

A segmentation of pulmonary nodules based on improved fuzzy C-means clustering algorithm

Yang Tiejun¹, Cheng Jinfeng^{2,a} and Zhu Chunhua³

¹HeNan University of Technology, School of Information Science and Engineering, 450001, China

²HeNan University of Technology, School of Information Science and Engineering, 450001, China

Abstract. According to reports, lung cancer is gradually becoming the first cancer that threatens human life. The early stage of lung cancer is in the form of pulmonary nodules. The key issue in computer-aided diagnosis of lung tumors is to correct and accelerate rapid segmentation of diseased tissue. Therefore, this paper proposes a robust fuzzy c-mean clustering algorithm for pulmonary nodules segmentation, which can effectively improve the adaptive degree of local domain pixels. Since the information of the domain pixels does not necessarily have a positive correlation with the central pixels, the reference mechanism of domain window pixel information needs to be redefined. The robust fuzzy c-means clustering algorithm redefines the grayscale of the spatial pixel points in the domain and selects different fuzzy factors according to the reference standard. Based on this, the weights of different fuzzy factors are updated according to the characteristics of pixel points and gray fluctuation in pixel domain. The experimental results show that this method is superior to other typical algorithms in the segmentation of pulmonary nodules.

1 Introduction

Lung cancer is one of the most malignant tumors with the fastest growth in morbidity and mortality and the greatest threat to population health and life. Over the past 50 years, many countries have reported significant increases in the morbidity and mortality of lung cancer. Released statistics by the National Cancer Center in 2015 shows that incidence of lung cancer is close to 17.09%, and men accounted for 70.3%, women accounted for 29.7%, and the mortality rate is as high as 21.68%, ranking first among various tumors [1]. Lung cancer will become the biggest threat to human health. If patients with lung cancer in early get standardized surgical treatment, the 5-year survival rate is as high as 90%. In the stage I treatment of lung cancer, the 5-year survival rate of patients is 60%, but in stage II-IV treatment of lung cancer, the 5-year survival rate reduces to 5% from 40% [2]. In order not to miss the best treatment period, you need to find it as early as possible and treat it as soon as possible. Pulmonary nodules are the most common form of early lung cancer. In order to improve the true positive rate of lung nodules, CT images are used for the auxiliary diagnosis of pulmonary nodules. At present, CT image data shows an explosive growth trend, which is bound to increase the workload of doctors, leading to missed diagnosis and misdiagnosis. Studies have shown that accurate and effective lung CT image segmentation can reduce the amount of calculation, improve the efficiency of the entire diagnostic system, reduce missed diagnosis and

misdiagnosis, and play an important role in the process of lung disease and function evaluation [3].

In lung CT images, pulmonary nodules are usually seen as tissues that is brighter than the surrounding lung parenchyma, and similar to a sphere or an ellipsoid. According to the shape and position distribution of lung nodules in the lung parenchyma, they can be divided into the following categories: picture (4) is Solitary Pulmonary Nodule (SPN), picture (2) is Lung Wall Attachment Pulmonary Nodule (LWA), picture (1) is Ground Glass Opacity (GGO), and picture (3) is Juxta-vascular Pulmonary Nodule (JV). As shown in Figure 1.

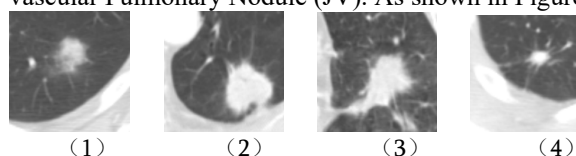


Figure 1. Examples of various types of pulmonary nodules

Because of many noise interferences and poor resolutions in CT images, the segmentation accuracy of many traditional algorithms is relatively low, and the mis-segmentation rate is relatively high. Medical image segmentation techniques range from early manual segmentation to semi-automatic segmentation, then to automatic segmentation. During this period, many scholars at home and abroad have proposed many methods for segmentation of pulmonary nodules. Such as threshold method, regional growth, fuzzy clustering, neural network and so on.

