Feature Extraction and Analysis on Xinjiang High Morbidity of Kazak Esophageal Cancer by Using Comprehensive Feature

Murat Hamit *murat.hamit@xjmu.edu.cn*
College of Medical Engineering Technology
Xinjiang Medical University
Urumqi, 830011, China

Fang Yang *1090466411@qq.com*
College of Medical Engineering Technology
Xinjiang Medical University
Urumqi, 830011, China

Abdugheni Kutluk *akutluk@hotmail.co.jp*
College of Medical Engineering Technology
Xinjiang Medical University
Urumqi, 830011, China

Chuanbo Yan *ycbsky@163.com*
College of Medical Engineering Technology
Xinjiang Medical University
Urumqi, 830011, China

Elzat Alip *elzat003@qq.com*
College of Medical Engineering Technology
Xinjiang Medical University
Urumqi, 830011, China

Weikang Yuan *454069946@qq.com*
College of Medical Engineering Technology
Xinjiang Medical University
Urumqi, 830011, China

Abstract

Image feature extraction technology has been widely applied in image data mining, pattern recognition and classification. Esophageal cancer is a common digestive malignant tumor, China is one of the world's highest incidence and mortality rates of esophageal cancer among the countries, Xinjiang Uygur Autonomous Region is a high incidence area of esophageal cancer, and the kazak is the esophageal cancer high-risk groups. In this paper, we selected 60 advanced esophageal X-ray barium images, half of them are constricted esophagus and the rest are ulcerous esophagus. Firstly, image preprocessing approaches were used to preprocess images. Secondly, extracting the gray-scale histogram features and the GLCM features of the images, then composing the two features into comprehensive feature. Finally, using Bayes discriminant analysis to verify the classification ability of the comprehensive feature. The classification accuracy for constricted esophagus was 86.7%, for ulcerous esophagus was 93.3%.

Keywords : Xinjiang High Morbidity of Kazak, Esophageal Cancer, Comprehensive Feature, Feature Extraction, Image Classification.

1. INTRODUCTION

Esophageal cancer is a common digestive malignant tumor and its worldwide morbidity and mortality in a common cancer were ranked in the sixth and eighth, respectively [1]. The geography difference on its pathogenesis is the most obviously among all cancers, China is one of the world's highest incidence and mortality rates of esophageal cancer among the countries[2], the world's annual increase of 300 thousand patients with esophageal cancer, about half occurred in China[3]. Esophageal cancer is one of the most common cancers in China, the incidence and mortality are above all cancers in the world, esophageal cancer has become a seriously threat to people’s life and health, it has been proved to be one of the focus cancer of research [4]. Xinjiang Uygur Autonomous Region is a high incidence area of esophageal cancer, the kazak is the esophageal cancer high-risk groups, whose esophageal cancer mortality rate up to 155.9/106, higher than the average level 15.23/106 in China, so this disease is the regional focus malignant tumors to prevent [5].

CAD can assist the clinician to discover lesions and improve diagnostic accuracy through medical image processing technology as well as other possible physiological and biochemical methods, combining with computer analysis and calculation technology. Image feature extraction technology a cross discipline, has been widely applied in image data mining, pattern recognition and classification. It’s not only included in the computer vision technology, but also involved in the image processing, the aim is to extract the image invariant features through computer analysis and solve practical problems [6]. The gray level histogram method and the GLCM method are usually used in the image feature extraction. There is no related research on image feature extraction of high morbidity of kazak esophageal cancer in Xinjiang. Therefore, this study means a lot to kazak esophageal cancer in Xinjiang.

2. METHODLOGY

In this paper, we selected two kinds of advanced esophageal X-ray barium images, which were acquired from First Affiliated Hospital, Xinjiang Medical University of China. Classifying them under the clinician’s guidance, Selecting 60 images, half of them are constricted esophagus and the rest are ulcerous esophagus. For these selected images, firstly, image preprocessing approaches were used to preprocess images before the analysis algorithm was applied in order to keep the useful image information under different conditions, converting RGB image to the grayscale intensity image by eliminating the hue and saturation information while retaining the luminance, through median filter to remove the image noise and using histogram equalization to enhance the contrast; secondly, to extract the gray-scale histogram features and the GLCM features of the images, then composing the two features into comprehensive feature; Finally, using Bayes discriminant analysis to verify the classification ability of the comprehensive feature.

2.1 Image Preprocessing

Medical images obtained from the hospital not only contain valid information, but there are also some noise and the variation caused by the rotation and translation. So, the images cannot be used directly. The purpose of image preprocessing is to process the image obtained from the medical apparatus, removing the noise due to external disturbance, enhancing the contrast between the normal tissue and the pathological tissue, getting the interest region of strong visuality and contrast, thus providing the better support for the doctor's clinical diagnosis and the subsequent image processing[7]. In this paper, for the purpose of image feature extraction, the original image needs to be preprocessed to convert color, remove noise and enhance the contrast, reducing the influence of the factors mentioned above.

