research on medical image processing, biometric authentication, signal processing, GPS and its application, and nonlinear oscillation theory.

Yoshitaka Uchino (Non-member) received a B.S. in 2003 from the Department of Applied Physics and Physico-Informatics, Faculty of Science and Technology, Keio University, and received a M.S. in 2005 from the School of Science and Technology, Keio University. Now he works at UNIQLO Co. His research interest during his university career was computer-assisted diagnosis of gastric tumors.

Teruaki Oka (Non-member) received the degree of M.D. (medical doctor) in 1979 from the Department of Medicine, Iwate Medical University, and entered the medical school at University of Tokyo in 1979. He became an assistant professor at the Graduate School of Medicine, University of Tokyo in 1998, and now serves as Chief of the Division of Pathology and Clinical Laboratory, Kanto Central Hospital. He is a member of the Japanese Society of Pathology and a councilman of the Japanese Society of Clinical Cytology.