Hindawi
Mobile Information Systems
Volume 2022, Article ID 1429635, 12 pages
https://doi.org/10.1155/2022/1429635

Research Article

Language and Literature Layout Integrating Text Image Preprocessing Algorithm

Lirong Wang,[1] Cui Fu ,[2] and Gaixia Fu[3]

[1]*College of Humanities and Education, Shaanxi Energy Institute, Shaanxi, Xianyang 712000, China*
[2]*School of Mathematics and Statistics, Xidian University, Xi'an 710071, China*
[3]*School of Hejiaying, Weiqu Street,Chang'an District, Xi'an 71099, China*

Correspondence should be addressed to Cui Fu; cuifu@stu.xidian.edu.cn

Received 19 December 2021; Accepted 20 January 2022; Published 21 February 2022

Academic Editor: Hye-jin Kim

The Chinese language and literature layout has important application value. How to quickly and efficiently deal with the huge amount, diverse forms, and complex content of Chinese language and literature is a problem that people pay more attention to. In addition, a large number of digital resources exist in the form of images instead of text encoding. How to efficiently manage and use this document information, especially the fast retrieval of document content, is an important research direction. This paper mainly discusses the processing of Chinese language and literature layout combined with text image preprocessing algorithm and researches into the skew correction of document images. There are many types of documents, and many tilt correction methods are based on prior knowledge. If such a sample quality situation occurs, the image needs to be preprocessed. Using denoising to remove useless noise information and using correction technology to leave and strengthen the classified effective information, we can carry out the next step of retrieval and recognition. The preprocessing avoids interference and destruction of the feature description algorithm by other factors as much as possible and guarantees the effect of recognition and retrieval to a certain extent. This paper proposes a method of document image tilt correction based on the content of the document. According to different document contents, we select the corresponding strategy to estimate the document tilt. This paper uses two-dimensional wavelet transform, runs length smoothing and thinning preprocessing to extract the lines and text lines in the document image, and uses the least square method of linear parameters to estimate the inclination angle of the document. This method has the characteristics of high accuracy of tilt estimation and strong adaptability. The accuracy of feature line matching in geometric structure classification can reach 97%. This research helps to promote the continuous development of Chinese language and literature.

1. Introduction

In order to extract the information contained in the Chinese language and literature layout, firstly, the related technology of text image processing is used for image processing, and the text recognition technology is used to extract information from numbers and text. This requires designing a reasonable algorithm to extract some special symbols of information.

Chinese language and literature layout recognition includes image processing, pattern recognition, statistical decision theory, and improvement of image automation and intelligence. It lays the foundation for the development of Chinese language and literature layout image processing system. Through the implementation of this research, the development of related technologies can be promoted to meet the actual needs of business units. It has very important theoretical and practical value.

If the text information is stored in the form of images, the amount of information will become very large, and the storage space required is larger than that of electronic text files, which is not conducive to the rapid recognition of text information. With the advancement of computer technology, these deficiencies will gradually be overcome. Wu et al. believe that FSC can obtain more correct matches than RANSAC in a smaller number of iterations. A large number of experimental studies have proved that the algorithm they proposed is robust [1]. Xu et al.

proposed a fast version of the Nonlocal Concentrated Sparse Representation (FNCSR) algorithm. They used prelearned dictionaries instead of adaptive dictionaries to build runtime [2]. In order to effectively extract the text area of the passenger car body, Zhao proposed an improved binarization algorithm based on the Bernsen algorithm for the body text features. The algorithm he proposed is performed on a sample of passenger cars. Compared with the classic local threshold binarization methods such as Bernsen, Niblack, and Sauvola, the subjective visual evaluation and objective image quality assessment (IQA) results show that the method he proposed performs better in the preservation and differentiation of character regions [3]. Fan J proposed a new point matching algorithm to align two SAR images. The proposed simulated deformation and real SAR images were used to evaluate performance [4]. Zhang et al. proposed a fully automatic superpixel generation algorithm by simplifying the 3D triangle mesh modeled from the 2D input image. The experimental results prove the effectiveness of the proposed method [5]. In order to overcome this shortcoming, Meng S solved the HSI denoising problem by combining Tucker decomposition and principal component analysis (PCA) [6]. Bawane et al. believe that there have been many successful applications of the first two generations of neural networks. For example, in today's world, people are more and more interested in active research in the field of neural networks. The use of a popular model of SNN to classify and recognize objects and various handwritten characters is described. Leak integration and trigger nonmaximum suppression are used to extract features from images [7]. With the advancement of science and technology, information processing technology and image processing technology are changing with each passing day; especially with the widespread use of computers and the increasing popularity of the Internet, people's communication methods have become diverse. However, image information processing, as one of the important ways of information transmission, still has a very broad market and huge application scale. Not only has its role not been replaced; it has been also strengthened in some areas of application.

The layout of Chinese language and literature needs to be preprocessed by tilt correction. This is because OCR technology is very sensitive to oblique and deformed text. In the processing of text and images, the text will be somewhat distorted. This is a serious interference to the later formation of regional positioning and may also affect the final recognition effect. The article first analyzes and summarizes the research background and significance of text image retrieval technology. This paper introduces the research status of various text recognition contents at home and abroad and the preprocessing technology of text images. In order to solve the segmentation problem in preprocessing and improve the scanning projection method of rows and columns, this paper adds a recognition and removal module of irrelevant information and discusses the redivision of the Chinese text of the table content.

2. Exploring Methods of Chinese Language and Literature Layout

2.1. Layout Information Processing System. With the rapid development of computer storage technology and the rapid progress of computer vision research, for some very important formal documents, the main storage mode is text images. Compared with text files, image files can express scenes more intuitively and truthfully, and it is not easy to tamper with and forge. For example, image files such as handwritten signatures and bank statements can also reflect the reliability and validity of the image files. The structure of the Chinese language and literature layout information processing system is shown in Figure 1. It can be seen that the system is mainly composed of three parts: document scanning and input, analysis and understanding of Chinese language and literature layout, and Chinese language and literature layout reconstruction.

2.1.1. Document Scanning and Input. Acquire the Chinese language and literature layout image and input it into the Chinese language and literature layout information processing system.

2.1.2. Layout Analysis and Interpretation of Chinese Language and Literature. The obtained Chinese language and literature layout image is divided into images, charts, tables, texts, etc., according to the characteristics of different regions, and the regional coordinates of the image and table regions are recorded.

