Frontiers in Signal Processing, Vol. 3, No. 4, October 2019
https://dx.doi.org/10.22606/fsp.2019.34006

Research on Tibetan Document Segmentation Based on GMM+K-means Algorithm

Chengliang Jiang, Huazhang Wang*

College of Electrical & Information Engineering, Southwest Minzu University, Chengdu, China
Email: 2420119863@qq.com

Abstract. The Tibetan language, as the language of the tubo period, recorded the life, history and other important events of the Tibetan people, and is a treasure of Tibetan culture. Aiming at the problem of the loss of Tibetan documents caused by the yellowing, blackening and rotting of papers due to the old age, a new method for the segmentation of Tibetan documents is proposed. In order to better protect Tibetan documents and reveal the contents recorded in the documents. The method uses the improved NLM (non-local means) algorithm to de-dry the pre-processing, and uses the automatic region-blocking GMM (Gaussian Mixture Model)+K-means multi-feature fusion algorithm to segment, using multi-region classification extraction as post-processing of the Tibetan documents. Experimental results show that compared with k-means, GMM and other algorithms, this method can more effectively segment the text in Tibetan literature, proving the effectiveness and accuracy of this method.

Keywords: Tibetan documents, NLM, GMM, K-means

1 Introduction

In the long process of historical development, a large number of Tibetan historical documents have been formed in Tibet and other Tibetan areas. They record a lot of historical information of Tibetan culture, which is of great reference value for studying Tibetan culture. With the increase of the number of years, a large number of documents appear yellow, rot, mildew and other phenomena, resulting in some words difficult to identify.

With the development of science and technology, researchers in Tibetan literature urgently need a technology that can automatically extract Tibetan document texts. And the segmentation extraction of the literature is the first step in this technology. Currently, the main segmentation methods can be roughly divided into four categories: threshold segmentation, edge detection segmentation, region-based segmentation and clustering segmentation. Among them, the threshold segmentation method obtains the results by comparing the determined threshold value with the selected pixel value. This method is simple in calculation and fast in speed, but only applicable to the relatively single background image. The edge detection segmentation method performs segmentation by detecting the edge of the image to form a contour, such as Canny operator[1], Robert operator detection[2], LOG operator detection[3], etc. The method is fast and efficient, but the method is not applicable to images with high noise. Based on the regional segmentation method, it is mainly divided into the regional growth method and the regional splitting and merging method. The basic idea is to start from a group of growth points and merge the adjacent pixels or regions with similar properties of the growth points with the growth points to form new growth points. The process is repeated until it cannot grow. The method of region splitting and merging first divides the image into any disjoint regions, and then merges or splits them to meet the restriction conditions. The feature regions are separated by splitting and the same feature regions are combined by merging. Due to multiple splits and merges, this method usually causes disadvantages, such as the boundary of the segmentation area being destroyed. The clustering segmentation method classifies the patterns by the similarity criterion, make the similar patterns are divided into one class as much as possible, and the dissimilar patterns are divided into different classes as much as possible. Compared with other methods, this method is more complicated to calculate, but its accuracy is relatively high. With the development of computers, it has also been widely used in engineering.

As one of the treasures of Chinese ethnic culture, ethnic literature segmentation research is relatively

rare, and there are few research results, some researchers have proposed using adaptive binary potential function target fuzzy C-means algorithm to segmentation uygur text [4], the algorithm can meet the initial requirements, but for the background is too complicated, the image segmentation effect of image segmentation is not good. In addition, some researchers proposed to use wavelet difference value and k-means method to extract color image text [5]. This method can handle the extraction of most color text, but a large amount of manual information extraction is required, so it cannot be fully automated.

