Journal of Medical Systems (2019) 43: 306
https://doi.org/10.1007/s10916-019-1429-8

IMAGE & SIGNAL PROCESSING

Colorectal Cancer Diagnostic Algorithm Based on Sub-Patch Weight Color Histogram in Combination of Improved Least Squares Support Vector Machine for Pathological Image

Kai Yang[1,2] · Bi Zhou[1] · Fei Yi[1] · Yan Chen[1] · Yingsheng Chen[1,2]

Received: 27 February 2019 / Accepted: 25 July 2019 / Published online: 14 August 2019

Abstract
In order to improve the diagnostic accuracy of colon cancer, a novel classification algorithm based on sub-patch weight color histogram and improved SVM is proposed, which has good approximation ability for complex pathological image. Our proposed algorithm combines wavelet kernel SVM with color histogram to classify pathological image. Firstly, the pathological image is divided into non-overlapping sub-patches, and the features of sub-patch histogram are extracted. Then, the global and local features are fused by the sub-patch weighting algorithm. Then, the RelicfF based forward selection algorithm is used to integrate color features and texture features so as to enhance the characterization capabilities of the tumor cell. Finally, Morlet wavelet kernel-based least squares support vector machine method is adopted to enhance the generalization ability of the model for small sample with non-linear and high-dimensional pattern classification problems. Experimental results show that the proposed pathological diagnostic algorithm can gain higher accuracy compared with existing comparison algorithms.

Keywords Colon cancer · Pathological image · Diagnostic · Color histogram · Morlet wavelet · Support vector machine · Sub-patch weight · RelicfF strategy

Introduction

Colorectal cancer (CRC), also known as bowel cancer and colon cancer [1], is the development of cancer from the colon or rectum (parts of the large-intestine) [2]. A cancer is the abnormal growth of cells that have the ability to invade or spread to other parts of the body [3]. Signs and symptoms may include blood in the stool, a change in bowel movements, weight loss, and feeling tired all the time, which harm the health of the people [4]. Therefore, how to achieve early diagnosis has become a hot field of research direction. Early detection and early treatment are helpful to patient.

Recently, medical imaging and pathology play an important role in the diagnosis and treatment of colon cancer [5]. Intestinal examination includes CT, MRI, PET, ultrasound scanning, blood test, molecular testing of tumors [6], and colonoscopy biopsyy [7]. However, the final cancer diagnosis is to obtain pathological sections by biopsy, and then to determine whether there is a cancer area in the pathological image by the pathologist [8]. Colorectal pathological section is obtained by puncture biopsy, and then the tissue is stained with hematoxylin and eosin. The pathological image of large-intestine can be obtained by taking pathological section of large-intestine with CCD microscope. Hematoxylin and eosin stain or haematoxylin and eosin stain (H&E stain or HE stain) is one of the principal stains in histology. It is the most widely used stain in medical diagnosis and is often the gold standard; for example, when a pathologist looks at a biopsy of a suspected cancer, the histological section is likely to be stained with H&E [9]. A combination of hematoxylin and eosin, it produces blues, violets, and reds. Finally, the pathologist analyzes the tissue section image and combines his long-term accumulated clinical diagnosis experience to determine whether the tissue has cancerous.

This article is part of the Topical Collection on *Image & Signal Processing*

✉ Yingsheng Chen
chengyingsheng@hotmail.com

[1] Department of Radiological Intervention, Shanghai Sixth People's Hospital East Campus Affiliated to Shanghai University of Medicine & Health Science, Shanghai 201306, China

[2] Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

In order to reduce the human differences in the diagnosis process, and relieve the pressure of pathologists to read high-intensity images, computer-aided diagnosis of colon cancer has far-reaching significance [10]. Pathological examination is the golden standard for cancer diagnosis and classification, and one of the most important medical examinations [11]. For computer-aided diagnosis of colon cancer, many studies have focused on developing computational quantification tools to provide objective analysis and evaluation [12]. Because the diagnostic accuracy of the system is only related to the core recognition algorithm in the diagnostic system, once the core algorithm is determined, no matter how many times a pathological image is diagnosed, the same result will be obtained, which can reduce the difference of human reading and avoid misdiagnosis [13]. Therefore, computer-aided identification of intestinal cancer is particularly important.