2.1.3. Reconstruction of the Layout of Chinese Language and Literature. The processed areas are reconstructed into document formats such as RTF or XML according to the processed documents, so that the original visual layout can be reproduced next time [8].

To correctly extract the table lines and text in the table document image, the image must be binarized. On the one hand, binary images contain rich image information; on the other hand, nonbinary images contain too much redundant information, which is difficult to process directly. The feature dimension in binary images is much lower than that of gray or color images, which greatly simplifies text feature extraction.

2.2. Image Binarization. After scanning the document, the image usually needs to be binarized. Due to the influence of some factors during image acquisition or input, noise appears. In addition, when paper documents are converted into document images, there may be problems such as human errors and scanner failures. This causes the document image to have a certain inclination, which brings great difficulty to the recognition of the image. Because the noise and the tilt of the image bring inconvenience to the image recognition, the image must be preprocessed before the image is recognized and extracted. Minimize the

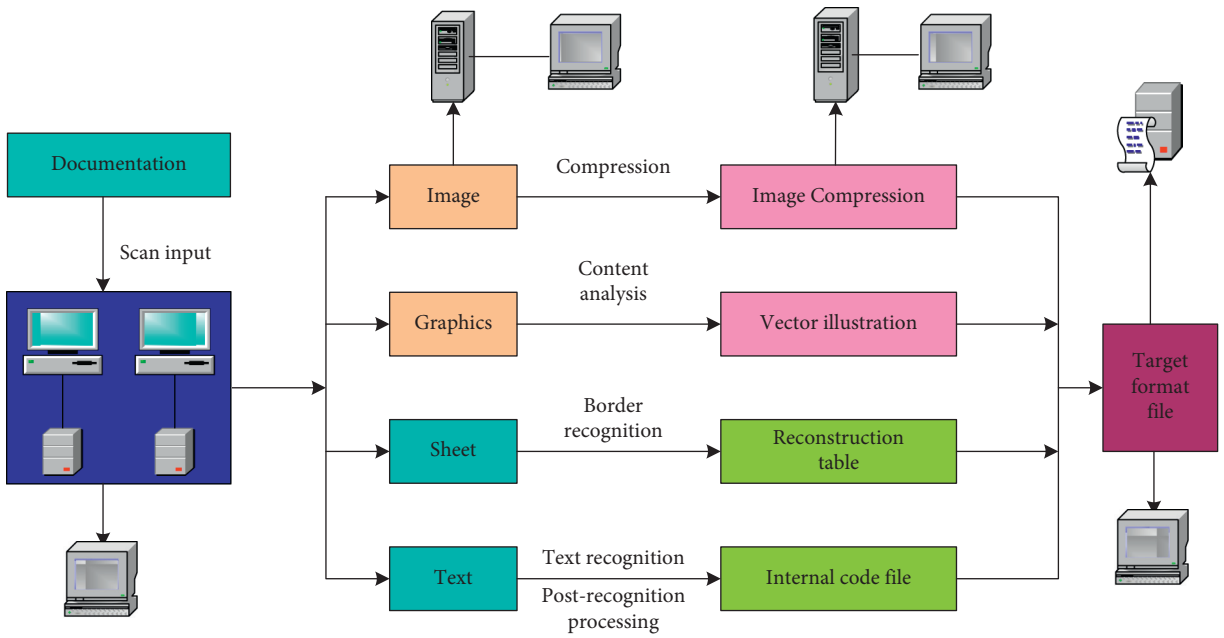


Figure 1: Structure of Chinese language and literature layout information processing system.

unfavorable factors for image recognition, and binarization is a very important technology in image preprocessing.

If the grayscale range of the image is $[M_1, M_K]$, the formula is [9]

$$F(i,j) = \begin{cases} 1, & F(i,j) \le M, \\ 0, & F(i,j) > M. \end{cases} \tag{1}$$

After the binarization process, the image will lose some grayscale information, but this does not change the original image much. The processed image can still fully express the information expressed by the original image. For the study of table recognition, the most important thing is the text and table lines. As long as the binarization process does not change the structure of text and table lines, so that the original image does not have breakpoints, it can proceed to the next step.

The maximum gray scale of the image is k_B, and then [10]

$$T_0 = \frac{k_A + k_B}{2}, \tag{2}$$

where k_A is the minimum gray scale.

$$k_c = \frac{\sum_{k(i,j)\le T} M(i,j) * N(i,j)}{N(i,j)}, \qquad k_d = \frac{\sum_{k(i,j)\le T} M(i,j) * N(i,j)}{M(i,j)}. \tag{3}$$

The average gray level of the image foreground is k_c. The average gray level of the background is k_d.

This paper proposes a document image binarization algorithm based on mathematical morphology. Because the background is the high-frequency part of the image, the open operation can be used to infer the background. Subtract the background from the original image to obtain a zero-background image and eliminate the influence of uneven background. Then, apply the Otsu (threshold) algorithm to the zero-background image to determine the threshold.

The ratio of the total number of pixels $f(x, y)$ belonging to the foreground to the total pixels is β_0, and the average gray value of the foreground pixels is μ_0, μ_0; then [11],

$$\omega_0 = \frac{n_0}{(M \times N)}, \qquad \omega_1 = \frac{n_1}{(M \times N)}, \tag{4}$$

where n_0 represents the total number of foreground pixels of the image and n_1 represents the total number of all background pixels.

The interclass variance G between the foreground and background of the figure is [12]

$$G = \omega_0 (\mu_0 - \mu)^2 + \omega_1 (\mu_1 - \mu)^2. \tag{5}$$

We can obtain the following:

$$G = \omega_0 \omega_1 (\mu_0 - \mu_1)^2. \tag{6}$$

So far, we have got the variance value between classes under this threshold. Then, through the traversal method, the threshold value corresponding to the maximum value of the variance between classes is found, which is the target threshold value.

2.3. Tilt Detection and Correction

2.3.1. Tilt Detection. This paper uses a document image tilt detection algorithm based on straight line fitting, which is described as follows:

(1) Text Line Positioning. In this article, the inclination angle detection will be performed based on the alignment of the curve of the text line with the line, so the curve of the text line will be found first. Before finding the text line, use mathematical morphology to inflate the image to prevent the text line curve from breaking.