Fuzzy C-means Clustering (FCM) [4] is the most widely used image segmentation algorithm at

^a Corresponding author: 782640105@qq.com

present. The traditional fuzzy C-means clustering is very sensitive to noise, and does not consider the spatial positional relationship of pixels, and easily converges to local extremum. In order to solve these problems, many scholars at home and abroad have improved the traditional FCM algorithm by combining the degree of correlation between pixel points and neighboring pixels. Ahmed et al introduced a spatial constraint on pixel color features in the objective function of FCM algorithm, and proposed a Fuzzy C-means Clustering with Spatial information (FCM-S) [5]. The algorithm has a certain capability of noise immunity, but each iteration must perform a series of calculations on the spatial information of the domain. Chen and Zhang [6] performed mean filtering and median filtering on the image before each iteration, and proposed FCM-S1 and FCM-S2, which reduced the computational complexity of FCM-S, filtering will also smooth the details of the image while suppressing noise, resulting in mis-segmentation of subsequent FCM clustering. Szilagy [7] proposed an enhanced fuzzy C-means clustering (EnFCM) based on image gray level. The EnFCM algorithm computes a linearly weighted sum image of the original image and its local neighborhood mean image. First, the image was filtered by means of means, and then the image was segmented by fuzzy c-mean clustering on its gray histogram. The number of gray levels of the image is much smaller than that of the image pixels. This algorithm effectively reduces the complexity of the operation and improves the segmentation speed of the image. However, in the mean filtering process, the original texture details will be lost, resulting in image edge blurring. Cai [8] proposed a fast generalized FCM clustering algorithm (FGFCM), a local similarity measure method that combines local spatial distance and gray-scale difference. It weakens the blur of the edge of the image to a certain extent, but the parameters that this algorithm cannot be automatically obtained through experiments are introduced, and the segmentation precision is not ideal. Stelios [9] proposed a local information-based FCM algorithm (Fuzzy local information C-means, FLICM), which combines local spatial information and gray-scale information to define a new type of fuzzy factor, similarity measurement without any parameters. The factor has better adaptability, and the original image is used in the iterative process, which avoids the loss of detail caused by the preprocessing and improves the segmentation effect of the image. However, due to the limitations of its constructed local spatial information, some details are lost during image segmentation.

In order to further improve the accuracy and robustness of FCM image segmentation algorithm, this paper proposes a Robust Fuzzy C-mean Clustering algorithm to improve the local area pixel adaptive degree. The robust fuzzy c-means clustering algorithm re-specifies the grayscale values of the pixels in the domain space by grayscale, and selects different fuzzy factors by reference standard. Based on this, the weights of different fuzzy factors are adaptively updated according to the characteristics of pixel points and the gray level fluctuations in the pixel domain.

2 Traditional fuzzy c-mean clustering algorithm

Fuzzy clustering refers to the process of dividing a set of data sets into categories according to specified features without training samples. The fuzzy c-means clustering algorithm is a combination of fuzzy theory and cluster analysis. The FCM algorithm is a process of repeatedly modifying the cluster center and the membership matrix, so it is also called dynamic clustering. It is an unsupervised machine self-learning algorithm.

The main idea of FCM is to set the objective function. After iterative optimization, the membership degree of the target point to its category is obtained, and then the membership degree is used to classify the target sample. We use all the pixel gray values in the image as the target data, and then continuously iteratively optimize the objective function to obtain the classification result of the pixel gray value and complete the accurate segmentation of the image. Define the objective function J_{FCM} :

$$J_{FCM} = \sum_{i=1}^N \sum_{j=1}^c u_{ij}^m d^2(x_i, v_j) \quad (1)$$

Where, N is the number of pixels in the image; c is the preset number of clusters, satisfying $2 < c \leq N$; u_{ij} is the fuzzy membership of the i th pixel x_i relative to the j th cluster center v_j ; m is the weight index acting on the fuzzy membership degree, which controls the degree of blurring of the data partition; v_j is the j th cluster center; d_{ij} is the Euclidean distance from the i th pixel to the j th cluster center; the constraint condition of the objective function minimization is:

$$u_{ij} \in [0,1] \cup \sum_{j=1}^c u_{ij} = 1 \quad (2)$$

According to the Lagrange multiplier method, it is necessary to obtain a minimum value for the objective function under the constraint condition:

$$v_j = \frac{\sum_{i=1}^N (u_{ij})^m x_i}{\sum_{i=1}^N (u_{ij})^m} \quad (3)$$

$$u_{ij} = \frac{1}{\sum_{k=1}^c \left(\frac{d_{ij}}{d_{ki}}\right)^{\frac{2}{m-1}}} \quad (4)$$