Automatic analysis algorithm of large-intestine pathological biopsy image has become the focus of domestic and foreign scholars. The development of colon cancer computer aided diagnosis system is very meaningful, which can be automatic classification of large-intestine pathological biopsy image and auxiliary medical diagnosis [14]. Combining the guidance of pathologist with the structural characteristics of pathological images, some effective methods for colorectal cancer diagnosis using pathological images are proposed by domestic and foreign scholar. Literature [15] proposed a novel recognition algorithm for colon cancer using pathological images based on low level features, where an improved GLRLM combined with HOG (Histogram of Oriented Gradient) are used for recognizing cancerous images. On the one hand, Color image is quantized into three colors through K-means by making full use of the color features and then the run length texture features are calculated, overcoming the disadvantages of the traditional GLRLM, which directly quantizes it into fixed grayscale gradations without fully considering the characteristics of the image. On the other hand, the gradient of some images are more obvious. Thus, HOG and improved GLRLM are combined. Then, mRMR for feature selection and SVM for test are executed. Experimental results show that the accuracy of combined algorithm is higher than that of the single improved GLRLM algorithm. Similarly, Literature [16] combines the cancer images characteristics of HOG GLRLM and histological characteristics and uses feature selection algorithm to reduce the impact of irrelevant features considering the redundancy of the information contained in different feature sets. In literature [17], recognition algorithm for colon cancer using pathological images is proposed by combining statistical characteristics and Delaunay features based on object-oriented method: In order to take the background knowledge and the content of image into account. First, PCA-KMEANS is used for preprocessing. Then, Heuristic search is carried out for segmentation, In this procedure three objects are segmented (lumen with epithelial cell cytoplasm, nucleus, stroma). After that, features based on statistics and Delaunay are extracted. Finally, mRMR method for feature selection and SVM for classification are carried out. Experimental results show that the pathological features gain higher accuracy compared with low level features.

Since the texture information used in the existing large-intestine pathological image classification algorithm can not describe the detailed features of the sample, the extracted features can not reflect the essential characteristics of the data well. In addition, various existing classification algorithms do not solve the problem of classification of unbalanced categories, which leads to more classifications tending to learn the characteristics of large-scale samples, while ignoring the classification of small samples, but the colon cancer classification diagnosis of pathological images with medium-small categories of samples often contain more valuable information. Because the SVM algorithm can avoid over-learning problems and has good generalization performance [18], it solves the unique advantages exhibited by small sample, nonlinear and high-dimensional mode classification problems, while the wavelet function has good time-frequency [19] and has very strong approximation and fault tolerance for training samples. Therefore, this paper proposes a novel colon cancer classification algorithm based on wavelet analysis theory and SVM architecture, which has good approximation ability for complex samples. The algorithm combines wavelet kernel SVM and color histogram to classify pathological images. Firstly, the pathological image is segmented pixel by pixel, and the sub-region histogram features are extracted. The sub-patch weighted algorithm proposed by this paper is used to fuse global and local features. The ReliefF-based forward selection algorithm integrates color features and texture features to enhance the characterization capabilities of the tumor. Finally, the Morlet wavelet kernel-based least squares support vector machine is used to classify features to enhance the generalization ability of small sample with nonlinear and high-dimensional mode classification problems.

Color histogram features based on sub-patch weight

Color is the first feature of the image presented to human vision. The global color histogram (GCH) proposed in literature [20] is the most commonly used one. Because of its simplicity in computation and invariance in rotation and translation, this method is widely used in image color features. However, the global color histogram has a disadvantage that cannot be ignored. It does not take into account the spatial distribution features of the image, which will lead to differences in object recognition. In order to weaken this influencing factor and increase the spatial distribution features, the image can be divided into 3×3 sub-patches in literature [13],

and then the color histogram feature of each sub-patch is extracted respectively. Through the image patch processing, the extracted color features are endowed with certain spatial features; literature [21] holds that the importance of different regions in an image is different, so different sub-patches are pre-weighted. However, the spatial distribution of different images is not invariable, and the importance of each sub-patch area is also changing to some extent. Therefore, the weight of each sub-patch is set as a fixed value in literature [22], which will undoubtedly affect the accuracy and does not have universal applicability.

Color space quantization

In this paper, HSV color space which is more suitable for human visual features is selected and non-uniform quantization is carried out. The specific formula for converting RGB color space to HSV color space is shown in formula (1)–(3):

$$\mathrm{H}=\begin{cases}\arccos\dfrac{(2R-B-G)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}, B\le G\\ 2\pi-\arccos\dfrac{(2R-B-G)}{2\sqrt{(R-G)^2+(R-B)(G-B)}}, B>G\end{cases} \quad (1)$$

$$\mathrm{S}=\frac{\max(R,G,B)-\min(R,G,B)}{\max(R,G,B)} \quad (2)$$

$$\mathrm{V}=\frac{\max(R,G,B)}{255} \quad (3)$$

where H represents hue, and range from 0° to 360°; S is for saturation; V represents the brightness or value, and the range is [0, 1].

In order to reduce the vector dimension of the histogram, where the hue H is divided into 8 parts and the saturation S and brightness V are both 3 parts, it is quantized as a 72-dimensional color feature vector $f = 9H + 3S + V$. The range of f is $(0, 1, \cdots, 71)$, and the specific quantization process is shown in eq. (4).

$$\mathrm{H}=\begin{cases}0,\ H\in[0^\circ,45^\circ)\\ 1,\ H\in[45^\circ,90^\circ)\\ 2,\ H\in[90^\circ,135^\circ)\\ 3,\ H\in[135^\circ,180^\circ)\\ 4,\ H\in[180^\circ,225^\circ)\\ 5,\ H\in[225^\circ,270^\circ)\\ 6,\ H\in[270^\circ,315^\circ)\\ 7,\ H\in[315^\circ,360^\circ)\end{cases},\quad S=\begin{cases}0,\ S\in[0,0.3)\\ 1,\ S\in[0.3,0.6)\\ 2,\ S\in[0.6,1)\end{cases},\quad V=\begin{cases}0,\ V\in[0,0.3)\\ 1,\ V\in[0.3,0.6)\\ 2,\ V\in[0.6,1)\end{cases} \quad (4)$$