(2) Straight Line Fitting. Taking into account the influence of random noise, this paper uses the least square method to fit straight lines [13].

According to the principle of least squares, the estimated value θ_g can be obtained:

$$\theta_{\mathrm{g}} = \left(x^T x\right)^{-1} x^T y. \tag{7}$$

Then, the inclination angle θ_Q of different fitted straight lines can be obtained:

$$\theta_Q = \tan^{-1} a_i. \tag{8}$$

2.3.2. Tilt Correction. After detecting the inclination angle, the inclined Chinese language and literature layout must be corrected to facilitate subsequent Chinese language and literature layout analysis and character recognition processing. The tilt correction is to rotate the Chinese language and literature page in the opposite direction according to the corresponding tilt angle. The process of obtaining the original Chinese language and literature layout is the reconstruction of the Chinese language and literature layout.

After the processing, the tilted document image can be corrected according to the detected tilt angle. The specific algorithm can be described as follows:

(1) Obtain the inclination angle θ of the document image according to the slope of the text line curve:

$$\theta = \frac{1}{N} \sum_{i=1}^{N} \theta_i. \tag{9}$$

(2) Rotate the image with the lower left point (origin) as the center of rotation. Set the coordinate system of the input image as the rectangular coordinate system XOY and the corrected coordinate system as the rectangular coordinate system $X'OY'$, $X'OY'$. We can get [14]

$$\begin{aligned} X &= X\cos\theta + X\sin\theta, \\ Y' &= Y\cos\theta - Y\sin\theta. \end{aligned} \tag{10}$$

(3) Calculate the extreme points of the vertical and horizontal coordinates of the rotated image. Among them, the calculation formula of the tilting process is [15]

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix}. \tag{11}$$

The calculation formula of the correction process is

$$\begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}. \tag{12}$$

As a result of the experiment, the image is rotated by this algorithm, and a relatively good effect is achieved. The original tilted image is corrected after being rotated. The specific rotated image is shown in Figure 2.

2.4. Chinese Language and Literature Layout Font Feature Extraction. Under normal circumstances, when extracting the features of Chinese characters, directly extract the features of the strokes of the Chinese characters, which can effectively eliminate the influence of the topological structure of the Chinese characters on the font recognition. Among them, wavelet transform is a kind of directional filter, which can decompose the strokes of Chinese characters, which facilitates the feature classification of the Chinese language and literature layout.

The time window of continuous wavelet is [16]

$$S(a,b) = [b + aE(\psi) - |a|\Delta(\psi), |a|\Delta(\psi)]. \tag{13}$$

Here, $E(\psi)$ is the center and $\Delta(\psi)$ is the width of the window.

The frequency window is

$$P(a,b) = [b - aE(\psi), b + |a|\Delta(\psi)]. \tag{14}$$

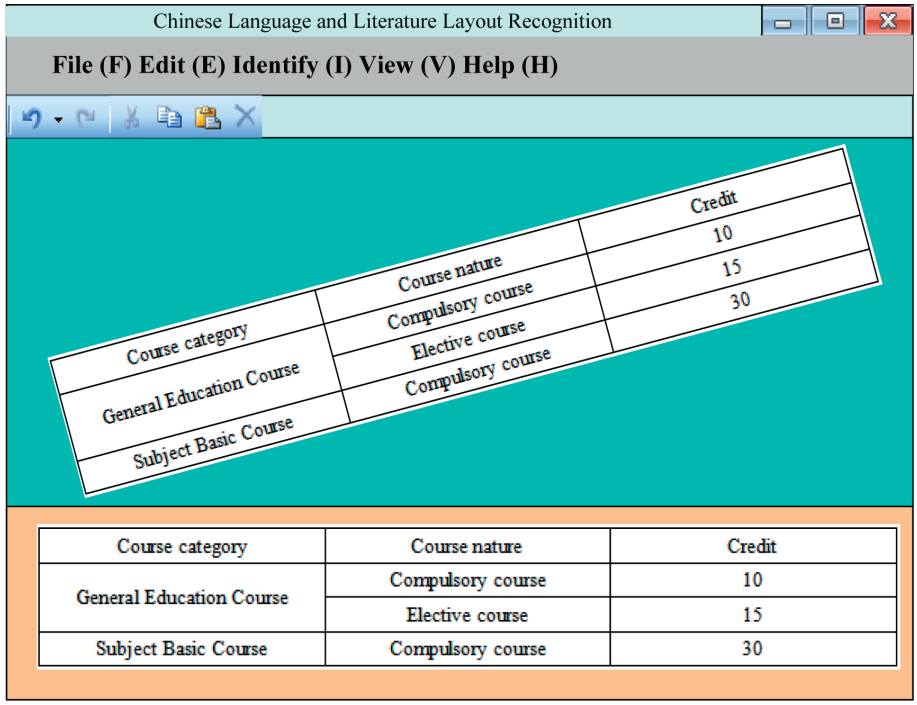
Its time-frequency window is a variable rectangle [17]:

$$T[a,b] = [b - aE(\psi), b + |a|\Delta(\psi)] \times \left[\frac{b - aE(\psi)}{a}, \frac{b + |a|\Delta(\psi)}{b}\right]. \tag{15}$$

In the $L^2(R)$ space, the scaling function and wavelet function constitute an orthonormal system, and two spatial columns are formed based on them [18]:

$$\begin{aligned} W &= C\{2\psi(2-n); n \in Z\}, \\ G &= C\{2\beta(2-n); n \in Z\}. \end{aligned} \tag{16}$$

2.5. Learning the Geometric Structure of the Chinese Language and Literature Table Layout. The study of the geometric structure of the Chinese language and literature form layout is to obtain prior knowledge such as the position of the form fill-in information in the layout. To realize the automatic learning of the Chinese language and literature table layout structure, it is necessary to recognize the handwritten and printed text, and then the computer automatically determines the position of the handwritten text in the Chinese language and literature layout to complete the automatic learning of the table. Finally, after manually verifying that the learning result is correct, the prior knowledge of the learned form is saved in the form knowledge base. In the form learning method based on handwriting and print recognition, the text block in the form needs to be extracted first. The horizontal and vertical dividing lines in the table are not conducive to the extraction of strings, so the straight line removal module is



Course category	Course nature	Credit
General Education Course	Compulsory course	10
	Elective course	15
Subject Basic Course	Compulsory course	30

Figure 2: Specific rotated image.

used to detect straight lines and remove them. The recognition process of the Chinese language and literature table layout is shown in Figure 3.

