IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received June 22, 2020, accepted July 28, 2020, date of publication August 4, 2020, date of current version August 20, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3014219

Ancient Chinese Character Image Segmentation Based on Interval-Valued Hesitant Fuzzy Set

XUEDONG TIAN, TENGYING SUN, AND YANMEI QI
School of Cyber Security and Computer, Hebei University, Baoding 071002, China

Corresponding author: Xuedong Tian (xuedong_tian@126.com)

This work was supported by the National Natural Science Foundation of China under Grant 61375075, the Natural Science Foundation of Hebei Province of China under Grant F2019201329, and the Key Project of the Science and Technology Research Program in University of Hebei Province of China under Grant ZD2017208.

ABSTRACT To address the low segmentation accuracy caused by the rich glyph styles of ancient Chinese characters and the complex layout of ancient Chinese books, which affects the retrieval and recognition results, an algorithm for the layout image analysis of ancient Chinese books and Chinese character image segmentation is proposed. The initial segmentation results were obtained through the projection method of the layout of ancient Chinese books, and the connected component analysis of the above results was carried out to determine the rough divided blocks of under-segmentation and over-segmentation. Considering under-segmentation of adhesive Chinese characters, the improved K-means clustering method was used to segment adhesive blocks to obtain single-character images. To address the over-segmentation of character components separation, a method based on interval-valued hesitant fuzzy set is proposed. This method analyzed the features of the connected component in the block, characterized the over-segmentation connected component. The hesitant fuzzy distances between other connected components and the standard merge evaluation interval number were calculated in sequence. The connected component with the smallest distance was preferentially merged with the over-segmentation connected component until no over-segmentation connected component remained in the block. The experimental segmentation accuracy was 89.94%.

INDEX TERMS Ancient Chinese books, Chinese character segmentation, interval-valued hesitant fuzzy set, k-means clustering, layout analysis, layout image.

I. INTRODUCTION

Recently, from continuous advancements in ancient Chinese book research, computer technology is popularly used to address problems. Because ancient Chinese books were handwritten with the complex layout, and rich glyph styles of ancient Chinese characters, it is necessary to analyze the layout image of ancient Chinese books. In order to ensure the following image retrieval and recognition of Chinese characters in ancient Chinese books, the complete images of ancient Chinese characters must be segmented in the image segmentation stage.

In the research of text image segmentation, Qaroush *et al.* [1] presented an indirect segmentation-based algorithm in investigating text image segmentation. The proposed algorithm employed a projection profile method with the Interquartile Range statistical method to distinguish character parts in character part segmentation. It used a set of statistical and topological features invariant under font variations to distinguish actual segmentation points from all potential segmentation points in character segmentation. The method is suitable for segmenting Arabic character, Arabic character parts, and overlapping characters. Nguyen and Masaki [2] proposed a method based on a robust recognition model to segment characters in handwritten Japanese texts. Multi-segmentation points and over-segmentation after the first rough segmentation using the vertical projection and the Stroke Width Transform methods, and fine segmentation using bridge separation and Voronoi diagrams of text were addressed by searching a character with the optimal path for character recognition model. Under the assumption of a good character width estimation, this method can be suitable for segmenting adhesive characters. In addition, [3]–[7] studied text segmentation in other languages. Chinese characters are characterized by complex structures and large numbers hence there are different text segmentation methods

The associate editor coordinating the review of this manuscript and approving it for publication was Liangxiu Han.

in the segmentation process compared to others. Segmentation methods are popularly based on statistics [8], stroke characteristics [9], connected component analysis [10], and recognition [11], [12]. But it is effective to use an integrated method than a single method in addressing some complex segmentation. Li *et al.* [13] proposed a segmentation algorithm based on the analysis of clustering structure and stroke for the handwritten adhesion Chinese characters. Initially, the dividing line was determined to extract the adhesion stroke to analyze its type (straight line or curve), adhesion points and segmented direction. Furthermore, a background thinning algorithm was used to determine the segmentation curve. This method of addressing the adhesion of handwritten Chinese characters was robust and resistant to noise. Xu *et al.* [14] proposed a method for single-touching Chinese handwriting with learning-based filtering. It initially detected candidate segments by skeleton and contour analysis, and designed a filter by supervised learning to remove unreasonable candidate segments. This method can be used to segment adhesive Chinese characters irrespective of the length of the characters.

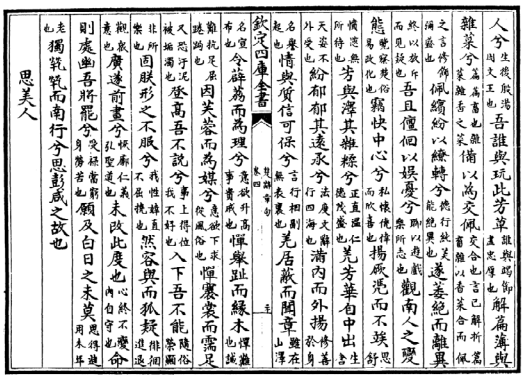
Researchers have studied the ancient Chinese character image Segmentation. The methods of projection, piecewise projection and segmentation of strokes features [15], and AP clustering algorithm can be applied in ancient Chinese character image segmentation [16]. Zhou *et al.* [17] put forward a multi-step segmentation method to segment characters of ancient Chinese books. Firstly, the projection method was employed to obtain the no adhesion characters from the rough divided blocks. Then, for the adhesive characters, the segmentation was performed by searching and modifying the segmentation path in the local neighborhood of initial segmentation path with minimum weight segmentation path algorithm. This method can have a good result on the vertical adhesion of Chinese characters, but it had the poor result in segmentation the horizontal adhesion of Chinese characters. Wu *et al.* [8] put forward multi-step segmentation method based on variable window for ancient Chinese character. Projection method was used to roughly segment characters of ancient Chinese books. The method of variable window was used to seek out segmentation path of every character in the character string and segment adhesive or overlapping characters. This method had a poor result in segmentation the horizontal adhesion of Chinese characters. Liu and Jin [18] proposed an improved Drop-fall algorithm to solve the character segmentation problem for ancient Korean and Chinese books. K-means clustering was initially performed on all possible starting drip points in the adhesion range. This method used the point closest to the cluster center as the final starting drip point. Finally, dripping was used to segment the adhesion characters. The segmented characters are merged according to the statistical threshold. Although this method has significant segmentation result on adhesion in Korean and Chinese characters, it can merge errors if processing over-segmentation characters.