The specific steps of the FCM algorithm are as follows:

1. Input image X as the target data, set the value of m and c , set the iteration stop threshold $\varepsilon > 0$;

2. Random initialization of fuzzy clustering matrix $U^{(0)}$;

3. Set the loop counter $t=0$;

4. Calculate fuzzy membership matrix and cluster center according to formulas (3) and (4);

5. If $\max(U^{(t)} - U^{(t+1)}) < \epsilon$, the algorithm terminates, otherwise $t=t+1$ is executed, and step (4) is repeated to continue the operation;

3 Adaptive and robust fuzzy c-mean clustering algorithm

In order to improve the accuracy and robustness of the FCM clustering algorithm, the algorithm selects the applicable fuzzy factor by introducing a weighting factor that can automatically determine whether the central pixel is a noise point or an edge point in the objective function of the FCM algorithm. The weighting factor is defined according to the characteristics of the pixel points and the gray level fluctuations in the pixel domain:

$$a_k = \begin{cases} \frac{\sigma_i}{\max(\sigma)}, & |x_k - \bar{m}| \leq \sigma_i \\ 0, & |x_k - \bar{m}| > \sigma_i \end{cases} \quad (5)$$

Where, x_k is a certain domain pixel of the central pixel of the image; The algorithm processes the eight-domain pixel, and all the domain pixels constitute a domain system set $L(x_i)$, that is $x_k \in L(x_i)$; $|x_k - \bar{m}|$ indicates the absolute distance between the domain pixel x_k and the mean square value \bar{m} of the domain. \bar{m} is the mean square value of the domain pixel system $L(x_i)$; σ_i is the variance of the domain pixel system $L(x_i)$; If the distance is greater than the variance σ_i of the domain pixel, it indicates that x_k is the noise point or the edge point of the domain pixel; The value of weight factor is zero at this time; Otherwise, x_k is the pixel inside the region, the weighting factor can be obtained according to the grayscale fluctuations in the region; A weight matrix $a = \{x_k | k \in N\}$, $\sigma = \{\sigma_i | i \in N\}$ is obtained by traversing each pixel in the image, where N is the sum of the image pixel; Since the change of the weighting factor m of the membership degree has little effect on the result of the segmentation image of the algorithm, the value of the algorithm m is selected as two.

The objective function of the algorithm in this paper is defined as:

$$J_{ARFCM} = \sum_{i=1}^N \sum_{j=1}^c [u_{ij}^2 d^2(x_i, v_j) + a P_{ij} + \bar{a} Q_{ij} + G] \quad (6)$$

Where x_i is the gray value of the pixel in the center of the image; u_{ij} is the membership matrix belonging to the j th class of the center pixel x_i ; v_j is the cluster center of the j th class; $d^2(x_i, v_j)$ is the Euclidean distance between the pixel and the cluster center; The blur factor P_{ij} is defined as:

$$P_{ij} = u_{ij}^2 \frac{b}{N_R} \sum_{r \in N_i} \|x_r - v_j\|^2 \quad (7)$$

Where b is the control parameter of the intensity influence of the domain information; N_R is the number of pixels of the domain set; x_r is the gray value of the domain pixel; The blur factor Q_{ij} is defined as:

$$Q_{ij} = u_{ij}^2 \sum_{r \in N_i} (1 - u_{ij})^2 \|x_r - v_j\|^2 \quad (8)$$

Each pixel should obey the constraint formula (2).

The constraint Q_{ij} not only takes the neighboring pixels of the central pixel as reference information, but also limits the fuzzy membership value. The constrained pixel completely does or does not belong to the j th class of the clustering center, which avoids ambiguity and makes the central pixel's attribution more clear. There is also a fixed blur factor G_{ij} in the objective function, defined as:

$$G_{ij} = \sum_{\substack{x_r \in N_i \\ i \neq r}} \frac{1}{d_{ir}} (1 - u_{jr})^2 \|x_r - v_j\|^2 \quad (9)$$

Where d_{ir} is the distance from the pixel point to the center of the first cluster; the blur factor G_{ij} includes spatial information and gray level information, and the influence on the objective function is controlled through the distance from the neighborhood pixel to the central pixel, the robustness of the segmentation and the noise sensitivity are improved, and there is no choice of parameters.