After the image preprocessing, the quality of X-ray barium image was improved significantly, the contrast of normal tissue and pathological tissue was enhanced apparently and the details of image were more clearly (see figure 1 and 2 lower right), which laying a good foundation for the subsequent image extraction of high morbidity of kazak esophageal cancer in Xinjiang, furthermore, improving the subsequent image classification accuracy.

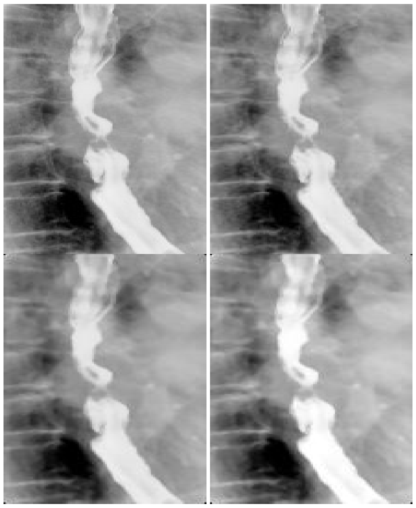

FIGURE1: Preprocessing results of constricted esophageal X-ray. Upper left: Original image; Upper right: Color transformed; Lower left: Removed noise; Lower right: Enhanced.

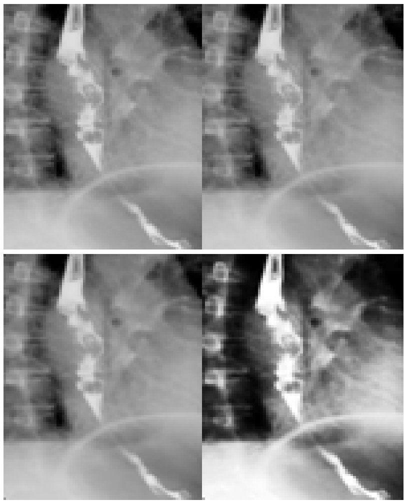

FIGURE 2: Preprocessing results of ulcerous esophageal X-ray. Upper left: Original image; Upper right: Color transformed; Lower left: Removed noise; Lower right: Enhanced.

2.2 Gray-scale Histogram Feature Extraction

A digital image histogram is a gray-scale discrete function, the following formula represents the image histogram definition [8].

$$H(i) = \frac{n_i}{N}, i = 0, 1, \cdots, L - 1$$

i represents gray level, L represents the number of gray level types, *ni* represents the number of *i*, *N* represents the total number of pixels in an image. The equation describes the percentage of the number of i account the total number of pixels in an image. The abscissa is the gray level, the vertical axis is the frequency of the gradation. Calculating the following statistics to reflect the image characteristic values [9] on the basis of the gray-scale histogram.

(1) Average value: Mean reflects the average gray value of an image.

$$u = \sum_{i=0}^{L-1} iH(i)$$

(2) Variance: Variance, which is a measure for the width of the histogram, reflects the discrete distribution of a gray-scale image numerically. That is, the difference between the gray level and the average.

$$\sigma^2 = \sum_{i=0}^{L-1} (i-\mu)^2 H(i)$$

(3) Skewness: Skewness reflects the degree of asymmetry in the histogram distribution, the greater skewness represents the histogram distribution is more asymmetric.

$$\mu_s = \frac{1}{\sigma^3} \sum_{i=0}^{L-1} (i-\mu)^3 H(i)$$

(4) Kurtosis: Kurtosis, which measures whether the distribution of gray-scale image is very focused on the average gray nearby, reflects the image gray-scale distribution when it closes to the mean. The smaller Kurtosis represents the histogram distribution is more concentrative.

$$\mu_k = \frac{1}{\sigma^4} \sum_{i=0}^{L-1} (i-\mu)^4 H(i)$$

(5) Energy: Energy reflects the uniform degree of gray-scale distribution, the more uniform gray-scale, the larger energy.

$$\mu_N = \sum_{i=0}^{L-1} H(i)^2$$

2.3 GLCM Feature Extraction

GLCM reflects the microtexture of an image area. It described the gray correlation of pixel pairs by a certain spatial relationship [10-11]. GLCM is defined as that starting at the gray level of pixel i from the image, calculating the probability of the pixel j from the distance d and the angleθ, reflecting the spatial correlation of gradation between two points in the image.