This study is mainly aimed at the segmentation of Tibetan documents. We know that Tibetan literature is generally distributed in a determinant. Due to the age, the Tibetan literature appears to be partially blackened and blurred, if the document image is not captured by professional equipment (such as manual manual shooting), the image may be blurred, tilted, etc. due to jitter and other reasons, and the expected effect is not achieved. Aiming at previous algorithms, this paper firstly cuts the pre-selected image by perspective transformation. In the early stage of segmentation, we performed improved non-local mean filtering (NLM) on the image to eliminate most of the noise of the image, and then increase the text edge information by image enhancement. In the segmentation stage, K-means algorithm [6] and gaussian mixture model (GMM) algorithm [7] are used. According to the characteristics of block-blackening in Tibetan literature, the image is divided into n blocks of equal size. Then, the mean value of each block is calculated. First, we set a demarcation threshold (t). When the average value of this block image is less than the demarcation threshold (t), K-means algorithm will be used; when it is greater, GMM+ K-means algorithm will be used. And finally image processing integration is performed. The flow chart is shown in Figure 1.

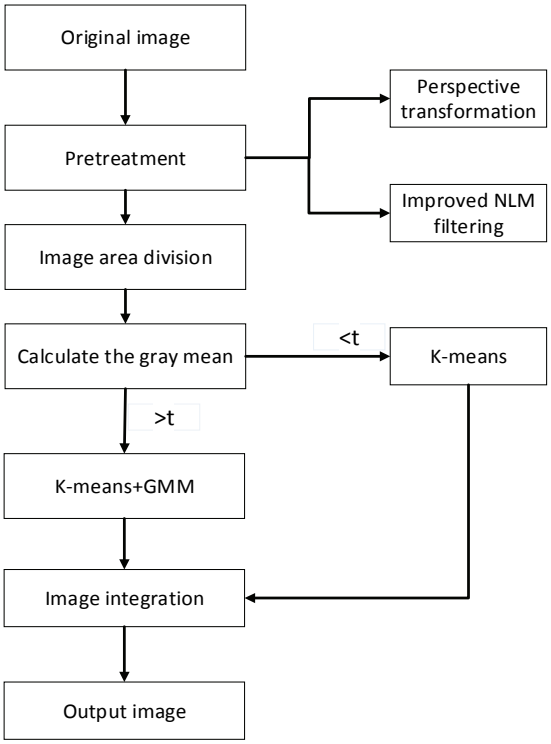

Figure 1. Overall flow chart of the research plan

2 Image Preprocessing – Image Correction and Image Enhancement

2.1 Image Correction

Aiming at the problem of photography Angle in some literatures, the phenomenon of image tilt is caused. In the application of modern technology, perspective transformation is the most widely used. The basic formula is as follows:

$$\begin{bmatrix} X' & Y' & W' \end{bmatrix} = \begin{bmatrix} U & V & W \end{bmatrix} * \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{21} & a_{32} & a_{33} \end{bmatrix} \tag{1}$$

First, we used the mouse to find out the edges and corners of the document, and recorded 4 coordinates. And then, according to the coordinate perspective transformation, get the image to be processed.

2.2 Image Enhancement

The Tibetan literature has caused a large number of small points due to its age, and such small points can be regarded as noise. Mean filtering, Gaussian filtering, median filtering, Gaussian filtering, etc., as classical filtering algorithms, have an important impact on image segmentation. However, they all have some shortcomings. For example, mean filtering adopts the method of taking the neighborhood of pixel points as the mean to get the mean value of each pixel point. Although this method can remove noise, it will also lose a lot of image texture information and image edge information. Median filtering algorithm takes the median value of selected pixel neighborhood class as the target value. This method can achieve a good effect when targeting isolated noise points, but it is obviously insufficient when dealing with regional noise. In recent years, many new dewatering methods have been proposed, such as using wavelet difference value in wavelet theory to denoising [8][9] and using partial differential equation to denoising, etc. [10][11]. In general, this method is better than the classical method, but it still eliminates a lot of texture and edge information. In 2005, Buades[12] et al. proposed a non-local Means (NLM) algorithm, which used non-local self-similarity to remove noise for the first time. The basic idea is: for each pixel in the image, take the image block of fixed size with the point as the center, search for the image block similar to it in the whole image, and carry out weighted average processing of weights between the two image blocks, the details are as follows:

A noisy image can be expressed as:

$$v(i) = u(i) + n(i) \tag{2}$$

where, $v(i)$ is the observed image, $u(i)$ is the real image, and $n(i)$ is the noise. For any pixel i in the image, NLM can obtain its estimated value according to the weighted average of all pixels:

$$NLM(i) = \sum_{j \in i} w(i,j)v(j) \tag{3}$$

where, the similarity of pixel value i and j depends on the weight $w(i,j)$, and $0 \ll w(i,j) \le 1$, $\sum_{j \in i} w(i,j) = 1$.