Similar to the traditional FCM algorithm, the Lagrange multiplier method can be used to obtain the membership matrix U^n and cluster center v_j^n that make the objective function J_{ARFCM} obtain the local minimum under the constraint condition (2):

$$u_{ij} = \frac{\|x_i - v_j\|^2 + \frac{b}{N_R} a \sum_{x_r \in N_i} \|x_r - v_j\|^2 + G_{ij}}{\sum_{j=1}^c \left(\|x_i - v_j\|^2 + \frac{b}{N_R} a \sum_{x_r \in N_i} \|x_r - v_j\|^2 + G_{ij} \right)} \quad (10)$$

$$v_j = \frac{\sum_{i=1}^N u_{ij}^2 \left[x_i + a \frac{b}{N_R} \sum_{x_r \in N_i} x_r + \bar{a} \sum_{x_r \in N_i} (1 - u_{ij})^2 x_r \right]}{\sum_{i=1}^N u_{ij}^2 \left[1 + ab + \bar{a} \sum_{x_r \in N_i} (1 - u_{ij})^2 \right]} \quad (11)$$

The flow of the ARFCM algorithm:

- (1) Input image data, determine the number of clusters c , set iterative stop threshold ε , and let $t=0$;
- (2) Initialize the fuzzy clustering matrix by a random method;
- (3) Calculate the weighting factor from equation (5);
- (4) Calculate the cluster center v_j^n and the membership matrix U^n from equations (10) and (11);
- (5) Calculate the fuzzy membership matrix U^{n+1} by the formula (10), if $\{U^{n+1} - U^n\} < \varepsilon$ the algorithm stops iterating, otherwise let $t=t+1$ and return to step (3) to continue the operation.

4 Experimental results and discussion

The experimental environment of this paper is as follows: The operating system is Windows 10; The processor: Intel(R) Core(TM) i7-6700CPU 3.4GHZ 3.41GHZ; The actual experiment of the segmentation effect is performed by Matlab R2008a. In the experiment, setting the parameter fuzzy weighted index $m=2$; Segmentation window $N_R=9$; Number of clusters $c=3$; Target function iteration stop threshold $\varepsilon=0.00001$; Controlling parameter $b=0.7$ of field information influence intensity; The target picture pulmonary nodule CT images is mainly from the database of the Lung Image Database Consortium (LIDC). We selected 40 low-dose lung CT images from the LIDC database. In order to verify the validity and robustness of the proposed algorithm, the red curve is used to mark the pulmonary nodule boundary according to the annotation file provided by the database. We compared the segmentation results with the FCM algorithm, FCM-S algorithm, FLICM algorithm and rFCM algorithm.

Figure 2 shows the segmentation effect of Solitary Pulmonary Nodule. Figure 2-1 shows an original CT image of a pulmonary nodule. Figure 2-2 shows the pulmonary nodules boundary that marked by red curve based on the annotation file provided by the database. Figure 2-3 shows the extracted lung parenchyma. When comparing with various traditional algorithms, we find that the edges of the segmentation results with Figure 2-5 (FCM algorithm) and Figure 2-6 (FCM-S algorithm) have obvious sawtooth edges, there is no effective segmentation of some fine textures. Figure 2-7 and Figure 2-8 show that the edge information of the pulmonary nodules is lost when we used the FLICM algorithm and rFCM algorithm to segment pulmonary nodules, and the over-segmentation is more serious. The algorithm of figure 2-9 (ARFCM algorithm) shows that the effect of noise can be well eliminated, and the segmentation edge is smoother, and the shape of the

pulmonary nodules obtained by segmentation is close to the expert's artificial segmentation results.