In summary, the projection method and the connected component search method are reliable for the segmentation of Chinese characters in ancient Chinese books. However, it is difficult to get a complete image of single Chinese characters, because there is separation of internal components in a single Chinese character. If over-segmentation components are going to be merged, they are affected by multiple attributes. The problem of merger belongs to a multi-attribute decision-making problem, and interval-valued hesitant fuzzy sets can be used to solve the problems. Therefore, this study proposes a segmentation algorithm for Chinese characters in ancient Chinese books based on the interval-valued hesitating fuzzy set used to address the over-segmentations of Chinese characters in layout images of ancient Chinese books. The interval-valued hesitating fuzzy set decision method is used to establish the merging model to deal with the over-segmentation characters. The multi-step segmentation methods were adopted in this study. Firstly, Projection method was used to divided blocks roughly. Secondly, The K-means classification method was used to address the adhesive characters, and the interval-valued hesitating fuzzy set decision method was used to establish the merging model to deal with the over-segmentation characters.

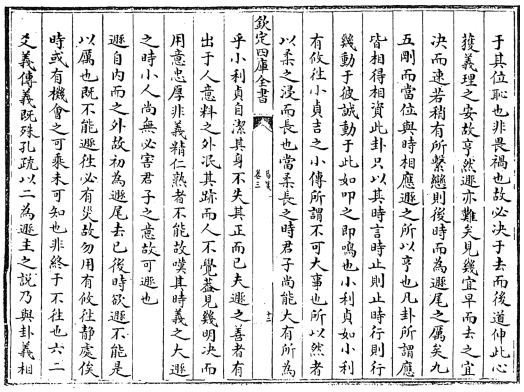
II. ROUGH DIVISION OF LAYOUT IMAGES IN ANCIENT CHINESE BOOKS

Layout images of ancient Chinese books are shown in Figure 1. There are the situations of adhesion and overlap of characters in the layout. Therefore, the projection method was used to segment the layout images but get an incomplete Chinese character image. Combined with the characteristics of the layout of ancient Chinese books for layout image analysis, the input is the layout image of ancient Chinese books, as shown in Figure 1. There is noise in the images of ancient Chinese books, and there is a tilt when scanning the images of ancient Chinese books. Remove the image noise of ancient Chinese books use a denoising method, and correct skew of the image employ the Hough transformation method which could find straight lines in an image [19], [20]. The rough divided block is obtained by column projection followed by row projection of the layout of ancient Chinese books, and searching the connected component in the block [21], [22]. The output is the sequence of the connected component, and the type of rough divided block is defined according to the connected component in the block. It is worth noting that Chinese characters consist of one or more components.

There are several types of connected components in the rough divided block, Table 2 shows that the rough divided block includes five types: complete block, overlapping block, single component block, adhesive block, and separation block. (a) If there are two or more connected components in a block, the block maybe an overlapping block, a separation block, or a complete block with two complete characters. In these three kinds of blocks, there are overlapping components in the overlapping block; there are two components which ratio of the width and height is limited to r_l and r_u in the complete block, and heights of components are limited to h_l and h_u; and the other is the separation block. (b) If there is one



(a) A column of a typesetting

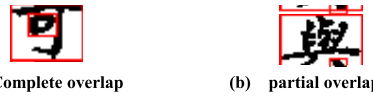


(b) A column of mixed typesetting

FIGURE 1. Layout images of Chinese characters in ancient Chinese books.

connected component in a block, the block maybe a complete block, a single component block or an adhesive block. In a complete block, the component has the height limited to h_l and h_u, and the ratio of the width and height limited to r_l and r_u. In a single component block, the component has the height less than h_l. In an adhesive, the conditions are shown in Table 1. Where r_l, and r_u refer to the lower threshold of the ratio of the width and height, and upper threshold of that, respectively; h_l, and h_u refer to the lower threshold of the height of an ancient Chinese character, and upper threshold of that, respectively. The parameters are statistics from the ancient Chinese books.

Different kinds of blocks need to be processed with different strategies. The complete block requires no operation in the next step. There is under-segmentation in adhesive blocks, and there is over-segmentation in overlapping blocks, single component blocks and separation blocks, so these four types require more operations in the next step. The connected components of overlapping blocks are merged by estimating the common relationship between the connected components. As shown in Figure 2, the relationship includes complete overlap and partial overlap. For a single component block, the Euclidean distance between the single component block, and the upper and lower two blocks with a single connected component is calculated, and the connected component to the smaller distance is selected to merge with the single component block.



(a) Complete overlap (b) partial overlap

FIGURE 2. Overlapping block.

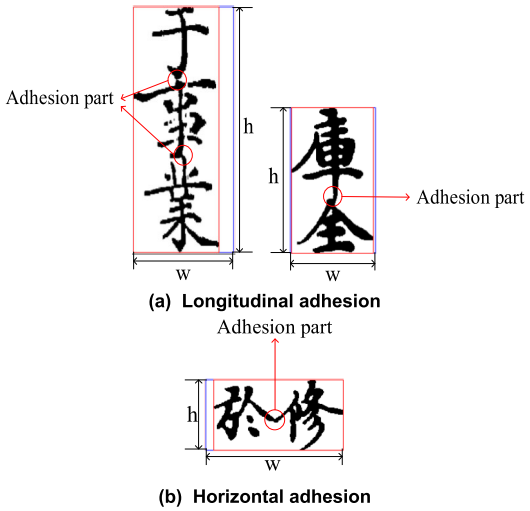


FIGURE 3. Adhesive block.

III. THE SEGMENTATION OF ADHESIVE BLOCKS

The failure of obtaining an image of a single character from an adhesive block by connected component searching is an under-segmentation problem. In ancient Chinese books, there are at least two characters in the longitudinal adhesion and two characters in the transverse adhesion as shown in Figure 3. The writing of Chinese characters is characterized by high cohesion. The pixels of the adhesive block can be classified according to the clustering result. Compared with other clustering algorithms, the K-means algorithm is a simple and fast unsupervised learning algorithm [23]. Therefore, the K-means clustering algorithm can be used to cluster adhesive parts to achieve significant clustering result to segment the adhesive Chinese characters.

In Figure 3, the height and width of the adhesive block are denoted by h, and w, respectively. Longitudinal adhesion is shown in Figure 3(a), and the number of characters in the adhesive block is determined by comparing the average height of statistics with that of the block. The number of characters is used as the number of clustering categories for the adhesive block. After the clustering, there are several categories from top to bottom. The upper and lower edge points in the longitudinal direction of class 1 are used as the upper and lower initial drop-fall points of the class 1 representative character to segment the first adhesive character with Drop-fall algorithm [24]; the upper and lower edge points in the longitudinal direction of class 2 are used as the upper and lower initial drop-fall points of the class 2 representative character to segment the second adhesive character with Drop-fall algorithm; and so on to get the subsequent adhesive characters. As shown in Figure 3(b), the horizontal adhesion block has two-character adhesion, and two clustering

TABLE 1. *K* Value Conditions.