The space GLCM usually can not be used as texture analysis feature due to its complex computation, so we extracted the texture feature on the basis of the GLCM. Haralick [12] proposed 14 kinds of GLCM texture quantitative methods on the basis of the character of texture. We calculated the following statistic characters:

(1) Angular second moment: Also known as energy, reflecting the uniformity of gray, the rougher texture moments, the greater energy.

$$ASM = \sum_{i=0}^{L-1} \sum_{j=0}^{L-1} [P(i,j,d,\theta)]^2$$

(2) Entropy: Using entropy to detect the complexity of the image space and internal uniformity, the thinner texture, the greater entropy.

$$ENT = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1} P(i,j,d,\theta)\log P(i,j,d,\theta)$$

(3) Inertia Moment: Also known as Contrast, which represents the total change of gray in a small image area. This parameter reflects the degree that high value matrixes away from the diagonal line.

$$CON = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}(i-j)^2 P(i,j,d,\theta)$$

(4) Correlation: Correlation described the similarity of GLCM between rows and columns, which reflects the extend length of a certain gray value along a certain direction, the longer extension, the greater correlation value.

$$COR = \frac{\sum_{i=0}^{L-1}\sum_{j=0}^{L-1} i\cdot j\cdot P(i,j,d,\theta) - \mu_x\mu_y}{\sigma_x\sigma_y}$$

μ_x represents the mean of gray value, μ_y represents the smooth mean,
σ_x^2 represents the grayscale variance, σ_y^2 represents the smooth variance.

$$\mu_x = \sum_{i=0}^{L-1} i \sum_{j=0}^{L-1} P(i,j,d,\theta)$$

$$\mu_y = \sum_{i=0}^{L-1} j \sum_{j=0}^{L-1} P(i,j,d,\theta)$$

$$\sigma_x^2 = \sum_{i=0}^{L-1}(i-\mu_x)^2 \sum_{j=0}^{L-1} P(i,j,d,\theta)$$

$$\sigma_y^2 = \sum_{i=0}^{L-1}(i-\mu_y)^2 \sum_{j=0}^{L-1} P(i,j,d,\theta)$$

(5) Inverse difference moment: Reflecting the concentration of high value matrixes on the main diagonal line, the greater value, the higher concentration.

$$IDM = \sum_{i=0}^{L-1}\sum_{j=0}^{L-1}\frac{P(i,j,d,\theta)}{[1+(i-j)^2]}$$

3. RESULTS AND DISCUSSION

3.1 Gray-scale Histogram Feature Extraction Result

Extracting the average vaule (avg), variance (var), skewness (sk), kurtosis (kur), energy (ene) of all the X-ray images selected, and composing them into vector component(see table 1).

Image type	avg	var	sk	kur	ene
constricted esophagus	147.8329761	1501.590717	-0.82923649	-0.20165276	0.01113704
	141.8383579	2327.279287	-0.91476189	0.38049406	0.00748258
	145.4243755	4301.323428	-0.33738664	-0.5308574	0.00683868
	…	…	…	…	…
	159.1537156	2545.407865	-0.27509711	-1.37506984	0.00992652
ulcerous esophagus	151.0811004	732.2059987	-0.50390497	1.49034815	0.01210474
	153.7212397	778.7894774	-0.2101274	-0.18883096	0.01077037
	158.4681136	780.5992869	-0.60767563	0.45453657	0.01107275
	…	…	…	…	…
	140.3558579	2590.974543	-0.68579449	-0.10677265	0.00725628

TABLE 1: Two kinds of image features extracted by gray-scale histogram method.

3.2 GLCM Feature Extraction Result

Extracting the GLCM features of all the images selected, electing the Angular Second Moment, the Entropy, the Inertia Moment, Correlation and Inverse difference moment from 4 directions of pixel distance, d = 1, θ = {0 °, 45 °, 90 °, 135 °}, calculating the average value and variance of each feature at four directions and 10 features were obtained, then composing them into vector component (see table 2).

Image type	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
constricted esophagus	0.0386	0.0035	3.6072	0.1279	0.9848	0.4915	0.0468	0.0006	0.8676	0.0267
	0.0325	0.0042	3.7719	0.1494	0.8085	0.2943	0.0465	0.0002	0.8200	0.0357
	0.0304	0.0030	3.8440	0.1210	0.9388	0.3269	0.0462	0.0003	0.8039	0.0295
	…	…	…	…	…	…				
	0.0351	0.0056	3.6790	0.1975	0.7982	0.3763	0.0465	0.0002	0.8424	0.0429
ulcerous esophagus	0.0257	0.0043	4.0761	0.1926	1.8209	0.6707	0.0450	0.0008	0.7459	0.0483
	0.0272	0.0041	3.9627	0.1707	1.2833	0.474	0.0456	0.0006	0.7672	0.0436
	0.0342	0.0034	3.7483	0.1347	1.1431	0.5102	0.0463	0.0005	0.8326	0.0305
	…	…	…	…	…	…				
	0.0245	0.0047	4.1009	0.2099	1.6412	0.6626	0.0452	0.0007	0.7247	0.0570

TABLE 2: Two kinds of image features extracted by GLCM method.