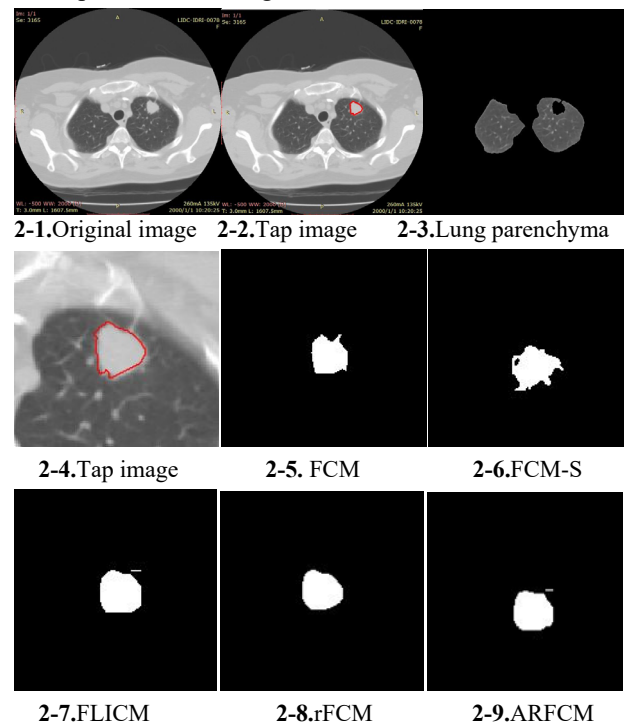
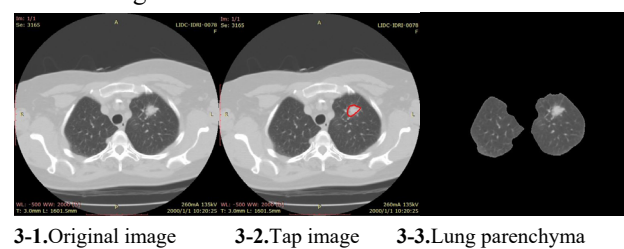


Figure.2 Segmentation effect of Solitary Pulmonary Nodule

Figure 3 shows the segmentation effect of Ground Glass Opacity (GGO) Pulmonary Nodule. Figure 3-1 shows an original CT image of a pulmonary nodule. Figure 3-2 shows the lung nodule boundary that marked by red curve based on the annotation file provided by the database. Figure 3-3 shows an extracted lung parenchyma. When comparing with various traditional algorithms, we find that figure 3-5 (FCM algorithm) and 3-6 (FCM-S algorithm) do not effectively segment some small textures, and the segmentation effect is not very good. Figure 3-7 (FLICM algorithm) and figure 3-8 (rFCM algorithm) have lost the edge information of lung nodules. The over-segmentation is more serious, but the segmentation result of rFCM algorithm has been significantly improved compared with other algorithms. The algorithm of figure 3-9 (ARFCM algorithm) shows that the noise can be effectively eliminated, the edge information of pulmonary nodules is preserved, and the segmentation edge is smoother. The shape of the segmented pulmonary nodule is close to the expert's artificial segmentation result.



3-1.Original image **3-2.**Tap image **3-3.**Lung parenchyma

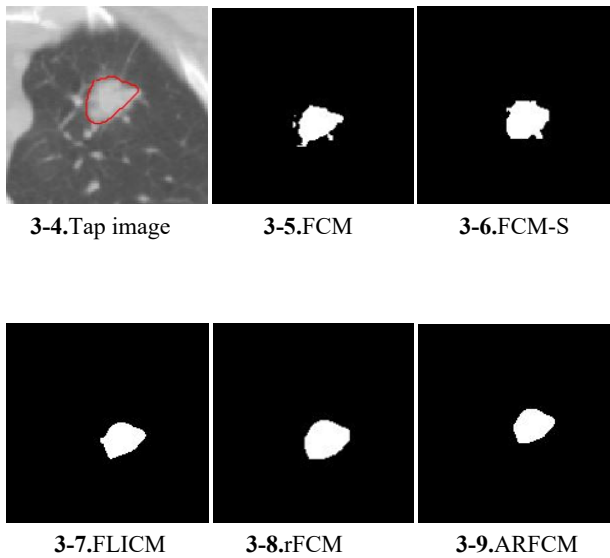


Figure.3. Segmentation effect of Ground Glass Opacity (GGO) Pulmonary Nodule

All of the above are qualitative analysis and compare the algorithm in this paper with the traditional algorithms. The following is the quantitative analysis of the segmentation effect of each algorithm. We compare the algorithms using pixel-based precision, recall and F1 values. The precision (P) and recall rate (R) is defined as follows:

$$\text{Precision (P)} = \frac{\text{Area} \cap \text{Area}_{real}}{\text{Area}_{real}} \times 100\% \quad (12)$$

$$\text{Recall (R)} = \frac{\text{Area} \cap \text{Area}_{real}}{\text{Area}} \times 100\% \quad (13)$$

$$\text{F-Measure} = \frac{(\alpha^2 + 1) \times P \times R}{\alpha^2 \times (P + R)} \times 100\% \quad (14)$$