Using N_i to represent an image block centered on pixel i, the gray value vector of pixel point i and pixel point j is determined by the similarity between gray value vector $v(N_i)$ and $v(N_j)$. The similarity can be calculated by gauss weighted Euclidean distance, denoted as d:

$$d = \left\| v(N_i) - v(N_j) \right\|_{2,a}^2 \tag{4}$$

where, a is the standard deviation of gaussian kernel function, $a>0$.

According to the gaussian weighted distance, the weights of pixels i and pixels j are defined as:

$$w(i,j) = \frac{1}{z_i} \exp(-d / h^2) \tag{5}$$

where, $Z_i = \Sigma_j - \left\| v(N_i) - v(N_j) \right\|_{2,a/h^2}^2$ is the normalization constant, and the attenuation rate of the indirect control value of parameter h.

Compared with other classical algorithms, this algorithm achieves good results, but it has the disadvantages of large computation, too complex, and consuming too much time. Later, people introduced integral graph for calculation and proposed fast non-mean algorithm (FNLM), which greatly increased the speed but also reduced the dryness effect. In addition, the desiccating effect of NLM mainly depends on the extraction of image similar blocks and the weight distribution strategy. Whether it is NLM or FNLM, Euclidean distance is used in the calculation of the inner distance of the image, while Euclidean distance cannot achieve a good effect in the extraction of similar blocks of the image.

To solve such problems, we proposed a method in literature [13], added a new correlation coefficient index to measure the similarity between adjacent image blocks, so as to reduce the shortcomings caused by Euclidean distance.

$$W(i,j)=\frac{\sum((i-\overline{N_i})(j-\overline{N_j}))}{\sqrt{\sum((i-\overline{N_i})(i-\overline{N_i}))}\sqrt{\sum((j-\overline{N_j})(j-\overline{N_j}))}} \tag{6}$$

$$w(i,j)=\frac{1-W(i,j)}{2} \tag{7}$$

$$d=\frac{w(i,j)}{Num_j}\left\|v(N_i)-v(N_j)\right\|_{2,a}^2 \tag{8}$$

where, $\overline{N_i}$ and $\overline{N_j}$ represent the grayscale mean of the gray matrix N_i, N_j, respectively. Num_j indicates the number of j. The value of $w(i,j)$ is between [–1, 1]. When we bring d into (5), we get the final formula.

3 Proposed GMM+K-means Multi-feature Fusion Algorithm for Automatic Region Partitioning

The darkening part of Tibetan literature images is usually uneven. If the whole image is used for one-time segmentation, the probability of mis-segmentation is very large. According to the characteristics of determinant distribution in literature, the image is divided into several small blocks according to the specific number of rows, and then the average value of each block is analyzed to determine the specific operation method. If a block image is smaller than the threshold we set, a separate K-means algorithm is used to reduce the time running. If a block image is larger than the threshold, the GMM+K-means multi-feature fusion algorithm is used.

3.1 Introduction to the Algorithm

According to the central limit theorem in probability theory, a large number of random variables are approximately subject to gaussian distribution. If we extract multiple sets of feature vectors independent of each other in an image, each set of feature vectors will also obey a Gaussian distribution:

$$P(y\mid\theta)=\sum_{k=1}^{K}\alpha_k N(v\mid\mu_k,\Sigma) \tag{9}$$

where α_k is the kth Gaussian kernel weight, $\sum_{k=1}^{K}\alpha_k=1,(0\le\alpha_k\le 1)$.