Sub-patch weighted color histogram extraction

Color histogram based on sub-patch is an improvement on the image global histogram,which is to include the spatial distribution information of the image [23, 24]. This method usually divides the whole image into several patches in a certain way, and then calculates the color information of each patch separately. Then object recognition is carried out with the patch histogram, and spatial distribution information is introduced into the global histogram [25]. In order to improve the object characterization performance, this paper combines weight strategy on the basis of sub-patch, and the specific steps are as follows:

1) The image firstly is resized so that all images have the same size, and then the image is evenly divided into 3×3 sub-patches.
2) Extracting Harrris interest points from the image, the total number of image interest points is n. And then the number of interest points $d_i (i = 1, 2, \cdots, 9)$ contained in each of the nine sub-patches is counted, so as to calculate the weight $W_i (W_i = d_i/n_i)$ occupied by each sub-patch in image feature extraction.
3) Feature $f_i (i = 1, 2, \cdots, 9)$ is extracted for each sub-patch of, and then the weighted feature vector $\mathrm{F} = \sum_{i=1}^{9} W_i \times F_i$ of the divieded image can be obtained according to the weight W_i of each sub-patch obtained in the previous step.

Figure 1 is the processing diagram of the pathological image patch of colorectal cancer (CRC). According to the number of interest points in each patch area, the weight of each patch can be obtained.

The concept of texture feature is often used in image recognition and computer vision and other related fields. It mainly reflects some changes of gray level on the surface of the

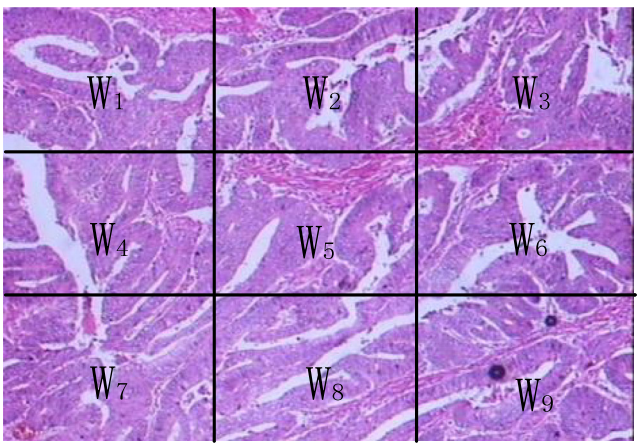

Fig. 1 Image texture feature extraction

object, which can demonstrate the proprietary properties of the object itself. By studying the psychological features of human visual perception, Tamur et al. proposed texture features that are more easily perceived. Texture feature vectors correspond to six texture attributes in psychology, including coarseness, contrast, directionality, linearity, regularity and roughness; the three components of roughness, contrast and directionality have better application value in texture synthesis, so only these three feature vectors are used as texture features [22].

Coarseness

Coarseness can reflect the most essential features of texture, mainly measuring the size of granularity in texture mode [26]. The smaller the granularity is, the finer the texture image is, and contrarily, the rougher the texture image is. The main calculation steps are as follows:

1) First, calculate the average gray value of the pixel within the window range of $2t \times 2t$, $t \in (1, 2, \cdots, 5)$ in the image, and $h(i,j)$ represents the gray value at the pixel point (i,j), as shown in eq. (5).

$$\mathrm{A}_i(x,y) = \sum_{i=x-2t-1}^{x+2t-1} \sum_{j=y-2t-1}^{y+2t-1} \frac{h(i,j)}{2^{2t}} \tag{5}$$

2) Calculate the average gray difference of each pixel in the horizontal and vertical directions between non-overlapping windows, as shown in (6), (7):

$$E_{t,h}(x,y) = |A_t(x-2^{t-1},y) - A_t(x+2^{t-1},y)| \tag{6}$$

$$E_{t,v}(x,y) = |A_t(x,y+2^{t-1}) - A_t(x,y-2^{t-1})| \tag{7}$$

Find the t value for each pixel that maximizes$E$, and then substitute t into $S_{best}(x,y) = 2t$ to get the optimal sizeS_{best} based on each pixel.

3) Finally, the mean value of the optimal size$S_{best}$of all pixels is calculated, which represents the roughness$\mathrm{F_{crs}}$of the image, as shown in eq. (8):

$$\mathrm{F}_{crs} = \sum_{i=1}^{m} \sum_{j=1}^{n} S_{best}(i,j) \tag{8}$$

Contrast

Contrast refers to the brightest and the darkest colors in the image area, which is proportional to the contrast [26]. By calculating the mean value, variance and other relevant statistical features of each pixel's neighborhood, the image contrast variable can be calculated, as shown in the following equations:

$$F_{con} = \delta / \sqrt[4]{\alpha} \tag{9}$$

$$\alpha = u/\delta^4 \tag{10}$$

where δ represents the standard deviation of the image gray value, δ^2 represents the variance, α represents a kurtosis value of the image gray value, and μrepresents the four-order moment.