K	Condition
2	$\frac{1}{3}h_0 < h < h_0$ and $\frac{w}{h} > r_u$
$\frac{h}{h_0}+1$	$h > 1.2h_0$

categories. The conditions for the number of clustering categories K are shown in Table 1, where h, and w refer to height of the adhesive block, and width of the adhesive block, respectively. And h_0 refers to the average height.

IV. MERGING OF SEGREGATED CHINESE CHARACTER COMPONENTS

A. RELATED CONCEPTS OF INTERVAL-VALUED HESITANT FUZZY SET

In practice, decision problems that cannot be represented by exact real numbers are evaluated with reasonable interval numbers. Closed interval $a = [a^-, a^+]$ is an interval number [25]. Like real numbers, interval numbers can also be compared in size, using the method of degree to judge the relative size of interval numbers [26].

The interval-valued hesitant fuzzy set for $X = \{x_1, x_2, \ldots, x_n\}$ is $\tilde{E} = \{< x, \tilde{h}_{\tilde{E}}(x) > | x \in X\}$, where $\tilde{h} = \tilde{h}_{\tilde{E}}(x)$ is the interval-valued hesitation fuzzy element represents the set of possible interval numbers that contains x in X [27]. Let $\tilde{A}$ and $\tilde{B}$ be two interval-valued hesitation fuzzy sets about $X = \{x_1, x_2, \ldots, x_n\}$, [28] proposed the Generalized interval hesitation ordered weighted Hamming distance measure of $\tilde{A}$ and $\tilde{B}$, and it is expressed as

$$d_{GIHOWD}\left(\tilde{A}, \tilde{B}\right) = [\sum_{i=1}^{n} \omega_i \left(\frac{1}{2l_{x_{\sigma(i)}}} \sum_{j=1}^{l_{x_{\sigma(i)}}} \left(\left| (\tilde{h}_{\tilde{A}}^{\sigma(i)}\left(x_{\sigma(i)}\right))^{+} - (\tilde{h}_{\tilde{B}}^{\sigma(i)}\left(x_{\sigma(i)}\right))^{+} \right| + \left| (\tilde{h}_{\tilde{A}}^{\sigma(i)}\left(x_{\sigma(i)}\right))^{-} - (\tilde{h}_{\tilde{B}}^{\sigma(i)}\left(x_{\sigma(i)}\right))^{-} \right| \right) \right)] \quad (1)$$

where $\tilde{h}_{\tilde{A}}^{\sigma(j)}\left(x_{\sigma(i)}\right)$ and $\tilde{h}_{\tilde{B}}^{\sigma(j)}\left(x_{\sigma(i)}\right)$ are the interval numbers with the j-th largest in $\tilde{h}_{\tilde{A}}(x_i)$ and $\tilde{h}_{\tilde{B}}(x_i)$, respectively, $l_{x_{\sigma(i)}} = \max(l_{\tilde{A}}(x_i), l_{\tilde{B}}(x_i))$, $l_{\tilde{A}}(x_i)$ and $l_{\tilde{B}}(x_i)$ represent the interval numbers in $\tilde{h}_{\tilde{A}}(x_i)$ and $\tilde{h}_{\tilde{B}}(x_i)$, respectively, w_i is determined by using the standard distribution idea as the position weight of the evaluation attribute [28], and σ is a ranking function defined in interval-valued hesitant fuzzy set.

B. MERGING MODEL BASED ON INTERVAL-VALUED HESITANT FUZZY SET

The number of ancient Chinese characters is huge with various writing styles, which results in many types of over-segmentation. The deep learning model is difficult to meet the need of the merge operations of over-segmentation in training. In the over-segmentation problem of Chinese character images in ancient Chinese books, there are many attributes that influence evaluation of the connected components in separation blocks. Hesitant fuzzy set can be introduced in multi-attributes decision to solve the problem of describing the correspondence between the connected components under the merged attributes. Additionally, it is difficult to quantify the merged evaluation indicators of the connected components. Generally, qualitative evaluation levels such as ''larger'', ''smaller'', ''more distant'', and ''closer'', and other qualitative evaluation levels are used to describe the size and distance, and other indicators. Because the hesitant fuzzy element in the hesitant fuzzy set decision-making method uses some value between 0 and 1 to represent the evaluation index, some evaluation results have significant one-sidedness. The interval value hesitant fuzzy set, if evaluating the attributes of an object, represents and uses the evaluation information in the form of a set of possible interval values to quantify each factor. Therefore, the interval-valued hesitation fuzzy set can effectively address the uncertainty in over-segmentation evaluations. The interval-valued hesitant fuzzy evaluation method is a comprehensive evaluation of the membership status of an evaluated object from several perspectives. It can effectively solve multi-attribute decision-making problems, is practicable, and can be used for the merged evaluation of connected components.

Definition 1: Suppose that the current over-segmentation connected component of the evaluation problem is c_0, and the set of the connected components to be evaluated is $C = \{c_1, c_2, \ldots, c_n\}$. The set of over-segmentation factors of each evaluation attribute is $P = \{p_1, p_2, \ldots, p_m\}$. The weight set $W = \{w_1, w_2, \ldots, w_m\}$, w_i is the importance of each attribute in the evaluated system and satisfies $\sum_{i=1}^{m} w_i = 1$. The evaluation result is $D = \{d_1, d_2, \ldots, d_n\}$, and $d_i (i = 1, 2, \ldots, n)$ denotes the merged evaluation value of c_i and c_0 under the evaluation attribute P.

C. MERGED EVALUATION ATTRIBUTES

That some Chinese characters are over-segmented in separation blocks is an over-segmentation, that is, one Chinese character has multiple connected components. The evaluation of the possibility of merging an over-segmentation connected component with its surrounding connected components is related to the layout of ancient Chinese books, the relative situation of connected components, and the shape and size of Chinese characters. Among these, the layout attributes of ancient Chinese books can be quantified by the midline and width indexes. The relative situation of connected components can be understood because the relative local attributes of Chinese characters can be quantified by the distance, pixel, offset, and relative size. The shape and size of Chinese characters can be quantified by the aspect ratio and the size after the merger, respectively.

Algorithm 1 Merging Algorithm for Separated Chinese Character Components in Ancient Chinese Books

Input: Position coordinates of the separation block

Output: A coordinate set of position of the connected component of every ancient Chinese character

Step 1 Input the position coordinates of the separation block, and determine the set of connected components to be evaluated, $C = \{c_1, c_2, \ldots, c_n\}$. Each object in the set is a connected component in the separation block except the current over-segmentation connected component.