2.6. Chinese Layout Analysis Algorithm Based on Hierarchical Extraction. The main development environment is as follows: computer hardware: Lenovo X1 Carbon with Intel core i7-10710U, memory DDR5; operating system: Windows 10 Standard Professional; primary development environment: Microsoft Visual Studio 2015; database environment: SQL Server 2014; web browsers: the Chrome browser is mainly used.

Text recognition mainly includes three parts: collecting text information, analyzing and processing the collected information, and judging and classifying information according to the analysis results. In a lot of research in the field of text recognition, it can be roughly divided into handwritten text recognition and printed text recognition. For the general Chinese language and literature layout, there is still a lack of an effective Chinese language and literature layout segmentation and region recognition method. For this reason, this paper proposes a Chinese layout analysis method based on hierarchical extraction. It divides the types of Chinese language and literature layout objects into levels and extracts the objects according to levels, transforming complex issues into single issues. Furthermore, for the text part, through the search of connected regions, the bottom-up processing is performed to ensure the correctness of the segmentation. In view of most Chinese language and literature layout analysis methods, the Chinese language and literature layout segmentation is completely separated from the region recognition process, so that the region recognition as the subsequent processing of the Chinese language and literature layout segmentation requires repeated data extraction. This increases the processing time of the system. This article regards region recognition as a supplement to the Chinese language and literature layout segmentation and makes full use of the data in the layout segmentation process. In this paper, different objects are extracted at different levels, and the Chinese language and literature layout segmentation is closely combined with region recognition, thereby improving the efficiency of the algorithm. The effect of Chinese language and literature layout processing is shown in Figure 4.

3. Results of Analysis of Chinese Language and Literature Layout

We have selected 20 representative Chinese language and literature layouts for experiments. Each image is rotated by −40° to 40° counterclockwise with an inclination angle of 5° as the step length, each Chinese language and literature layout forms 17 images, all 340 images are subjected to Chinese language and literature layout tilt detection statistics, and the average value is calculated (obtaining the tilt angle and the detection angle as shown in Figure 5(a)). It can be seen that the advantage of this algorithm is high accuracy. The average error of this algorithm is −0.14, and the average accuracy rate is 99.3% (the error rate and accuracy rate obtained are shown in Figure 5(b)).

The selected layout types include Chinese layouts, English layouts, and mixed Chinese and English layouts, and

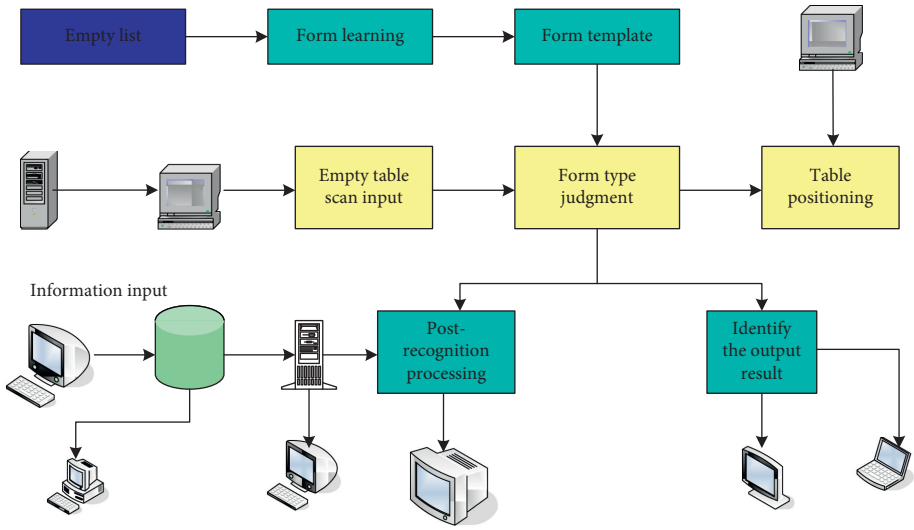


Figure 3: Recognition process of form layout.

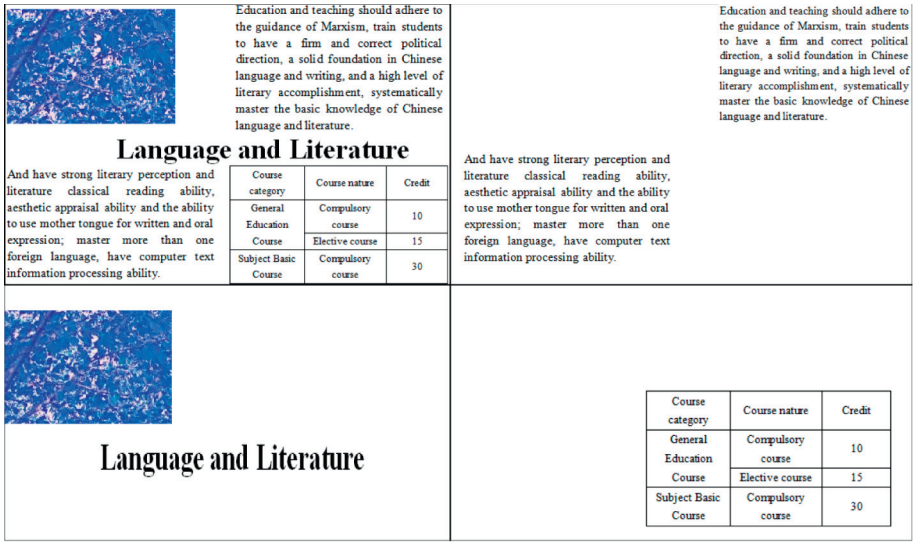


Figure 4: Layout processing effect of Chinese language and literature.

the detection results of the algorithm are not affected by its content, which shows that this algorithm has nothing to do with the content of the layout. The results of page content detection are shown in Figure 6.

The layout analysis algorithm of hierarchical extraction is based on the MFC environment and implemented with Visual C++ 6.0. The 110 layouts in the sample library are segmented, and area recognition experiments are carried out. The experimental result test is carried out by matching the template generated by the algorithm with the standard template. The average processing time of each image is less than 1 second, and the error statistical rate of the algorithm is shown in Table 1.