Directionality

Directionality refers to the centralization or divergence of texture in the direction. First, the gradient vector at the position of each pixel should be calculated. The magnitude and direction of gradient vectorΔGare defined as shown in eqs. (11) and (12):

$$|\Delta \mathrm{G}| = \frac{(|\Delta_H| + |\Delta_V|)}{2} \tag{11}$$

$$\theta = \tan^{-1}\left(\frac{\Delta_V}{\Delta_H}\right) + \frac{\pi}{2} \tag{12}$$

where Δ_{H} represents the change of gradient vector ΔGin the horizontal direction, and Δ_{V}represents the change of gradient vector ΔGin the vertical direction. Their change is generated by convolution of images with two 3×3 matrices respectively.

$$\mathrm{H}_D(k) = N_\theta(k / \sum_{i=0}^{n-1} N_\theta(i \tag{13}$$

In formula (13), $\mathrm{H}_D(k)$ represents the histogram, $N_\theta(k)$ is the number of pixels satisfying $|\Delta \mathrm{G}| \geq t$when the value range of θ is $(\pi(2k-1)/2^n, \pi(2k+1)/2^n)$, where t is the size of the given threshold, and n is the number of quantized direction angle grades. For the histogram with obvious directional image, there will be a peak, so it is only necessary to calculate the sharpness of the peak in the histogram to obtain a direction of the image, as shown in eq. (14):

$$\mathrm{F}_{dir} = \sum_{p}^{n_p} \sum_{\varphi \in W_p} \left(\varphi - \varphi_p\right)^2 H_D(\varphi) \tag{14}$$

In formula (14), n_p denotes the number of peaks in the histogram. For a peak$p$, $W_p$denotes the quantization space, $\Delta_{\mathrm{p}}$is the corresponding value when the histogram is maximized in $W_p$interval, F_{dir} is the global direction of the image.

ReliefF based forward selection algorithm

The ReliefF algorithm [16] can be applied to multi-class sample cases and regression problems, and it also solves the problem of missing data. Although ReliefF algorithm can effectively select features with high correlation with categories, it does not consider the correlation between features, so there may be redundant features in feature subsets. In order to eliminate redundant features, this study adopts the ReliefF based forward selection algorithm (ReFS). First, features are sorted from high to low according to the weight obtained by applying ReliefF, and the sorted feature set is represented by FR. Then, the forward selection step is executed. According to the order in FR, starting with the first feature, the improved features of the classifier will be added to the feature subset one by one. Through these two steps, the optimal feature subset with correlation and non-redundancy features is obtained.

Normalized multi-feature fusion

Given set $S = \{s_1, s_2, \cdots, s_m\}$containing m samples, where each sample contains n features, $s_i = \{s_{i-1}, s_{i-2}, s_n\}$. it is supposed that the category label of sample s_i is $l_i \in L$, and L is the set of sample labels (only 2 categories can be used); The difference between two samples s_i and s_j on the feature t is defined as:

$$d(t, s_i, s_j) = \left| \frac{s_{it} - s_{jt}}{\max_t - \min_t} \right| \tag{15}$$

where *max* represents the maximum value of feature t in the sample set, and *min* represents the minimum value of feature t in the sample set. In view of the difference between the range of color histogram feature vectors and texture feature vectors, the sample features are normalized as the standard vectors so as to realize the equivalence of the two feature vectors' influence on similarity. And the extracted 72-dimensional color histogram vectors and 36-dimensional texture feature vectors are normalized to [0,1] intervals.

Morlet wavelet kernel-based least squares support vector machine

It is generally known that support vector machine transforms low-dimensional data into high-dimensional space through kernel mapping [27]. Its purpose is to find an optimal hyperplane for linear classification. However, the approximation of any function in high-dimensional space by support vector machine is not always so accurate, which makes the actual classification results too rough. The main reason is that the current kernel function cannot generate a set of complete orthogonal function, so it is necessary to introduce a basic function with complete space transformation [28].

Because the traditional support vector machine has the problem that the kernel function cannot fit all the data, it directly affects the classification effect. In recent years, scholars at home and abroad are studying the improved support vector machine algorithm. Therefore, based on the least square support vector machine, this paper introduces the Morlet wavelet kernel function. The sparse kernel function is helpful to improve the classification accuracy of the model and the convergence speed of iteration. According to literature [14], for any multidimensional wavelet function, tensor product theory can be adopted to decompose it into one-dimensional product form. The expression of wavelet kernel function is as follows:

$$K(x_i, y_i) = \prod_{i=1}^{N} h\left(\frac{x_i - y_i}{a}\right) \tag{16}$$

where $h(x)$ is the wavelet generating function and a is the corresponding scaling factor. Test and simulation show that the widely used wavelet function Morlet can meet the support vector kernel function condition, and its wavelet kernel function is:

$$h(x) = \cos(1.75x)\exp\left(-\frac{x^2}{2}\right) \tag{17}$$

Therefore, the wavelet kernel function based on Morlet is:

$$\begin{aligned} K(x_i, y_i) &= \prod_{i=1}^{N} h\left(\frac{x_i - y_i}{a}\right) \\ &= \prod_{i=1}^{N} \left(\cos\left(1.75\left(\frac{x_i - y_i}{a}\right)\exp\left(-\frac{\|x_i - y_i\|^2}{2a^2}\right)\right.\right. \end{aligned} \tag{18}$$