Gaussian Mixture Model (GMM) is a commonly used method of data clustering. In the image segmentation method based on the gaussian mixture model, the researcher establishes the gaussian distribution of the image according to the distribution of each group of feature vectors. For multidimensional feature vectors, there are the following forms:

$$N(v\mid\mu,\Sigma)=\frac{1}{\sqrt{2\pi|\Sigma|}}\bullet e^{-\frac{1}{2}(v-\mu)^T\Sigma^{-1}(v-\mu)} \tag{10}$$

where μ represents the mean value and Σ represents the covariance matrix. According to equations (9) and (10), the main parameters needed to solve the gaussian mixture model are gaussian kernel weight α, mean μ and covariance matrix Σ. In order to solve such problems, the current classical method is to introduce the EM algorithm (namely the expectation-maximization algorithm). The specific methods are as follows:

Set the image point $x=\{x_1, x_2, \ldots, x_n\}$, and in the application of GMM algorithm, first rewrite (9) as:

$$p\left(x\mid\alpha,\mu,\Sigma\right)=\sum_{k=1}^{K}\alpha_k N\left(v\mid\mu,\Sigma\right) \tag{11}$$

In order to better use the EM algorithm, we introduce the implicit variable Z and the posterior

probability $\gamma(Z_{nk})$ after known x, which can be sorted out by the bayesian formula:

$$\gamma\left(Z_{nk}\right)=\frac{\pi_k N\left(x_n \mid \mu_n, \Sigma_n\right)}{\sum_{j=1}^{K} \pi_j N\left(x_j \mid \mu_j, \Sigma_j\right)} \tag{12}$$

The maximum likelihood function of α, μ and Σ can be calculated as follows:

$$\left.\begin{aligned} \mu_k &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma\left(Z_{nk}\right)x_n \\ \Sigma_k &= \frac{1}{N_k}\sum_{n=1}^{N}\gamma\left(Z_{nk}\right)\left(x_n-\mu_k\right)\left(x_n-\mu_k\right)^T \\ \alpha_k &= \frac{N_k}{N} \end{aligned}\right\} \tag{13}$$

where:

$$N_k=\sum_{n=1}^{N}\gamma\left(Z_{nk}\right) \tag{14}$$

where N represents the number of points, $\gamma(Z_{nk})$ represents the posterior probability of point x_n belonging to cluster k, and N_k represents the number of points belonging to the kth cluster point.

In the EM algorithm, if we set different initial conditions, we will get different local maxima, so the initial value setting directly affects the test results. In order to get the best segmentation image in practical applications, this study uses the K-means algorithm to calculate the centroid vector μ_k, and then uses it as the mean initial value in the EM algorithm, and defines it as μ_k.

The specific method is as follows:

(1) Convert the original color image from the RGB space to the HSV space and make it into a sample set D.

(2) Set the input sample set as $D=\{x_1, x_2, \ldots, x_m\}$, cluster tree of clustering as k, maximum iteration number as N, and output as cluster classification $C=\{c_1, c_2, \ldots, c_k\}$.

(3) By initialization, the initial k centroid vectors $\{\mu_1, \mu_2, \ldots, \mu_k\}$ and the initialization cluster partition $C_t=\phi$, (t=1, 2, …, k) are obtained. Then calculate the distance $d_{ij}=\left\|x_i-\mu_j\right\|_2^2$ between the centroid vector μ_j (j=1, 2, …, k) and the sample x_i of each sample. Mark x_i as the category λ_i corresponding to the smallest distance d_{ij}, last updated $C_{\lambda_i}=C_{\lambda_i}\bigcup\left\{x_i\right\}$.

(4) Formulate the Square Error $E=\sum_{i=1}^{K}\sum_{x\in c_i}\left\|x_i-\mu_i\right\|_2^2$.

(5) Recalculate the centroid vector $\mu_j=\frac{1}{\left|C_{\lambda_i}\right|}\sum_{x\in c_{\lambda_i}} x$ according to C_{λ_i}. When all the centroid vectors are unchanged or the values obtained are less than the Squared Error, the centroid vector is the initial mean. Recorded as μ_k.

(6) Bring μ_k into formula (13) and find α_k, Σ_k. Then use it as the initial value of the GMM algorithm.

(7) Calculate the posterior probability ($\gamma(Z_{nk})$), and then get the new μ_k, α_k, Σ_k, and repeat until it converges to get the final result.