The layout analysis method based on hierarchical extraction has an average accuracy of 91.3%, and it has the characteristics of short analysis time and high computational efficiency. The results of this research level extraction and recognition are shown in Table 2.

The judgment of the logical structure depends on the results of the layout analysis of Chinese language and literature. For nonembedded layout, due to the high accuracy of Chinese language and literature layout analysis, the accuracy of sequence judgment is also high. The segmentation effect of embedded Chinese language and literature layout is slightly worse than that of nonembedded layout, so the judgment also produces errors. In addition, the determination of the nature of the page object also has a great influence on the logical judgment. The results of the order of Chinese language and literature layout are shown in Table 3.

In order to make the detection effect cover the real situation as much as possible, 100 samples are selected in this experiment. It contains 202 tables (the two connected tables and the identification of the special table are shown in Figure 7(a)). The resolution of the image is 110 dpi (there are black areas and fuzzy recognition of the table lines in the table, and the result is shown in Figure 7(b)).

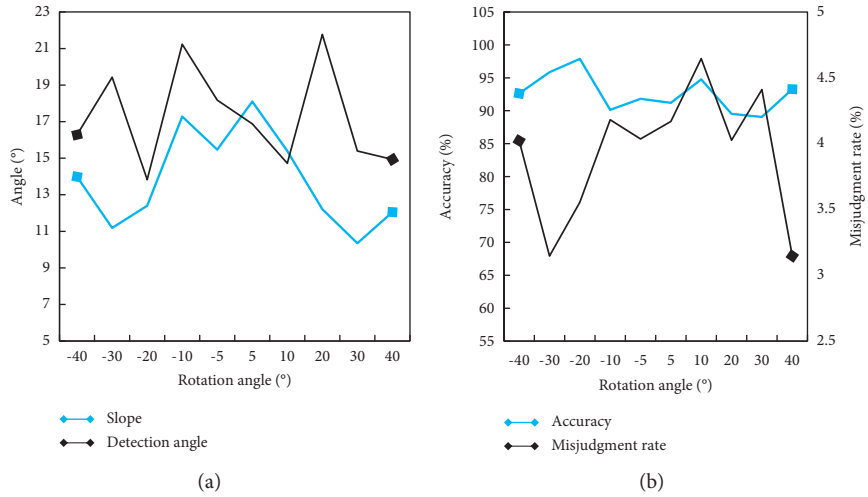


Figure 5: Inspection results after different counterclockwise rotations. (a) Obtaining the tilt angle and the detection angle. (b) Getting the rate of misjudgment and accuracy.

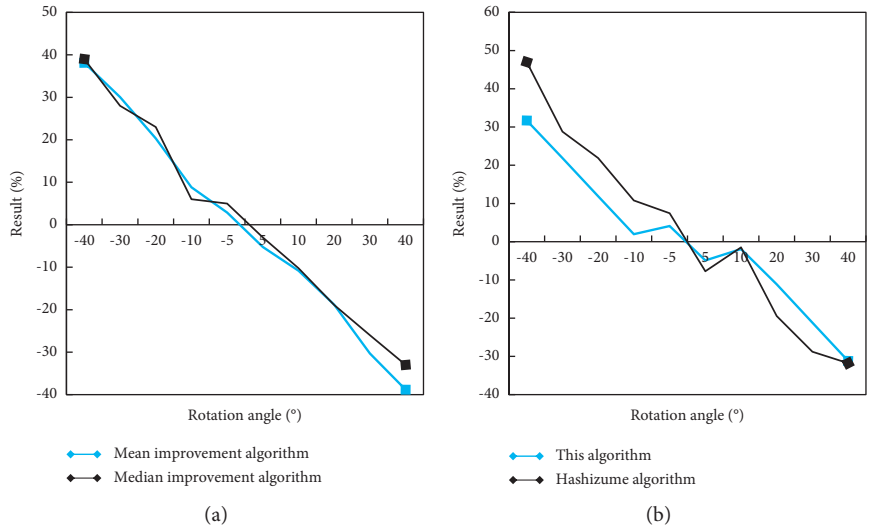


Figure 6: Layout content detection results. (a) Mean value improvement algorithm and median value improvement algorithm. (b) The algorithm in this paper and the Hash algorithm.

Table 1: Error statistical rate of the algorithm.

Error type	Error rate (%)	Total error rate (%)
Missing	0.86	22.24
Noise	0.24	2.84
Splitting	0.22	2.68
Merging	4.84	64.40
Spuriousness	2.61	22.26

We have selected two test sets in turn: test set I contains 50 pictures of Chinese language and literature, and test set II contains 100 pictures of Chinese language and literature. The accuracy of feature line matching in geometric structure classification can reach 97% (feature line matching and related feature line matching are shown in Figure 8(a)). The color matching recognition in the logical structure classification can reach 98% (color matching and key attribute matching are shown in Figure 8(b)).

Different magazines are selected as the experimental samples, and the experimental results of the sample tests are shown in Table 4.

Table 2: Hierarchical extraction and recognition results of this research.

Content	Region segmentation	Text area	Image area	Graphics
Number of regions	1940	1540	320	90
Identifying the correct number	1800	1460	301	75
Correct rate	92.8%	94.8%	94.1%	83.3

Table 3: Results of the order of the layout of Chinese language and literature.