Then the expression of the classification hyper-plane of the Morlet wavelet kernel-based least squares support vector machine is written as follows:

$$\min\left\{\frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j\prod_{i=1}^{N}h\left(\frac{x_i - y_i}{a}\right) + \sum_{i=1}^{n}\frac{\alpha_i^2}{2\gamma} - \sum_{i=1}^{n}\alpha_i y_i + b\sum_{i=1}^{n}\alpha_i\right\} \tag{19}$$

Experimental results and analysis

Experimental data

The pathological images of the large-intestine are the color pictures taken by the microscope's camera after HE staining of the tissue sections. The microscope is a 20-times objective lens with the resolution of 350×350. These 180 pathological images of large-intestine were

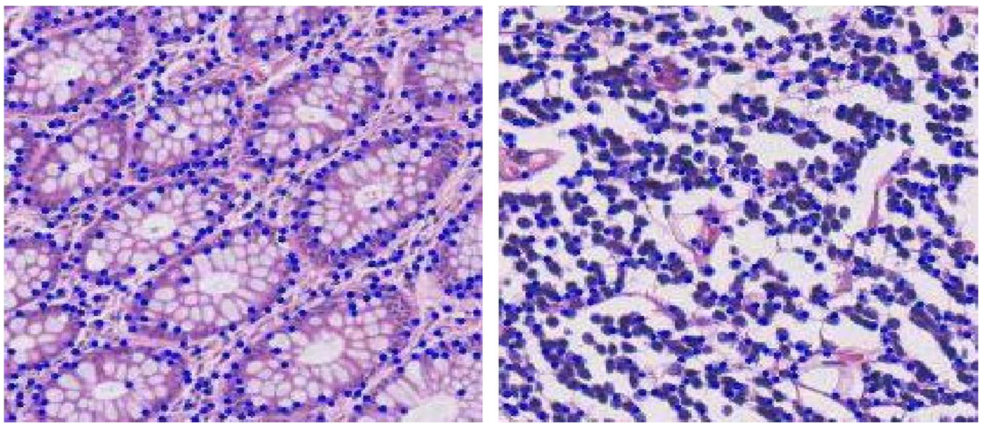

Fig. 2 Comparison for different large-intestine tissue section; (a) Normal pathological section; (b) Cancer tissue

marked by professional pathologists, where 90 were normal and 90 were cancerous. The experimental platform and software environment was Intel Core i5-430m CPU 2.3 GHz, with a memory of 4GB, and the simulation software was Matlab R2013b. Support vector machine (SVM) is based on LIBSVM software package developed by Dr. Lin zhiren from Taiwan. The training images are 108, of which 54 were normal and 54 are cancerous. The test images are 72, and the experimental results are obtained by taking the mean value for 10 times.

In order to compare the experimental results fairly, some existing state-of-art algorithms are selected as comparison algorithm, all of which are classified by SVM. In addition, to analyze the validity of the Morlet wavelet kernel-based least squares support vector machine, the proposed color histogram feature is also used in traditional SVM for classification detection, which is named Color_SVM. All the experiments were conducted on the same data set, and the evaluation methods are based on 10-fold cross-validation.

In this paper, the data set is divided into training set (70%) and test set (30%). The training set is used to select the optimum feature subset and construct SVM classification prediction model. The test set is used to evaluate the classification prediction performance of the proposed model. There are four indicators to evaluate the predictive performance of a classifier, namely, Area Under Receiver Operating Characteristic Curve (AUC), Sensitivity (SEN), Specificity (SPE) and Accuracy (ACC).

Table 1 Comparison for different features

Features	AUC	ACC	SEN	SPE
GLCM	0.8009	0.7176	0.6463	0.7902
HoG	0.7369	0.6878	0.5881	0.7876
LBP	0.8349	0.7757	0.7072	0.8436
CM	0.8072	0.7748	0.6458	0.7885
Proposed	**0.8523**	**0.8313**	**0.8197**	**0.8420**

Data features

According to the pathologist's expertise, a normal colonic epithelial cell surrounds a single layer to form the glandular region. When hematoxylin and eosin (HE) were stained, colonic tissue is observed by hematoxylin and eosin (HE) staining. The nuclei of these cells are purple-black and the glandular lumen is white [29]. The epithelial cells are rich in mucin, and as a result, their cytoplasm appears very bright. In this paper, these cytoplasms are represented by blue circles. For simplicity, as mentioned above, epithelial cytoplasm and glandular lumen are called cavity, which is expressed by blue circles. This expression is very close to normal tissue. Other types of objects rarely exist between the corresponding blue circles. The average clustering coefficient of the cavity is higher by using the color pattern. In addition, since the epithelial nuclei are arranged in a single structure, a circular connection between the corresponding purple circles may be generated, resulting in a low clustering coefficient between nuclei; goblet cells disappear in cancerous tissues, mucus is not abundant, and the nuclei are distributed in multiple layers, leading to a higher clustering coefficient. Statistical features are based on the segmentation of objects. For normal pathological images of large-intestine, the distribution of objects is more uniform, and the difference in area is not as large as that of cancerous images [30]. Due to different dimensions of area,