Step 2 Determine the set of over-segmentation attributes. There are eight factors in attributes, $P = \{p_1, p_2, \ldots, p_8\} = \{I_{mid}, I_{width}, I_{dist}, I_{pix}, I_{off}, I_{rel-s}, I_{shape}, I_s\}$.

Step 3 Determine the fuzzy relation matrix

$$\tilde{H} = \begin{bmatrix} h_{11} & \cdots & h_{1n} \\ \vdots & \ddots & \vdots \\ h_{m1} & \cdots & h_{mn} \end{bmatrix},$$

$\tilde{H}$ is the evaluation matrix, $\forall x_i$, $i \in N$, the row vector $(h_{i1}, h_{i2}, \ldots, h_{im}) \in [0, 1]^m$ is the fuzzy attribute vector D of x_i, and represents the evaluation value of c_i in the evaluation attribute p_i.

Step 4 Calculate the interval-valued hesitant fuzzy decision matrix and weight vector. Calculate the eight evaluation factors for each connected component using formulas (2) to (9). Calculate the weight vector W.

Step 5 Calculate the interval hesitation order weighted distance measure $d(c_i, c_0)$. Under each factor, the standard over-segmentation evaluation value is [1, 1], and the distance measure of c_i and c_0 is calculated by formula (1).

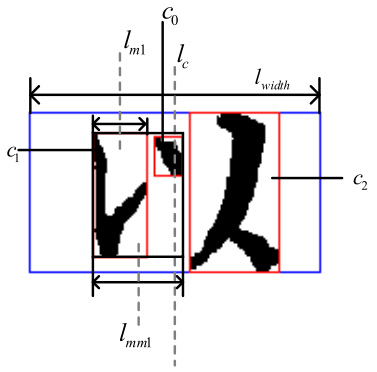
Step 6 Merge the connected components. The smaller $d(c_i, c_0)$, the higher the evaluation value of merging, and the target is preferentially merged with the current over-segmentation connected component.

Step 7 Repeat Steps 1-6 until there is no over-segmentation connected component in the separation block, and output the position coordinates of the connected component of every ancient Chinese character.

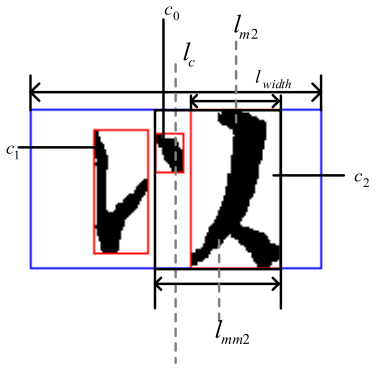
1) LAYOUT ATTRIBUTES OF ANCIENT CHINESE BOOKS

a: THE MIDLINE INDEX

The midline index, I_{mid}, can be used as the evaluation index of the merger of over-segmentation Chinese characters in the separation block. The I_{mid} evaluates the value of c_i and c_0 merged evaluation interval value by the relative positional relationship of the column midline where the block is roughly divided, the midline of c_i, and the midline of the whole of c_i and c_0. The evaluation value increases with decreasing distance between the midline of the whole of c_i and c_0 and the midline of the column. The evaluation value decreases with increasing distance between the midline of the whole of c_i and c_0 and the midline of the column. As shown in Figure 4, the connected components with proposed merger are considered as a whole region. Compared with the distance



(a) c_0 and c_1 merge



(b) c_0 and c_2 merge

FIGURE 4. Schematic of midline index.

between the midline of the column and that of c_i, the length which represents the distance between midline of the column and that of the whole region is decrease, and the evaluation value of the connected component increases. According to this law, design the evaluation function. Thereafter, the evaluation interval values of c_1 and c_2 are calculated. The connected component with a larger evaluation interval value is merged with c_0.

Definition 2: The evaluation function under I_{mid} is

$$\begin{aligned} & f_{I_{mid}}(c_0, c_i) \\ &= [f^-, f^+] \\ &= \begin{cases} [e^{\frac{-|l_{mmi}-l_c|}{l_{width}}}, e^{\frac{-|l_{mi}-l_c|}{l_{width}}}], |l_{mmi} - l_c| > |l_{mi} - l_c| \\ [e^{\frac{-|l_{mi}-l_c|}{l_{width}}}, e^{\frac{-|l_{mmi}-l_c|}{l_{width}}}], |l_{mmi} - l_c| \leq |l_{mi} - l_c| \end{cases} \end{aligned} \quad (2)$$

where l_{width} refers to the width of the column where the separation block is located, l_c to the middle line of the column where the separation block is located, l_{mi} is the midline of the connected component, l_{mmi} is the midline of the connected component after the proposed merger of c_0 and c_i.

b: THE WIDTH INDEX

In the separation block, the width index denoted by I_{width} can be used as the evaluation index for the merger of

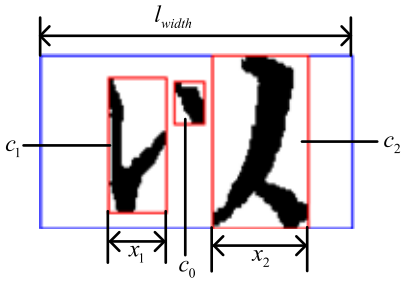


FIGURE 5. Schematic of width index.

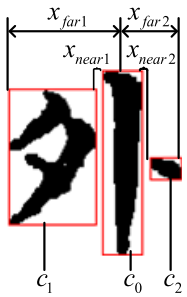


FIGURE 6. Schematic of distance index.

over-segmentation Chinese characters. The I_{width} evaluates the merged evaluation value of c_i and c_0 by the width of c_i and the width of column which the separation block is located. The merged evaluation value of c_i and c_0 increases with decreasing width of c_i. According to this law, design the evaluation function. From Figure 5, the size of x_2 is greater than that of x_1. Therefore, under the width index, the merging evaluation value of c_1 and c_0 is higher, and the merger of them is preferred.

Definition 3: The evaluation function under I_{width} is

$$f_{I_{width}}(c_0, c_i) = [f^-, f^+] = [\frac{l_{width} - x_i}{l_{width}}, \frac{l_{width} - x_i}{l_{width}}] \quad (3)$$

where l_{width} refers to the width of the column where the separation block is located, and x_i refers to the width of c_i.