Content	Nonembedded rectangular layout	Embedded rectangular layout	Other
Number of pages	55	30	15
Correct number	54	28	11
Correct rate	98%	93.33%	73.33%

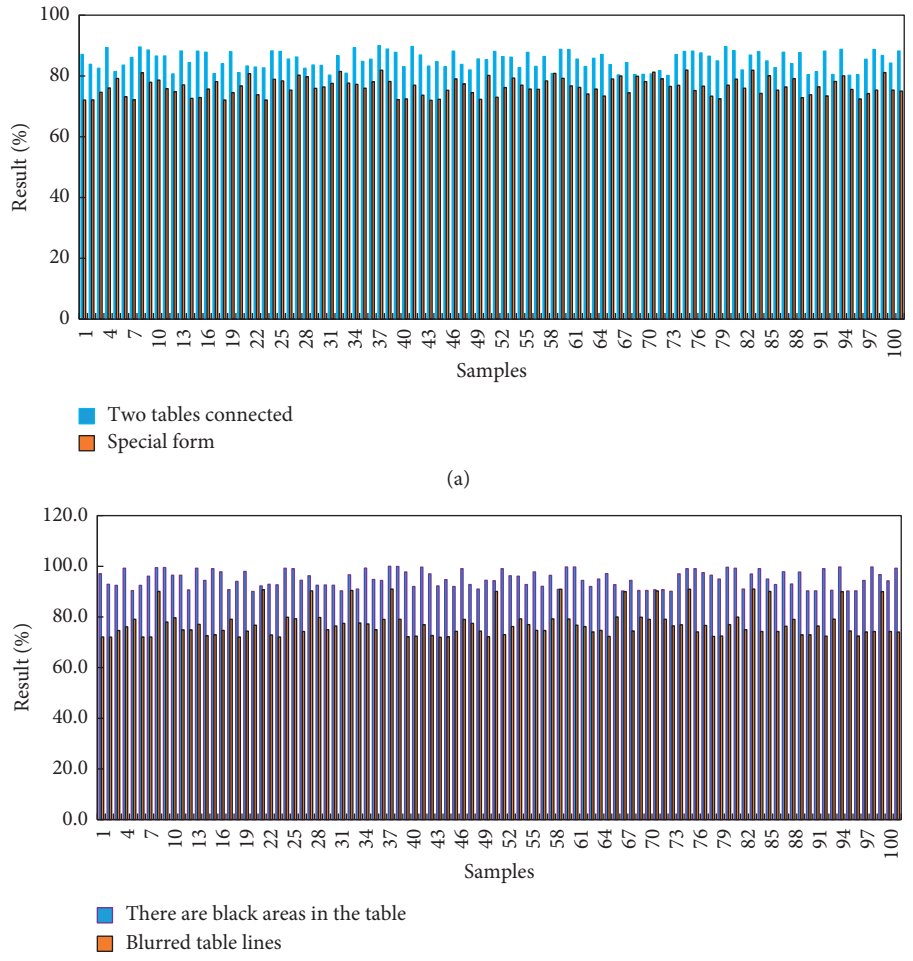


Figure 7: 100 sample test results. (a) The connection of two tables and the identification of special tables. (b) There are black areas in the table and fuzzy recognition of table lines.

Compared with the traditional bottom-up algorithm and top-down algorithm, the results of this experiment on the network side are shown in Table 5.

This article randomly selects 1000 Chinese characters from 6000 Chinese characters in the dictionary database as Chinese characters to be recognized. Through experimental analysis, we compared the effects of these two methods on the recognition accuracy and recognition speed of the algorithm. The experimental results are shown in Table 6.

Using the 50 oblique document images for experimentation, the method in this paper is compared with the experimental results of the Block Adjacency Graph algorithm. It can

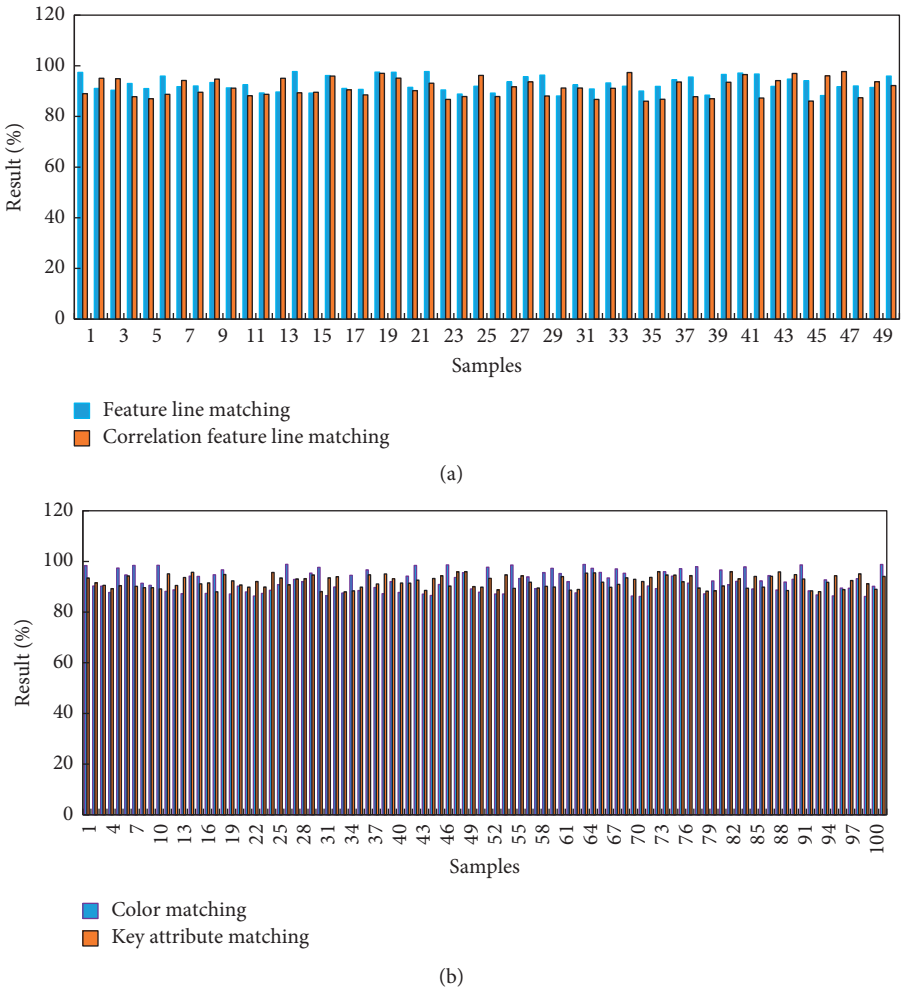


Figure 8: Geometric structure classification and logical structure classification results. (a) Feature line matching and related feature line matching. (b) Color matching and key attribute matching.

Table 4: Experimental results of the sample test.

Research content	Chinese books	Chinese magazines	Chinese newspapers
Experiment pages	314	261	64
Number of positive characters	127211	271211	176711
Number of text lines	6162	9611	6441
Correct rate of positive text merging	97.9%	99%	96%
The correct rate of text line merging	97.4%	97.2%	96.4%
Experiment page merging accuracy rate	97%	96.6%	95%

Table 5: Experimental results on the network side.