Table 2 Experiment results for deifferent k

Number	AUC	ACC	SEN	SPE
5	0.7309	0.7876	0.6526	0.7852
10	0.7669	0.7889	0.6881	0.7976
15	0.8126	0.7911	0.7072	0.8136
20	0.8372	0.8248	0.7409	0.8385
25	0.8523	0.8313	0.8197	0.8420
30	0.8836	0.8631	**0.8907**	0.8744
35	**0.8911**	**0.8655**	0.8875	**0.8801**
40	0.8878	0.8509	0.8861	0.8800

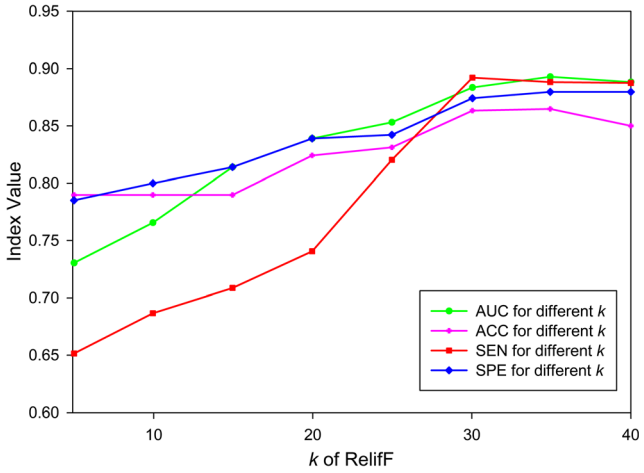

Fig. 3 Classification performance of classifiers with different k of ReliefF

mean value, variance and coefficient of variation are used to measure.

Figure 2 shows a normal pathological section of the large-intestine. The image samples of this experiment are obtained by hematoxylin-eosin staining of large-intestine tissues. The large-intestine glands are seen here in a nearly circular, well-defined, ordered pattern. The glandular cavity and epithelial cells are distributed in the center, and the stroma is distributed in it. The nuclei are located at the margin and between the glands. In normal pathological sections of the large-intestine, the cells are evenly distributed and orderly arranged, and the nuclei and epithelial cells are obvious, while the cells are scattered and disorderly arranged and the nuclei and epithelial cells disappeared in cancerous pathological sections of the large-intestine.

Accuracy is the percentage of the number of cases with correct diagnosis in the total number of cases. Although it can reflect the performance of discriminant diagnosis to some extent, it does not reflect the frequency of false negative and false positive misdiagnosis. Higher accuracy may also be many false negative or false positive. Therefore, sensitivity and specificity are added into the evaluation criteria. Sensitivity and false-positive varied synchronously while specificity and false-negative varied synchronously as diagnostic thresholds are changed.

Color histogram feature analysis

In this paper, color histogram combined with texture feature is selected for image description. In order to quantitatively analyze the effectiveness of selected color histogram, we selects Gray-level co-occurrence matrix texture feature (GLCM) [23], Color moment feature (CM) [24], HoG feature [19] and LBP feature for comparative analysis. All the classifier algorithms adopt the Morlet wavelet kernel-based least squares support vector machine. Table 1 shows the quantitative results of all algorithms, where the optimal results are represented in bold. It can be seen from the results in Table 1 that the proposed histogram feature can effectively describe the appearance of the target, while HoG feature and LBP feature are only used to describe the structure information of the image, and their accuracy is not high. This is because the case image has no obvious boundary information and is greatly disturbed by the background. Therefore, the experimental results have showed the effectiveness of the proposed algorithm.

Qualitative and quantitative analysis

In order to remove the irrelevant and redundant features in the original feature set, this paper adopted the forward selection algorithm for feature screening, i.e. different number of features is combined with Forward Selection respectively for experiments, and the results are listed in Table 2, where the optimal results are represented in bold. Comparing the predictive performance of the classifiers trained by each method, the classifier trained by the ReliefF based forward selection algorithm has the optimum classification performance, and its AUC, accuracy, sensitivity and specificity are all above 80%. Therefore, the ReliefF based forward selection algorithm is selected as the feature selection method in this paper. For the value of k in the nearest neighbor sample in the ReliefF algorithm, this study took 8 values, which are 5, 10, 15, 20, 25, 30, 35, 40, for experiment. Figure 3 shows the AUC, SEN, SE and ACC values of the classifier when k takes different values. According to the graph, when the number of neighbor samples K is 35, the evaluation indexes all get higher values.

In this study, the ReliefF based forward selection algorithm is used to remove the irrelevant and redundant features in the original feature set, and a total of 75 features are selected. The simulation results show that there is a significant correlation between these features and the types of intestinal cancer. The visual distribution of normalized feature sets of benign and malignant intestinal tumors also shows that the distribution of feature values is different intuitively, but it is very difficult to find a threshold from each feature as well as classify the benign and malignant intestinal tumors. Therefore, it is

Table 3 Classification accuracy for different algorithms

Algorithm	GLCM	HoG	LBP	CM	Proposed	Color-SVM
Accuracy	77.91%	71.90%	89.75%	83.28%	96.78%	96.01%

necessary to train SVM classifier with the optimal feature subset to obtain the benign and malignant classification model of large intestinal tumors.