2) RELATIVE SITUATION OF CHINESE CHARACTERS IN ANCIENT CHINESE BOOKS

a: THE DISTANCE INDEX

The distance index, I_{dis}, can be used as the evaluation index for the merger of over-segmentation Chinese characters. The I_{dis} evaluates c_i according to the principle that the merger increases with decreasing relative distance c_0. In Figure 6, the connected component has two borders. If use the near boundary of the reference to the current c_0 to evaluate, have $x_{near1} < x_{near2}$, so the first merger of c_1 and c_0, if use the far boundary of the reference to c_0 to evaluate, have $x_{far2} < x_{far1}$, so the first merger of c_2 and c_0. So with reference to different boundary to describe the index is in ambiguity. Therefore, the interval value can accurately describe the evaluation index, and the evaluation function is constructed

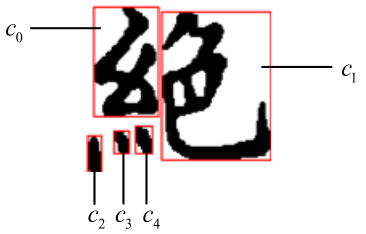


FIGURE 7. Schematic of pixel index.

using the side of the connected component to be evaluated near and far from the current over-segmentation connected component. Design the evaluation function can describe this index.

Definition 4: The evaluation function under I_{dist} is

$$f_{I_{dist}}(c_0, c_i) = [f^-, f^+] = [\frac{1}{\ln^2(x_{fari} + 1)}, \frac{1}{\ln^2(x_{neari} + 1)}] \quad (4)$$

where x_{fari} and x_{neari} refer to the distance values on the near side and the distance on the far side of c_0 and c_i, respectively.

b: THE PIXEL INDEX

In the separation block, the pixel index, I_{pix}, can be used as the evaluation index for the merger of over-segmentation Chinese characters. There are fewer pixels in c_i, the greater the possibility of over segmentation is, and the higher the merged evaluation value is. According to this law, design the evaluation function. In Figure 7, the connected component c_3 has the least pixels compared with the other connected components in the block, and the merged evaluation value in this index is the highest. Thus, c_0 and c_3 are merged preferentially.

Definition 5: The evaluation function under I_{pix} is

$$f_{I_{pix}}(c_0, c_i) = [f^-, f^+] = [\frac{x_{p0}}{x_{p0} + x_{pi}}, \frac{x_{p0}}{x_{p0} + x_{pi}}] \quad (5)$$

where x_{p0} is the number of pixels in c_0, and x_{pi} is the number of pixels in c_i.

c: THE OFFSET INDEX

The offset index denoted by I_{off} can be used as the evaluation index for the merger of over-segmentation Chinese characters. The smaller the offset of c_i is relative to c_0, the more merger is needed, and the higher the merger evaluation value is. The offset of c_i from c_0 is divided into two parts: upper offset and lower offset. It cannot be accurately represented by a real value; hence, the interval value is used to describe the offset. The schematic of the offset index is shown in Figure 8: the smaller the offset, the higher the evaluation value of c_i, and the higher the priority of merging it with c_0. According to this law, design the evaluation function.

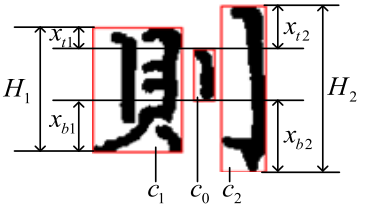


FIGURE 8. Schematic of offset index.

Definition 6: Evaluation function under I_{off} is

$$f_{I_{off}}(c_0, c_i) = [f^-, f^+] = \left[1 - \frac{\max(x_{ti}, x_{bi})}{H_i}, 1 - \frac{\min(x_{ti}, x_{bi})}{H_i}\right] \quad (6)$$

where x_{ti} and x_{bi} refer to the upper and lower boundary offsets of c_i relative to c_0, respectively. H_i refers to the overall height of c_i and c_0 after the proposed merger.

d: THE RELATIVE SIZE INDEX

If evaluating the possibility of the merger c_i and c_0, the relative size denoted by I_{rel-s} can be used as the evaluation index. The smaller the relative size of a connected component, the higher the merger evaluation value is, and the higher the priority of merging it with c_0. According to this law, design the evaluation function.

Definition 7: The evaluation function under the I_{rel-s} is

$$f_{I_{rel-s}}(c_0, c_i) = [f^-, f^+] = [\frac{x_{s0}}{x_{s0} + x_{si}}, \frac{x_{s0}}{x_{s0} + x_{si}}] \quad (7)$$

where x_{s0} is the size of the connected component c_0, and x_{si} is the size of the connected component c_i.

3) SHAPE AND SIZE OF CHINESE CHARACTERS

a: THE SHAPE INDEX

The shape index is I_{shape}. In this study, the separation blocks are mostly separated from the left and right sides. In merging, it is suitable to keep the connected component in a "slender" shape, that is, the height is slightly larger than the width, and the aspect ratio describes the shape characteristics. If this feature is used as the evaluation index, the shape is evaluated first. If the shape is "slender", the merging evaluation value is higher. In contrast, the "wide and short" shapes were rated as low. From Figure 9, the proposed merging results show that the "slender" shape is consistent with the characteristics of ancient Chinese characters than the "wide and short" shape, so c_0 from c_1 are merged in priority.

Square characters are characteristics of ancient Chinese books, hence the closer the connected component is to square characters in the process of merging, the higher its evaluation value. From Figure 10, the aspect ratio of the proposed combination of c_0 and c_1 is less than that of the proposed combination of c_0 and c_1 that is consistent with the characteristics of Chinese characters in ancient Chinese books. Therefore, the combination of c_0 and c_1 is preferred.

According to these laws, design the evaluation function.

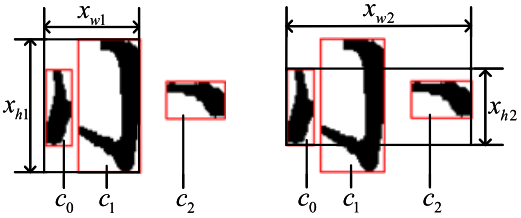


FIGURE 9. Schematic of shape index.

Definition 8: The evaluation function under I_{shape} is

$$f_{I_{shape}}(c_0, c_i) = [f^-, f^+] \begin{cases} [\frac{x_{wi}}{x_{hi}}, 1], x_{hi} \geq x_{wi} \\ [0, \frac{x_{wi}}{x_{hi}}], x_{hi} < x_{wi} \end{cases} \quad (8)$$

where x_{wi} and x_{hi} refer to the width and height after the merger of c_0 and c_i, respectively.

b: THE SIZE INDEX

The size of a Chinese character is the internal information of ancient Chinese characters. The size of the merged Chinese character is used as a merged evaluation standard, and the size index is I_s. The merged evaluation value is obtained by comparing the size of the whole of c_i and c_0 and the average size of ancient Chinese characters obtained from the statistics. Considering that merging the over-segmentation connected components to the greatest extent, the best premise of the merger is that the whole of c_i and c_0 is larger than the average and less than twice the average, followed by smaller than the average, and no combination is possible if it is larger than twice the average, the evaluation value is 0 namely. According to this law, design the evaluation function.