4 Image Post Processing

K-means algorithm and GMM algorithm are both for color image processing. If only binary classification is used in clustering, there will be a lot of wrong segmentation of the obtained images due to yellowing, mildew unevenness and other reasons. Therefore, we need to improve the classification number and reprocess it later. In this study, we used four types of segmentation, and the processing methods are as follows:

(1) Extracting the pixel position (marked as T) of the segmented image and the corresponding original image pixel position (labeled as S). For simplicity, we convert it to a single channel image.

(2) Make the segmentation image into a sample set ($T=\{t_1, t_2, \ldots, t_m\}$), the original image ($S=\{s_1$,

s_2, …, s_m}). Since T is divided into 4 categories (U={U_1, U_2, U_3, U_4}) by the segmentation algorithm, Now the four categories in T are mapped to the original image S to get Q={Q_1, Q_2, Q_3, Q_4}. Then calculate the mean of each type of pixel block corresponding to the original image. According to the darkest color feature of the literature, the image with the largest mean value is identified as text, the smallest as background, and the remaining two images as undetermined images.

(3) Assume that Q_1 corresponds to the largest mean of pixel blocks and the smallest mean of Q_4, then Q_1 corresponds to text and Q_4 to background. Now all the external edges of Q_1 in the original image are extracted. Calculate the mean value of the neighborhood pixels (the pixels after removing part Q_1) of each edge pixel point is μ. If μ is less than the set threshold t, this part is considered as an edge, and its pixel coordinate is integrated into Q_1. When the edge encounters another type of pixel block, the command terminates.

(4) In order to prevent the case where the entire area is changed to Q_1. The algorithm was set to stop automatically after calculating the neighborhood means for 5 times, and the rest of the algorithm was set as the background.

5 Test Results and Analysis

5.1 Filtering Process Experimental Results and Analysis

In this experiment, the original image direct segmentation, median filtering segmentation, NLM filtering segmentation and improved NLM segmentation results were compared. In order to show the effectiveness of the filtering process, the original image is compared in this study, and then the segmentation details are compared. At the same time, Peak signal-to-noise ratio (PSNR) and Structural Similarity Index (SSIM) are introduced to compare them. The final test results are the optimal results of various methods, in which the median filtering window is set as 5, and the size of similar Windows of NLM and improved NLM is set as 7*7. The results are as follows:

Figure 2. Comparison of three filtering effects

As can be seen from figure 2, since the image noise points used in this experiment are similar to the original image background, there is no big difference in the image filtering observed subjectively. However, from the detailed comparison after segmentation, it is found that the obvious noise points are reduced after filtering. Meanwhile, compared with the filtered images, it is found that the improved

NLM filtering algorithm used in this study can more effectively segment the severely blackened images. As shown in Table 1, compare them with PSNR and SSIM, The peak signal-to-noise ratio of the filtering algorithm used in this paper is 40.257, and the structural similarity is 0.9646574, which are the highest in the comparison of this experiment. The data show that the filtering algorithm used in this paper has better filtering effect.

Table 1. Comparison of PSNR and SSIM for each image.

Image name	PSNR	SSIM
Median	37.8796	0.942212
NLM	39.4138	0.962077
The improve NLM	40.257	0.964674

5.2 Segmentation Process Test Results and Analysis

This test compares the three methods of K-means algorithm, GMM algorithm and K-means algorithm + GMM algorithm. Set the number of iterations of the GMM algorithm to 100 times, demarcation threshold t=128. The specific test image and time used are as follows:

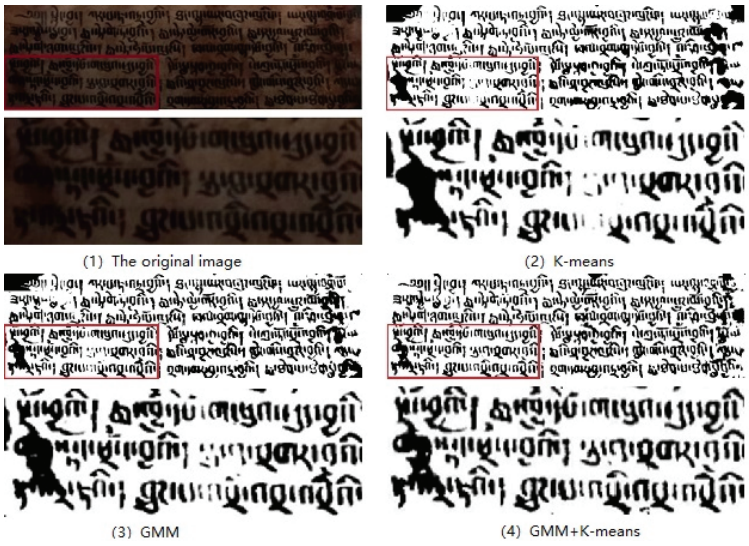

Figure 3. Comparison of various algorithms in Test 1

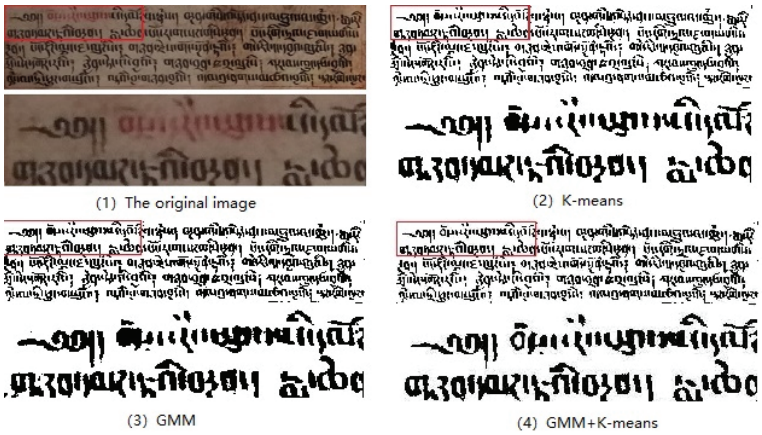

Figure 4. Comparison of various algorithms in Test 2

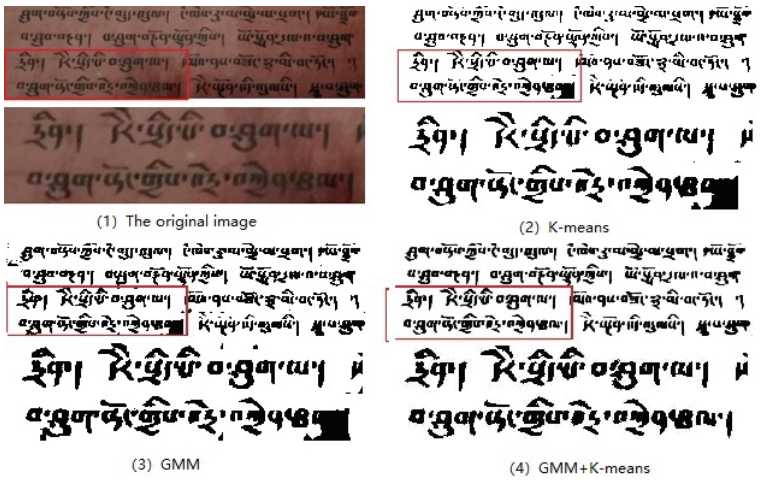

Figure 5. Comparison of various algorithms in Test3

Table 2. Time comparison of various algorithms

Image name	Using algorithm	Time
Test 1	GMM	27.662s
	K-means	14.854s
	K-means+GMM	22.102s
Test 2	GMM	24.6169s
	K-means	15.7221s
	K-means+GMM	19.864s
Test 3	GMM	21.1256s
	K-means	10.2563s
	K-means+GMM	15.1677s

According to the time comparison in table 2, GMM algorithm has the longest running time, and k-means algorithm has the lowest running time, K-means+GMM algorithm has the middle running time. Combined with the three experimental images in this paper, it can be clearly known from subjective observation that they are all uneven illumination, but they all have their own characteristics. In experiment 1, the black part of the image was more serious.In experiment 2, the yellow part of the image was more serious.The image in experiment 3 was the clearest of the three. By observing the segmentation results of three images, we can know that the algorithm used in this paper can achieve the best results. For example, the darkened part in the lower left corner of experiment 1, the red font part in experiment 2, and the uneven illumination part in the lower left corner of experiment 3. They are the best results of the three algorithms. Therefore, the algorithm in this paper has great advantages.