In this paper, the wavelet kernel SVM classifier is selected to train the benign and malignant classifier of large intestinal tumors, and the feature subset and corresponding tag of the training data were used as the input of the SVM classifier. After obtaining a classification model that can predict the benign and malignant intestinal tumors, Bootstrap is conducted 1000 times in the training set and the test set to evaluate the classification and prediction performance of the model.

The AUC, accuracy, sensitivity, specificity and 95% confidence interval of the training concentration model for intestinal cancer classification are respectively: 0.9919±0.0003, 0.9277±0.0017, 0.9537±0.0020, 0.9018±0.0028; the AUC, accuracy, sensitivity, specificity and 95% confidence interval of the test concentration model for intestinal cancer classification were respectively: 0.8521±0. 0047, 0.8313±0.0038, 0.8197±0.0059 and 0.8120±0.0052, where the classification accuracy of different algorithm is shown in Table 3.

Compared with the traditional ROI-based feature extraction, the feature extraction based on color space contains more abundant information, which is conducive to model training. By fusing with texture features and selecting features, the features in the selected feature subset are all valid features. They not only contain color information and spatial information of the tumor image, but also can express heterogeneous information of the tumor. The extraction of these information can improve the classification performance of the model. As a tool to quantify tumor heterogeneity, color histogram combined with texture features has a broad prospect in the prediction of tumor prognosis. Although the model has achieved good prediction performance estimation in both training set and test set, it still needs to be further verified in independent data sets of other institutions to determine its prediction performance. In addition, compared with the benign and malignant classification of intestinal tumors, the four-classification of intestinal tumors (extremely low risk, low risk, medium risk and high risk) is more accurate and has more clinical application value.

Conclusions

Since the texture information used in the existing large-intestine pathological image classification algorithm can not describe the detailed features of the sample, the extracted features can not reflect the essential characteristics of the data well. Therefore, in order to improve the diagnostic accuracy of colon cancer, a novel classification algorithm based on color histogram and improved SVM framework is proposed, which has good approximation ability for complex pathological image samples. Our proposed algorithm combines wavelet kernel SVM with color histogram to classify pathological image. Firstly, the pathological image is divided into sub-patch, and the features of sub-patch histogram are extracted. Then, the global and local features are fused by the sub-patch weighting algorithm Then, the RelicfF based forward selection algorithm is used to integrate color features and texture features to enhance the characterization capabilities of the tumor. Finally, Morlet wavelet kernel-based least squares support vector machine method is used to enhance the generalization ability of the model for small sample with non-linear and high-dimensional pattern classification problems. Experimental results show that the proposed pathological diagnostic algorithm can gain higher accuracy compared with existing comparison algorithms. In future work, we will optimize the algorithm and embed the module into the medicine equipment to feedback the diagnosis results in real time and accurately so as to improve the automation level of pathological examination.

Acknowledgments This study is supported by the National Natural Science Foundation of China (No.81571773, 81781771943 81771943), Shanghai municipal health and Family Planning Commission (No.201640191).

Compliance with ethical standards

Conflict of interest We declare that we have no conflict of interest.

Human or animals participants This article does not contain any studies with human participants or animals performed by any of the authors.

Informed consent Informed consent was obtained from all individual participants included in the study.

References

1. Hong, Y., Wei, H, and Zeng-Li, L., Research for the colon cancer based on the EMD and LS-SVM[C]. Fourth International Conference on Intelligent Computation Technology & Automation. IEEE Computer Society, 24(7):329-331, 2011.
2. Wang, H., and Huang, G., Application of support vector machine in cancer diagnosis[J]. *Med. Oncol.* 28(1 Supplement):613–618, 2011.
3. Chen, H., Tan, C., Wu, H. et al., Feasibility of rapid diagnosis of colorectal cancer by near-infrared spectroscopy and support vector machine[J]. *Anal. Lett.* 47(15):2580–2593, 2014.
4. Mizaku, A, and Land, W. H., Biomolecular feature selection of colorectal cancer microarray data using GA-SVM hybrid and noise perturbation to address overfitting[J]. Dissertations & Theses - Gradworks, 12(1):64-76, 2009.
5. Tamaki, T., Yoshimuta, J., Kawakami, M. et al., Computer-aided colorectal tumor classification in NBI endoscopy using local features[J]. *Med. Image Anal.* 17(1):123-129, 2013.
6. Li, S., Fevens, T., Krzy Ak, A. et al., Automatic clinical image segmentation using pathological modeling, PCA and SVM[J]. *Eng. Appl. Artif. Intell.* 19(4):403–410, 2006.
7. Shi, W., Dongkai, J., Ke, W., Application of modified wavelet features and multi-class sphere SVM to pathological vocal