Layout segmentation algorithm	Accuracy (%)	Average time (ms)
Connected domain segmentation algorithm based on fusion contour projection	92	46.6
Traditional top-down algorithm	82	68.3
Traditional bottom-up algorithm	82	63.2

be seen from the research results that the average error of the content-based document image tilt estimation method proposed in this paper is very small and that the ratio of the tilt estimation error is less than 0.1, which is much higher than that of the subspace straight line detection. The comparison results of different algorithms are shown in Figure 9.

Table 6: Arbitrary selection of 1000 Chinese characters as the test results of Chinese characters to be recognized.

Normalization method	Recognition accuracy (%)	Recognition speed (words/s)
36×36 dot matrix	95.40	54
48×48 dot matrix	95.60	88
60×60 dot matrix	94.3	92

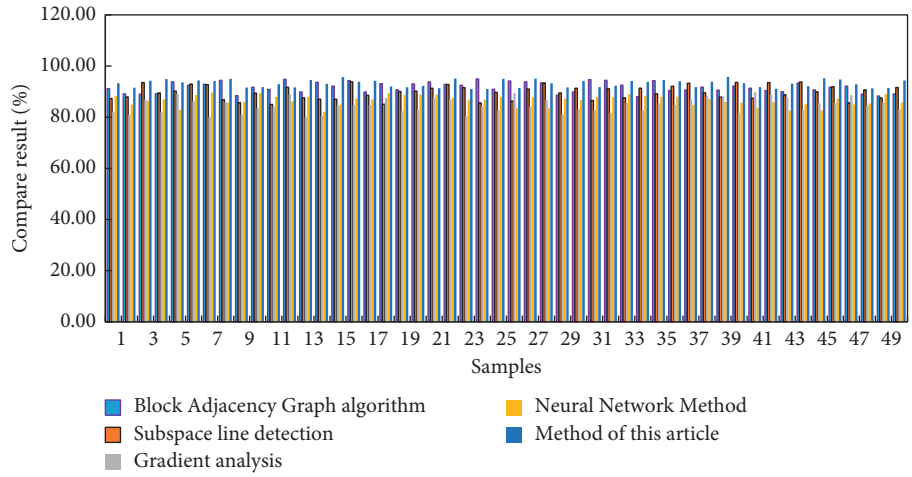


Figure 9: Research comparison results of different algorithms.

4. Discussion

In pattern recognition, there are many patterns that cannot be completed only by the recognition ability of the computer itself. Because pattern classification needs to play the role of people, it is better to learn when there is a teacher than when there is no teacher. Using human guidance on macro pattern recognition, computers are good at quantitative information, humans are good at qualitative information processing, and we give full play to their respective advantages. In the training process of teachers and learning, the role of people is to implement macro control, so that the network can learn consciously and maximize its possibilities. For the Chinese language and literature major, most of the scholars who conduct research are specialized in Chinese and literature. They have accumulated a lot of Chinese knowledge in Chinese research, have learned systematic Chinese, and have excellent language analysis and performance skills. However, the Chinese language and literature major does not focus on the practice of Chinese education as a second language, so students with a major background in Chinese language and literature are not proficient in second-knowledge language education and lack educational experience. They lack knowledge and ability to communicate between different cultures.

In recent years, the development of the "Chinese fever" and the strategic deployment of the "Belt and Road" have brought closer exchanges between countries. The demand for learning Chinese is increasing, and the number of learners is increasing. The Chinese language and literature page contains text, images, graphics, tables, and other types of areas. This is because the processing methods are different in various fields such as the text area for character recognition, the form reconstruction of the form area, and the analysis of the line chart area. In order to execute according to the area type of the subsequent processing, various attributes of the spatial area need to be classified. The method of texture analysis is to use texture analysis technology in image analysis. It ignores the specificity and contextual knowledge of various types in the image layout and directly analyzes the layout, for example, using texture analysis technology to directly segment and classify grayscale images, analyzing the texture through neural network, and selecting the characteristics of the structural elements of the layout. The disadvantage of this method is that the calculation is large and the accuracy is not high.

The major of Chinese language and literature has a long history and profound historical accumulation. Since the end of the last century, the major of Chinese language and literature has developed. Because the Chinese language and literature major is the cultural foundation of China, universities are generally named as the Chinese language and literature major. At the same time, teacher training schools are indispensable for the professional basic knowledge training of junior high school teachers. After years of construction and development, the discipline theory of this major is relatively mature, and the course is basically completed.

At present, Chinese language and literature are divided into two main categories in general universities. Although the focus of cultivation is different, part of the content of the curriculum is basically the same. In other words, the language courses and introduction of linguistics in modern China and ancient China reflect the differences in the

knowledge of Chinese (philosophy, ancient literature, modern literature, foreign literature, comparative literature, etc.). Of course, in order to cultivate students who will become teachers of vocational training and educational ability, educational courses such as Chinese education courses and educational theories have been opened. With the rapid development of the Internet, more and more information will be presented in different forms; in addition to the form of text, other forms mostly include pictures. However, in the pictures, there are a lot of mixed pictures and texts. For this kind of mixed image and text, on the one hand, the information they store is different and may be duplicated. On the other hand, its traditional extraction techniques for pictures and text are different. Therefore, the distinction between graphics and text can effectively improve the extraction rate and accuracy of mixed graphics and text images [19].

With the development and popularization of the Internet and mobile smart terminals, the requirements for media digitization are getting higher and higher. On the one hand, the mobile Internet has given birth to various new media, and social media occupy many media resources. On the other hand, the direct digitization of traditional media has now become a major trend. For example, Southern Weekend and Beijing News have established their own electronic print media and digital paper print media on the official website and APPS. Based on the purpose of effectively dividing the Chinese language and literature layout, this article effectively combines the bottom-up segmentation method with the top-down segmentation method. First of all, because the gray value of Chinese characters is almost the same as the background color, the algorithm expands the area of a word so that the gray value of Chinese characters will not change greatly. Then, it performs contour projection and finally merges uniform connected regions. In other words, because adjacent text is merged into a wider area, images and text can be effectively segmented.