3.3 Bayes Discriminant Analysis

Discriminant analysis is a statistical method, which can estimate the discriminate objects' category according to their observation results. It has been widely used in the medical filed. Robin Hanson used Bayes analysis method of unsupervised models' classification in different complexity.

Using the gray-scale histogram features and GLCM features extracted to classify the images through Bayes discriminant analysis. The classification accuracy for constricted esophagus was 86.7%, for ulcerous esophagus was 93.3% (see table 3).

Image Type	Bayes Discriminant Classification		total	accuracuy rate(%)
	correct	error		
Constricted Esophagus	26	4	30	86.7
Ulcerous Esophagus	2	28	30	93.3

TABLE 3: Discriminant analysis between two kinds of image by comprehensive features.

4. CONCLUSION

According to the difference between histogram distribution and texture distribution in different types of high morbidity of kazak esophageal cancer in Xinjiang, combining with the characteristics of esophageal cancer, we proposed the feature extraction based on histogram and GLCM method, combining the extracted gray-scale histogram features and GLCM features into the comprehensive feature, and then evaluated the feature's classification ability through Bayes discriminant analysis. Experimental results show that using comprehensive feature for image classification has a high accuracy, and feature classification ability is different when classifying different images, which provides a new direction for the research of computer aided diagnosis system for the high incidence of kazak esophageal cancer in Xinjiang Uygur Autonomous Region.

5. ABBREVIATIONS

CAD: computer aided diagnosis; GLCM: gray level co-occurrence matrix

6. ACKNOWLEDGEMENTS

This work was supported by the National Natural Science Foundation of China, Xinjiang, 30960097 and 81160182.

7. REFERENCES

[1] Parkin DM, Bray FI, Devesa SS. "The Cancer burden in the year 2000. The global picture". Eur J Cancer, vol.37, pp.S4 - S66, Aug. 2001.

[2] ZHAO Fengjuan, YUN Miaoying, ZHANG Yan, "XU Yu.Research progress on Xinjiang kazak esophageal cancer". Journal of central university for nationalities, vol.18, pp.85-89, Mar. 2009.

[3] CHEN Xiangchuan, PANG LiJuan, LI Feng. "Research progress on Xinjiang kazak esophageal cancer". Agricultural reclamation of medicine, vol.28, pp.384-387, May. 2006.

[4] ZHANG Huixia, CHEN Yan, YIN Dong, DENG Yanchao, MA Yanqing, JuLaiDi. "Discussion on risk factor of Xinjiang kazak esophageal cancer". Modern preventive medicine, vol.36, pp.1804-1806, Oct. 2009.

[5] GUO Hui, DING Jianbing, ZHANG Wei, ZHANG Tong: "Gene research progress on Xinjiang kazak esophageal cancer". Basic medicine and clinical, vol.30, pp.428-430, Apr. 2010.

[6] WANG Zhirui, YAN Cailiang: "Review on image feature extraction method". journal of jishou university, vol.30, pp.52-56, May. 2011.

[7] Murat Hamit, ZHOU Jingjing, YAN ChuanBo, LI Li, CHEN Jianjun, HU Yanting, KONG Dewei. "Feature extraction and analysis of Xinjiang local liver hydatid CT images based on gray-scale histogram". Tech review, vol.30, pp.79-83, May. 2010.

[8] JIN Hua. "The research of segmentation and feature extraction method for medical images based on density clustering". M.A. thesis.: Jiangsu university, Zhenjiang 2005.

[9] WANG Shuqin. “Research on feature selection and extraction of Liver CT aided diagnosis system”. M.A. thesis, Shanghai jiaotong university, Shanghai 2010.

[10] TONG Longzheng, WANG Lei, CHEN Hairong etc. “Gray Level Co-occurrence Matrix Analysis on liver fibrosis images”. capital medical university, vol.24, pp.240-242, Mar. 2003.

[11] JIN Jing, SHI Li. Intelligent diagnosis of hepatocellular carcinoma based on the CT images[J]. system engineering research of China and clinical rehabilitation, vol.13, pp.5919-5922, Mar. 2009.

[12] Haralick R M. Statistical and structural approaches to texture. Proceedings of the IEEE, vol.67, pp.786-840, May. 1979.