In equation (12), Area_{real} is pulmonary nodules area that we use the red curve of the pulmonary nodule boundary to mark the lung nodule area according to the annotation file provided by the database. As the gold standard, Area is the pulmonary nodule area obtained by the segmentation algorithm. When the parameter $\alpha^2 = 1$, it is the most common evaluation index F1 value, which combines the results of precision and recall. When F1 is high, it proves that the segmentation effect of the algorithm is better.

$$\text{F1} = \frac{2 \times P \times R}{P + R} \times 100\% \quad (15)$$

Figure 4 is a comparison of the precision of each algorithm; Figure 5 is a comparison of the recall rate of each algorithm; Figure 6 is a comparison of the F1 values of each algorithm. It can be seen from the following figure that the precision of ARFCM algorithm is higher than other algorithms, and the F1 value is higher than 85%, and the effectiveness of pulmonary nodule segmentation is improved.

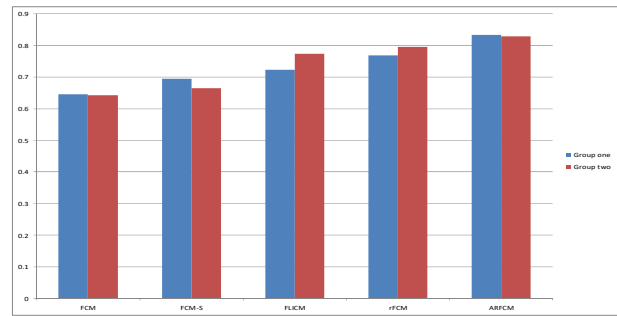


Figure.4. Comparison diagram of precision of each algorithm

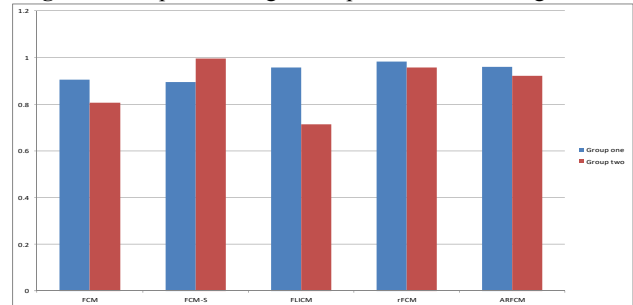


Figure.5. Comparison diagram of recall rate of each algorithm

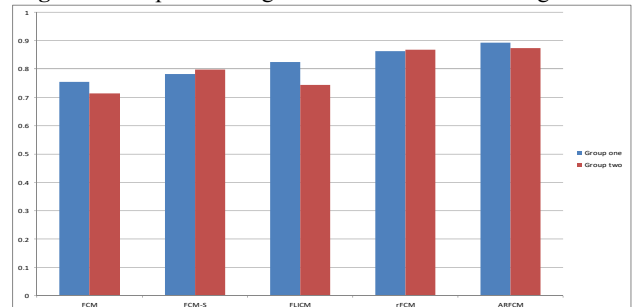


Figure.6. Comparison diagram of F1 values of each algorithm

5 Conclusions

This paper proposes an adaptive robust fuzzy C-means clustering method for lung nodule segmentation, which introduces an adaptive update weighting factor based on the characteristics of pixel points and gray-scale fluctuations in the pixel domain. The blur factor is selected by the weight value to reflect whether the domain pixel is positively correlated with the center pixel and whether the center pixel is a noise pixel. The algorithm has the advantage of being unconstrained by parameters while suppressing noise. The experimental results show that the proposed algorithm can overcome the shortcomings of traditional fuzzy C-means clustering in the image segmentation process, improve the segmentation accuracy and enhance the robustness of the algorithm.