detection[C]. Seventh International Conference on Natural Computation. IEEE, pp:1290-1298, 2011.
8. Majhi, B., Dash, R., and Nayak, D. R., Stationary wavelet transform and AdaBoost with SVM based pathological brain detection in MRI scanning[J]. *CNS Neurol. Disord. Drug Targets* 16(2):32-44, 2017.
9. Cataldo, S. D., Ficarra, E., and Macii, E., Automated discrimination of pathological regions in tissue images: Unsupervised clustering vs. supervised SVM classification[C]. International Joint Conference on Biomedical Engineering Systems and Technologies. Springer, Berlin, Heidelberg, pp:2100-2111, 2008.
10. Cataldo,Wang S, Huo J, et al. Bayesian Framework with Non-local and Low-rank Constraint for Image Reconstruction[C]// Journal of Physics Conference Series. pp:1-11, 2017.
11. Shunji, T., Junji, T., Atsuko, H. et al., Role of early phase helical CT images in the evaluation of wall invasion of colorectal cancer: Pathological correlation[J]. *Nihon Igaku Hōshasen Gakkai Zasshi Nippon Acta Radiologica* 60(3):87, 2000.
12. None, Colorectal cancer pathology reporting: A regional audit[J]. *J. Clin. Pathol.* 50(4):358–358, 1997.
13. Xia, Kai Jian, H. S. Yin, and J. Q. Wang. "A novel improved deep convolutional neural network model for medical image fusion." Cluster Computing, 23(20):1-13, 2018.
14. Song, B., Zhang, G., Wang, H., et al., A feasibility study of high order texture features with application to pathological diagnosis of colon lesions for CT Colonography[C]. Nuclear Science Symposium & Medical Imaging Conference. IEEE, 2013.
15. Robnik-ŠIkonja, M., and Kononenko, I., Theoretical and empirical analysis of ReliefF and RReliefF[J]. *Mach. Learn.* 53(1–2):23–69, 2003.
16. Beretta, L., Santaniello et al., Implementing ReliefF filters to extract meaningful features from genetic lifetime datasets[J]. *J. Biomed. Inform.* 44(2):361–369, 2011.
17. Wang, C., Guan, Y., Zuo, C., et al., Value of the texture feature for solitary pulmonary nodules and mass lesions based on PET/CT[C]. International Conference on Bioinformatics & Biomedical Engineering. 2010.
18. Song, B., Zhang, G., Zhu, H., et al., A feasibility study of high order volumetric texture features for computer aided diagnosis of polyps via CT colonography[C]. Nuclear Science Symposium and Medical Imaging Conference (NSS/MIC). pp:719-724, 2012.
19. Song, B., Zhang, G., Lu, H. et al., Volumetric texture features from higher-order images for diagnosis of colon lesions via CT colonography[J]. *Int. J. Comput. Assist. Radiol. Surg.* 9(6):1021–1031, 2014.
20. Thon, N., Haas, C. A., Rauch, U. et al., The chondroitin sulphate proteoglycan brevican is upregulated by astrocytes after entorhinal cortex lesions in adult rats.[J]. *Eur. J. Neurosci.* 12(7):2547–2558, 2010.
21. Jiang, X., Liang, Q., and Shen, T., A new color information entropy retrieval method for pathological cell image[C]// Computer & Computing Technologies in Agriculture Iv-ifip Tc 12 Conference. 0.
22. Jiang, X., Liang, Q., and Shen, T., A new color information entropy retrieval method for pathological cell image[J]. Computer and Computing Technologies in Agriculture IV, 22(12):872-880, 2016.
23. Xia K J, Yin H S, Zhang Y D. Deep Semantic Segmentation of Kidney and Space-Occupying Lesion Area Based on SCNN and ResNet Models Combined with SIFT-Flow Algorithm[J]. Journal of Medical Systems, 2019, 43(1):2.
24. Sammouda, M, and Mukai, K., Diagnosis of liver cancer based on the analysis of pathological liver color images[C]. Medical Imaging: Image Processing. International Society for Optics and Photonics, pp:12-21, 2000.
25. Malekian, V., Mokhtari, M., Sadri, S., et al., Detection of collagenous colitis based on histopathology image segmentation of colon[C]// Iranian Conference on Machine Vision & Image Processing. IEEE, 2011.
26. Sammouda, M., Sammouda, R., Niki, N. et al., Cancerous nuclei detection on digitized pathological lung color images[J]. *J. Biomed. Inform.* 35(2):92–98, 2002.
27. Zheng, L., Wetzel, A. W., Gilbertson, J. et al., Design and analysis of a content-based pathology image retrieval system[J]. *IEEE Trans. Inf. Technol. Biomed.* 7(4):249–255, 2004.
28. Kande, G. B., Subbaiah, P. V., and Savithri, T. S., Unsupervised fuzzy based vessel segmentation in pathological digital fundus images[J]. *J. Med. Syst.* 34(5):849–858, 2010.
29. Ramella, G., Baja, G. S. D. Color histogram-based image segmentation[M]. Computer Analysis of Images and Patterns. Springer Berlin Heidelberg, 2011.
30. Jin, Y., Fayad, L., and Laine, A. F., Contrast enhancement by multiscale adaptive histogram equalization[J]. *Proc. SPIE Int. Soc. Opt. Eng.* 4478:206–213, 2001.