Definition 9: The evaluation function under I_s is

$$f_{I_s}(c_0, c_i) = [f^-, f^+] \begin{cases} [0, 1 - \frac{s_0 - s_i}{s_0}], & s_i - s_0 < 0 \\ [1 - \frac{s_i - s_0}{s_0}, 1], & 0 \leq s_i - s_0 < s_0 \\ [0, 0], & s_i - s_0 \geq s_0 \end{cases} \quad (9)$$

where s_i is the size of the connected component after the merger of c_0 and c_i, and s_0 is the global average size of Chinese characters in ancient Chinese books.

V. EXPERIMENTAL RESULTS AND ANALYSIS

A. EXPERIMENTAL SETTINGS

The accuracy of ancient Chinese character image segmentation is used to evaluate the advantages and disadvantages of the segmentation method. In this study, the accuracy P of text segmentation is defined as

$$P = \frac{N_T}{N_{sum-pri}} \times 100\% \quad (10)$$

where $N_{sum-pri}$ refers to the total number of Chinese characters in the layout image sample of ancient Chinese books, and

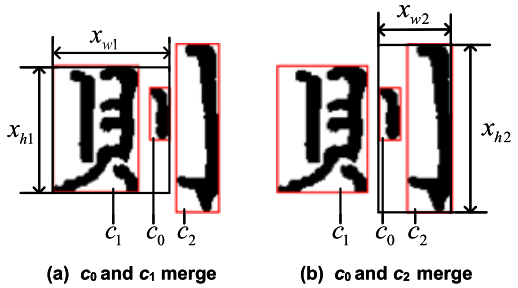


FIGURE 10. Schematic of "slender" shape.

N_T refers to the number of Chinese characters with correct segmentation.

B. EXPERIMENTAL RESULTS

"Si ku Quan shu" is a representative ancient literature. Si Ku Quan Shu of Wenyuan Pavilion is a source of experimental data for researchers to analyze the layout of ancient Chinese books and retrieve Chinese characters. A randomly selected 100 layout images from 1500 volumes and 28110 Chinese characters were used as experimental samples.

Unlike modern layout writing specifications, the layout images of ancient Chinese books are in a vertical layout, and Figure 1 in Section II is the two layout images in the experimental sample. From Figure 1, the main features of the ancient Chinese characters layout image. First, the typeset of the column text is different, the number is uncertain; second, the characters are stuck, broken stroke, and components are separated.

After processing the layout image of ancient Chinese books, all rough divided blocks are obtained. Among them, the complete block includes one Chinese character or two Chinese characters. The type of overlapping block includes over-lapping completely and partly between components, also there is over-lapping between Chinese characters. The adhesive block includes two kinds of horizontal and vertical adhesion, and the separation block is that there is separation of Chinese character components. Some experimental results are shown in Table 2.

After rough division, several operations were carried out to merge the single component block with the connected component with the least Euclidean distance, merge connected components in overlapping blocks, segment characters in adhesive blocks, and merge in separation blocks to obtain single characters. The experimental results of the segmentation of Chinese characters are shown in Table 3.

The algorithm in this study is used to segment the layout images of ancient Chinese books, and the accuracy of each stage in the segmentation process is calculated using formula (10). The Chinese character segmentation results obtained are shown in Table 4.

The data in Table 4 shows that the segmentation accuracy of the algorithm is 89.94%, particularly in the separation merging stage, and the accuracy is significantly improved, with a range of 19.45%. Therefore, the method of interval-valued hesitation fuzzy set can solve the over-segmentation of separation in the segmentation of Chinese characters.

TABLE 2. Some experimental results.

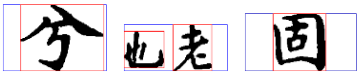
TypeID	Block Type	Some examples
1	Complete	
2	Single component	
3	Overlapping	
4	Adhesive	
5	Separation	

TABLE 3. Experimental results of the segmentation part of chinese characters.

TypeID	Block Type	Some examples
2	Single component	
3	Overlapping	
4	Adhesive	
5	Separation	

TABLE 4. The chinese character segmentation results.

Segmentation stage	P(%)
Rough segmentation	62.73
Overlap merging	65.22
Adhesion segmentation	70.49
Separation merging	89.94

C. COMPARATIVE ANALYSIS

The experiment used the algorithms in [29] and [8] to compare with the algorithm in this study. The character segmentation algorithm proposed in [29] is divided into two steps: rough segmentation determines the approximate

(a) Results of literature [29] (b) Results of literature [8] (c) Results of this study

FIGURE 11. The partial interception of the segmentation results.

segmentation position, and further fine segmentation is performed based on the method of connected component analysis and adhesion point assessment. The algorithm proposed in [8] is also divided into two steps: roughly segmentation of ancient Chinese character, and the method of variable window was used to seek out segmentation path of every character in the Chinese character string. With the algorithm in this study, the layout characteristics of ancient Chinese books are considered. After rough division, the interval-valued hesitation fuzzy set theory is introduced to address over-segmentation Chinese characters, multiple functions under multi-attribute indexes are designed, and the connected components are iteratively merged according to the interval-valued hesitant fuzzy distance until Chinese characters are not over-segmented.

1) SEGMENTATION RESULTS

Figure 11(a) shows the partial interception of the segmentation results of the page images of an ancient Chinese book using the algorithm in [29], Figure 11(b) shows that from using the algorithm in [8], and Figure 11(c) shows that from using the algorithm in this study. The Chinese characters of the overlapping and adhesion problems can be well divided with the three algorithms. The Chinese characters with radical separation structure have wrong segmentation results using the algorithms in [29] and [8] from Figure 11(a) and Figure 11(b). It evident that notwithstanding the up-down or left-right structure of Chinese characters, over-segmentation is solved, and the over-segmentation connected components are merged under the algorithm in this study from Figure 11(c).

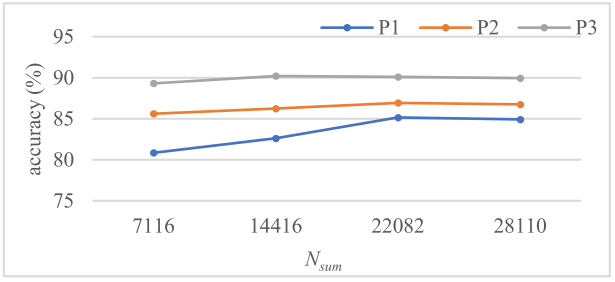


FIGURE 12. Contrast line chart of the segmentation accuracy.

TABLE 5. The Segmentation accuracy of Chinese characters in Ancient Chinese Books.