References

1. WANG Qingchen, ZHENG Rongzhou, BAAD P D, et al, Cancer statistics in China, 2015 [J]. CA: A Cancer Journal for Clinicians, 2016, 66(2): 115-132.

2. Takahiro Nakajima, Kazuhiro Yasufuku. Early Lung Cancer: Methods for Detection [J]. *Clinics in Chest Medicine*, 2013, 34(3): 373-383.
3. SHARMAN P, WOOD-BAKER R. International lung disease due to fumes from heat-cutting polymer rope [J]. *Occup Med*, 2013, 63(6): 451-453.
4. Bezdek J C. Pattern recognition with fuzzy objective function algorithms [M]. New York: Plenum Press, 1981.
5. Ahmed M N, Yamany S M, Mohamed N, et al. A modified fuzzy C-means algorithm for bias field estimation and segmentation of MRI data [J]. *IEEE Transaction on Medical Imaging*, 2002, 21(3): 193-199.
6. Chen S C, Zhang D Q. Robust image segmentation using FCM with spatial constraints based on new kernel-induced distance measure [J]. *IEEE Transactions Systems, Man, and Cybernetics-Part B: Cybernetics*, 2004, 34(4): 1907-1916.
7. Szilagy L, Benyo Z, Szilagyii S M, et al. MR brain image segmentation using an enhanced fuzzy C-means algorithm [C] // *Proceeding of 25th Annual Informational Conference of IEEE EMBS. Cancun: IEEE Press*, 2003, 1: 724-726.
8. Cai W, Chen S, Zhang D Q. Fast and robust fuzzy C-means clustering algorithm incorporating local information for image segmentation [J]. *Pattern Recognition*, 2007, 40(3): 825-838.
9. Krinidis S, Chatzis V. A robust fuzzy local information C-means clustering algorithm [J]. *IEEE Transaction on Image Processing*, 2010, 19(5): 1328-1337.
10. CHAMA C K, MUKHOPADHYAY S, BISWAS P K, et al. Automated lung field segmentation in CT image using mean shift clustering and geometrical features [J]. *Medical Imaging, Computer-Aided Diagnosis*, 2013, 53(11): 1-10.
11. KORFIATIS P, KAZANTZIA A, KALOGEROPOULOU C, et al. Optimizing lung volume segmentation by texture classification [C] // *Proceedings of the IEEE/EMBS Region & International conference on Information Technology Application in Biomedicine*. [S.I.]: IEEE, 2010: 1-4.
12. Cao L, Zhan J, Yu X E, et al. Fast lung segmentation algorithm for thoracic CT based on automated thresholding [J]. *Computer Engineering and Applications*, 2008, 44(12): 178-181.
13. Wang B, Gu X M, Yang Y, et al. Automated lung segmentation for chest CT images based on random walk algorithm [J]. *Journal of Computer Applications*, 2015, 35(9): 2666-2672.
14. Pulagam A R, Kande G B, Ede V K R, et al. Automated lung segmentation from HRCT scans with diffuse parenchymal lung diseases [J]. *Journal of Digital Imaging*, 2016, 29(4): 507-519.
15. Lei T, Wang Y, Fan Y Y, et al. Vector morphological operators in HSV color space [J]. *Science China Information Science*, 2013, 56(1): 1-12.
16. Zarinbal M, Fazel Zarandi M H, Turksen I B. Interval type-2 relative entropy fuzzy C-means clustering [J]. *Information Sciences*, 2014, 272(10): 49-72.
17. Lei T, Wang Y, Wang G H, et al. Multivariate mathematical morphology based on fuzzy extremum estimation [J]. *IET Image Processing*, 2014, 8(9): 548-558.
18. Lei T, Fan Y Y. Noise gradient reduction based on morphological dual operators [J]. *IET Image Processing*, 2011, 5(1): 1-17.
19. Hwang C, Rhee F C H. Uncertain fuzzy clustering: interval type-2 fuzzy approach to C-means. *IEEE Transactions on Fuzzy System*, 2007, 15(1): 107-120.
20. News Center. The current situation of lung cancer incidence in 2014, and the mortality trend of lung cancer in the future [EB/OL]. 2014-04-09 [2016-05-10].