Chinese language and literature layout and classification methods have their own advantages and disadvantages. At present, there is no Chinese language and literature layout analysis method suitable for various Chinese language and literature layout conditions. In addition, in the previous analysis methods of Chinese language and literature layout, the division and the classification of Chinese language and literature layout are independent. First, the layout division of Chinese language and literature is performed, and then the features of each field are extracted to classify the layout of Chinese language and literature. Because division and classification need to be calculated separately, a lot of calculations are required. However, no matter which text recognition method is adopted, the important thing is the selection of features and the corresponding classification algorithm. Because the existing manual processing of network transmission images is time-consuming, the use of image processing technology and user information for research and extraction technology, and the use of user information for classification and retrieval of fax images have improved the automation level of image processing [20].

5. Conclusion

This paper preprocesses the Chinese language and literature layout image (image gray scale, binarization, image noise removal, layout correction) and performs Chinese language and literature layout segmentation on the image. This paper uses the theory and technology of digital image processing and pattern recognition to preprocess, analyze, and recognize Chinese language and literature layout images. Through user information extraction technology, the characteristics of the main objects of this article have been deeply explored. It proposes a layout positioning method based on the best coordinate system. It also uses stable background information distributed in the layout as positioning marks, which are composed of thousands of positioning coordinates. In addition, it divides a layout into several small areas, and each coordinate system places points in nearby small areas. This can effectively eliminate the influence of image distortion in the layout. The character recognition accuracy of high-quality images becomes higher, but the character recognition accuracy of the distorted and noisy Chinese language and literature layout is not high. It is necessary to further improve the character recognition accuracy of low-quality images. For the distinction and recognition of charts in documents, especially tables, and some complex multilevel formulas, the article has not yet carried out special research, mainly for images, titles, and texts, so some other special content needs to continue to be studied later.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] Y. Wu, W. Ma, M. Gong, L. Su, and L. Jiao, "A novel point-matching algorithm based on fast sample consensus for image registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 12, no. 1, pp. 43–47, 2017.

[2] S. Xu, X. Yang, and S. Jiang, "A fast nonlocally centralized sparse representation algorithm for image denoising," *Signal Processing*, vol. 131, pp. 99–112, 2017.

[3] Y. Zhao, "An improved binarization method for passenger car limited-load character area," *International Core Journal of Engineering*, vol. 5, no. 10, pp. 188–196, 2019.

[4] J. Fan, Y. Wu, F. Wang, P. Zhang, and M. Li, "New point matching algorithm using sparse representation of image patch feature for SAR image registration," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 55, no. 3, pp. 1498–1510, 2017.

[5] Y. Zhang, L. Ma, Y. Zhou, and C. Zhang, "Automatic superpixel generation algorithm based on a quadric error metric in 3D space," *Signal, Image and Video Processing*, vol. 11, no. 3, pp. 471–478, 2017.

[6] S. Meng, L. T. Huang, and W. Q. Wang, "Tensor decomposition and PCA jointed algorithm for hyperspectral image

denoising," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 7, pp. 897–901, 2017.

[7] P. Bawane, S. Gadariye, S. Chaturvedi, and A. A. Khurshid, "Object and character recognition using spiking neural network," *Materials Today Proceedings*, vol. 5, no. 1, pp. 360–366, 2018.

[8] K. H. Kim, S. Lee, J. B. Shim et al., "A text-based data mining and toxicity prediction modeling system for a clinical decision support in radiation oncology: a preliminary study," *Journal of the Korean Physical Society*, vol. 71, no. 4, pp. 231–237, 2017.

[9] Y. Wu, Q. Miao, W. Ma, M. Gong, and S. Wang, "PSOSAC: particle swarm optimization sample consensus algorithm for remote sensing image registration," *IEEE Geoscience and Remote Sensing Letters*, vol. 15, no. 2, pp. 242–246, 2018.

[10] X. Chai, J. Fu, J. Zhang, D. Han, and Z. Gan, "Exploiting preprocessing-permutation-diffusion strategy for secure image cipher based on 3D Latin cube and memristive hyperchaotic system," *Neural Computing & Applications*, vol. 33, no. 16, pp. 10371–10402, 2021.

[11] M. Annabestani and M. Saadatmand-Tarzjan, "A new threshold selection method based on fuzzy expert systems for separating text from the background of document images," *Iranian Journal of Science and Technology, Transactions of Electrical Engineering*, vol. 43, no. JUL.SUPPL.1, pp. S219–S231, 2019.

[12] S. Alam and N. Yao, "The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis," *Computational & Mathematical Organization Theory*, vol. 25, no. 6, pp. 319–335, 2019.

[13] J. Wang, X. Zhi, X. Chai, and Y. Lu, "Chaos-based image encryption strategy based on random number embedding and DNA-level self-adaptive permutation and diffusion," *Multimedia Tools and Applications*, vol. 80, no. 10, pp. 16087–16122, 2021.

[14] J. Li, L. Gao, K. Xie et al., "Detection of functional homotopy in traumatic axonal injury," *European Radiology*, vol. 27, no. 1, pp. 325–335, 2017.

[15] G. Zhang and X. Yue, "The character of online visual recognition technology research," *Optical Technique*, vol. 44, no. 1, pp. 75–81, 2018.

[16] J. Huang, Z. Zhou, J. Shang, and C. Niu, "Heterogeneous domain adaptation with label and structural consistency," *Multimedia Tools and Applications*, vol. 79, no. 25, pp. 17923–17943, 2020.

[17] P. Pan and C. Patel, "The influence of native versus foreign language on Chinese subjects' aggressive financial reporting judgments," *Journal of Business Ethics*, vol. 150, no. 3, 2018.

[18] Y.-j. Lim, "Qualitative study on problems and development direction of activity-type Chinese language education using tandem learning method," *The Journal of Chinese Language and Literature*, vol. 106, no. 106, pp. 145–167, 2017.

[19] A. Ahmed, J. Page, and J. Olsen, "Enhancing Six Sigma methodology using simulation techniques," *International journal of lean six sigma*, vol. 11, no. 1, pp. 211–232, 2020.

[20] K. Bahmani, A. Akbari, and A. I. Darbandi, "Light and temperature requirements for germination assessment of various growth types of Iranian bitter fennels at different light and temperature regimes," *International Journal of Prosthodontics*, vol. 24, no. 2, pp. 109–117, 2018.