N_{sum}	P_1(%)	P_2(%)	P_3(%)
7116	80.85	85.61	89.31
14416	82.63	86.23	90.21
22082	85.16	86.93	90.1
28110	84.93	86.74	89.94

2) THE SEGMENTATION ACCURACY

The 100 layout images mentioned above were used as comparative experimental data. The algorithms proposed in [29] and [8] were simulated and the results were compared with that from the algorithm proposed in this study. The segmentation accuracy was calculated as shown in Table 5, and the contrast line chart of the segmentation accuracy is shown in Figure 12.

Where N_{sum} refers to the total number of Chinese characters in the dataset, P_1 refers to the accuracy of the segmentation of the algorithm in [29], P_2 refers to the accuracy of the algorithm in [8], and P_3 refers to the accuracy of

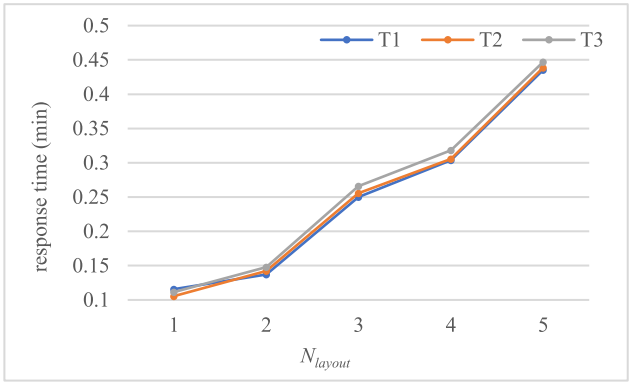


FIGURE 13. Contrast line chart of the response time.

the algorithm in this study. From Table 5 and Figure 12, the proposed algorithm can achieve higher accuracy in segmenting Chinese characters from the images of ancient Chinese books. In the proposed algorithm, layout, block relative, the morphological and size attributes of the characters are quantified and evaluated. After hesitant fuzzy distance calculation and connected component merging, the separated Chinese characters can be accurately merged. Therefore, the proposed algorithm is applicable in Chinese character segmentation of ancient Chinese book layout images.

3) RESPONSE TIME OF THE SEGMENTATION

The contrast line chart of the response time is shown in Figure 13.

Where N_{layout} refers to the number of the layout images of ancient Chinese books, T_1 refers to the response time of the algorithm in [29], T_2 refers to that of the algorithm in [8], and T_3 refers to that of the algorithm in this study. From Figure 13, although the response time of this study is slightly slower than the comparative methods which is used to perform evaluation of each attribute and calculate the interval-valued hesitation fuzzy distance for the better solutions of over-segmentation, it is within the acceptable level.

VI. CONCLUSION

Aiming at the layout image features of ancient Chinese books, particularly the layout image of ancient Chinese books represented by Si Ku Quan Shu of Wenyuan Pavilion, this study proposes a segmentation method for images of ancient Chinese books with rough divisions, fine divisions, and fine combinations. The experimental results show that in the process of ancient Chinese character segmentation, the adhesive characters are effectively segmented, and the over-segmentation components are merged accurately. Combined with the height feature and K-means algorithm, the segmentation of ancient Chinese books can be completed without influence from layout and adhesive morphology. Evaluation indexes and evaluation functions under multiple attributes are designed, and the interval-valued hesitation fuzzy set is constructed. The hesitation fuzzy distance between the connected component to be evaluated and the standard merging connected component is calculated, and the merger of the connected component with the smallest distance and the over-segmentation connected component can obtain the complete Chinese character. The experimental results show that the segmentation accuracy of this method is 89.94% and can achieve the expected effect.

ACKNOWLEDGMENT

The authors extend their appreciation to the National Natural Science Foundation of China under Grant 61375075, the Natural Science Foundation of Hebei Province of China under Grant F2019201329, and the Key Project of the Science and Technology Research Program in University of Hebei Province of China under Grant ZD2017208.

REFERENCES

[1] A. Qaroush, B. Jaber, K. Mohammad, M. Washaha, E. Maali, and N. Nayef, "An efficient, font independent word and character segmentation algorithm for printed Arabic text," *J. King Saud Univ. Comput. Inf. Sci.*, Aug. 2019, doi: 10.1016/j.jksuci.2019.08.013.
[2] K. C. Nguyen and N. Masaki, "Enhanced character segmentation for format-free japanese text recognition," in *Proc. 15th Int. Conf. Frontiers Handwriting Recognit. (ICFHR)*, Shenzhen, China, Oct. 2016, pp. 138–143, doi: 10.1109/ICFHR.2016.0037.
[3] N. K. Garg, L. Kaur, and M. K. Jindal, "Segmentation of touching modifiers and consonants in middle region of handwritten Hindi text," *Pattern Recognit. Image Anal.*, vol. 25, no. 3, pp. 413–417, Sep. 2015, doi: 10.1134/S1054661815030050.
[4] M. K. Sharma and V. P. Dhaka, "Segmentation of English offline handwritten cursive scripts using a feedforward neural network," *Neural Comput. Appl.*, vol. 27, no. 5, pp. 1369–1379, Jul. 2016, doi: 10.1007/s00521-015-1940-x.
[5] Y. S. Hwang, K. A. Moon, S. Y. Chi, D. G. Jang, and W. G. Oh, "Segmentation of a text printed in korean and English using structure information and character recognizers," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (IEEE-SMC)*, Nashville, USA, TN, USA, Oct. 2000, pp. 1586–1591, doi: 10.1109/ICSMC.2000.886248.
[6] Z. W. Jiang, X. Q. Ding, L. R. Peng, and C. S. Liu, "Uyghur character models with shared structure information for segmentation-free recognition under low data resource conditions," *J. Electron. Inf. Technol.*, vol. 37, no. 9, pp. 2103–2109, Sep. 2015, doi: 10.11999/JEIT150019.
[7] G. Tuerxun, Y. Aysa, T. Yibulayin, and K. Ubul, "Combination of connected regions and overlapping degree based Uyghur document image text segmentation," *Comput. Eng. Des.*, vol. 37, no. 7, pp. 1892–1897, Jul. 2016, doi: 10.16208/j.issn1000-7024.2016.07.036.
[8] X. J. Wu, Z. L. Zhong, and S. L. Zhou, "Multi-step segmentation method based on variable window for ancient handwritten Chinese character," *Comput. Eng. Des.*, vol. 37, no. 4, pp. 1102–1106, Apr. 2016, doi: 10.16208/j.issn1000-7024.2016.04.049.
[9] B. Li, J. Q. Wang, H. Y. Wei, Y. G. Sun, and X. N. Wang, "Automatic font normalization for handwritten women's script," *J. Chin. Inf. Process.*, vol. 29, no. 2, pp. 142–149, Mar. 2015.
[10] B. Y. Niu, L. L. Huang, and J. Hu, "Detection and recognition algorithm for license plate in natural scene," *J. Signal Process.*, vol. 32, no. 7, pp. 787–794, Jul. 2016, doi: 10.16798/j.issn.1003-0530.2016.07.005.
[11] C. L. Liu, M. Koga, and H. Fujisawa, "Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1425–1437, Nov. 2002, doi: 10.1109/TPAMI.2002.1046151.
[12] W. Y. Yang and S. W. Zhang, "An integrated segmentation and recognition algorithm for text in video," *Acta Automatica Sinica*, vol. 36, no. 10, pp. 1468–1476, Oct. 2010, doi: 10.3724/SP.J.1004.2010.01468.
[13] X. Y. Li, F. Yang, and W. B. Zhang, "Segmentation of connected handwritten Chinese characters based on structure cluster and strokes analysis," *Comput. Eng. Appl.*, vol. 44, no. 34, pp. 163–165, Dec. 2008, doi: 10.3778/j.issn.1002-8331.2008.34.050.
[14] L. Xu, F. Yin, Q. F. Wang, and C. L. Liu, "An over-segmentation method for single-touching Chinese handwriting with learning-based filtering," *Int. J. Document Anal. Recognit. (IJDAR)*, vol. 17, no. 1, pp. 91–104, Mar. 2014, doi: 10.1007/s10032-013-0208-1.
[15] Z. L. Zhang, X. J. Wu, and S. L. Zhou, "Study on the segmentation method of handwritten characters from historical chinese documents," *J. Zhengzhou Univ. (Eng. Sci.)*, vol. 36, no. 6, pp. 70–75, 2015, doi: 10.3969/j.issn.1671-6833.2015.06.014.