In summary, if we deal with Tibetan images that are not blackened and blurred, using K-means algorithm alone can also achieve good results, which can guarantee our efficiency. However, for Tibetan images with severe blackening and blurring, the method of this study not only guarantees the segmentation effect, but also tries to control the consumption of time, which has good practicability.

6 Conclusion

In Tibetan literature, characters are different from Chinese characters, it is not a font of the same size, and there is not much specific stroke order. Therefore, the segmentation of Tibetan text is more difficult than that of Chinese characters. Through the characteristics of Tibetan literature, this paper adopts more horizontal segmentation in automatic regional segmentation, so as to better segment the complete text. At the same time, aiming at the noise, blackening, blurring and other problems of Tibetan literature, this paper uses the algorithm of NLM after image correction. Using the improved weighting

scheme, make up for the deficiencies of this algorithm, by using image region segmentation and combining k-means algorithm and gaussian mixture model algorithm according to the characteristics of each image, the segmentation accuracy of the image is improved while time consumption is guaranteed. The experimental results show that the experiment can be used in practice.

Acknowledgements. The work of this paper is supported by the Southwest Minzu University Graduate Innovative Research Project (Master Program CX2018SZ91) and special fund project for basic scientific research operating expenses of central universities (2015NYB03). A special acknowledgement should give to Southwest Minzu University for its experimental conditions and technical support.

References

1. Song R, Zhang Z, Liu H. “Edge connection based Canny edge detection algorithm”. Pattern Recognition & Image Analysis, 2017, 27(4), pp.740-747.
2. GAO Yong-gang. “An Improved Edge Detection of Roberts Operators”. Journal of Chaohu University, 2009, 11(6), pp.31-32.
3. Ding K, Xiao L, Weng G. “Active contours driven by region-scalable fitting and optimized Laplacian of Gaussian energy for image segmentation”. Signal Processing, 2017, 134, pp.224-233.
4. YILIHAMU Yaermaimaiti, “Research on an improved image segmentation algorithm for Uyghur characters”, Modern Electronics Technique, pp.128-131, 2017(04).
5. Zhang Kaige, Miao Yi, Lei Jiankun, et al. “Extraction of color image texts combining wavelet interpolation and K-means”. Computer Technology and Development, pp.31-33. 2013(3).
6. Wu Suhui, Cheng Ying, Zheng Yanning, et al. “Survey on K-means Algorithm”, Data Analysis and Knowledge Discovery, pp.28-35,2011, 27(5)
7. Xiang Rihua, Wang Runsheng. “A Range Image Segmentation Algorithm Based on Gaussian Mixture Model”. Journal of Software, 14(7), pp.1250-1257. 2003.
8. Luisier F, Blu T. “SURE-LET multichannel image denoising: interscale orthonomal wavelet thresholding “.IEEE Transactions on Image Processing, 2008, 17(4). pp. 482-492.
9. Donoho D L, Johnstone J M. Ideal spatial adaptation by wavelet shrinkage[J]. Biometrika, 1994, 81(3), pp.425-455.
10. Bai J, Feng X C. “Fractional order anisotropic diffusion for image denoising”. IEEE Transactions on Image Pmcessing, 2007, 16(10), pp.2492-2502.
11. Perona P, Malik J. “scale-space and edge detection using all isotropic diffusion”. IEEE Transactions on Pattem Analysis and Machines Intelligence, 1990, 12(7), pp.629-639.
12. Buades A, Coll B, Morel J M. “A Non-Local Algorithm for Image Denoising”. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. San Diego, CA, USA：IEEE, 2005, pp60-65.
13. Zhao Qingping, Chen Debao, Jiang Enhua, et al. “Improved weighted non-local mean algorithm filter for image denoising”, Journal of Electronic Measurement and Instrument, 2014, 28(3), pp.334-339.