[16] Z. H. Zheng, C. Q. Huang, Y. Liang, L. C. Ran, and W. Y. Tian, "Chinese character segmentation based on AP clustering," *Intell. Comput. Appl.*, vol. 8, no. 1, pp. 65–67 and 71, Feb. 2018.
[17] S. F. Zhou, C. P. Liu, G. Liu, and S. R. Gong, "Multi-step segmentation method based on minimum weight segmentation path for ancient handwritten Chinese character," *J. Chin. Comput. Syst.*, vol. 33, no. 33, pp. 614–620, Mar. 2012.
[18] X. C. Liu and X. F. Jin, "Characters segmentation method of historical documents mixed in Korean and Chinese," *Comput. Eng. Appl.*, vol. 56, no. 11, pp. 135–141, Jun. 2020.
[19] D. Scazzoli, G. Bartezzaghi, D. Uysal, M. Magarini, M. Melacini, and M. Marcon, "Usage of Hough transform for expiry date extraction via optical character recognition," in *Proc. Adv. Sci. Eng. Technol. Int. Conf. (ASET)*, Dubai, United Arab Emirates, Mar./Apr. 2019, pp. 1–6, doi: 10.1109/ICASET.2019.8714306.
[20] Z. Shi, B. Xu, X. Zheng, and M. Zhao, "A Chinese character structure preserved denoising method for Chinese tablet calligraphy document images based on KSVD dictionary learning," *Multimedia Tools Appl.*, vol. 76, no. 13, pp. 14921–14936, Jul. 2017, doi: 10.1007/s11042-016-4284-3.
[21] Y. Wang, W. Wang, Z. Li, Y. Han, and X. Wang, "Research on text line segmentation of historical Tibetan documents based on the connected component analysis," in *Proc. Chin. Conf. Pattern Recognit. Comput. Vis. (PRCV)*, Guangzhou, China, 2018, pp. 74–87, doi: 10.1007/978-3-030-03338-5_7.
[22] C. Grana, L. Baraldi, and F. Bolelli, "Optimized connected components labeling with pixel prediction," in *Proc. Int. Conf. Adv. Concepts Intell. Vis. Syst.*, Lecce, Italy, 2016, pp. 431–440, doi: 10.1007/978-3-319-48680-2_38.
[23] J. G. Sun, "Clustering algorithms research," *J. Softw.*, vol. 19, no. 1, pp. 48–61, Jun. 2008, doi: 10.3724/SP.J.1001.2008.00048.
[24] G. Congedo, G. Dimauro, S. Impedovo, and G. Pirlo, "Segmentation of numeric strings," in *Proc. 3rd Int. Conf. Document Anal. Recognit.*, Montreal, QC, Canada, 1995, pp. 1038–1041, doi: 10.1109/ICDAR.1995.602080.
[25] Y. Q. Wei, J. S. Liu, and X. Z. Wang, "Concept of consistence and weights of the judgement matrix in the uncertain type of AHP," *Syst. Eng. Theory Pract.*, vol. 14, no. 4, pp. 16–22, 1994.
[26] Y. M. Wang, J. B. Yang, and D. L. Xu, "A preference aggregation method through the estimation of utility intervals," *Comput. Oper. Res.*, vol. 32, no. 8, pp. 2027–2049, Aug. 2005, doi: 10.1016/j.cor.2004.01.005.
[27] S. W. Chen and L. N. Cai, "Interval-valued hesitant fuzzy sets," *Fuzzy Syst. Math.*, vol. 27, no. 6, pp. 38–44, Jun. 2013.
[28] L. N. Cai, "Interval-valued hesitant fuzzy sets and its application to decision making," M.S. thesis, Dept Electron. Eng., Zhengzhou Univ., Zhengzhou, China, 2013.
[29] E. Z. Ni, M. J. Jiang, and C. L. Zhou, "Research on segmentation of historical Chinese books," *Comput. Eng. Appl.*, vol. 49, no. 2, pp. 29–33 and 38, Jan. 2013, doi: 10.3778/j.issn.1002-8331.1209-0246.

XUEDONG TIAN received the B.S. degree from the Department of Automation Engineering, Hebei University of Technology, China, in 1984, the M.S. degree from the Department of Electronics and Information Engineering, Hebei University, Baoding, China, in 1998, and the Ph.D. degree from the College of Physics Science and Technology, Hebei University, in 2007. He is currently a Professor with the School of Cyber Security and Computer, Hebei University. His research interests include information retrieval and pattern recognition.

TENGYING SUN was born in Baoding, Hebei, China, in 1995. She received the B.S. degree in Internet of Things engineering from the Baoji University of Arts and Sciences, Shanxi, China, in 2018. She is currently pursuing the master's degree with Hebei University, under the supervision of Prof. Tian. Her main research interests include intelligent image and text information retrieval.

YANMEI QI was born in Cangzhou, Hebei, China, in 1993. She received the B.S. degree in digital media technology from Shijiazhuang University, Hebei, in 2017. She is currently pursuing the master's degree with Hebei University, under the supervision of Prof. Tian. Her main research interests include intelligent image and text information retrieval.