TELKOMNIKA Indonesian Journal of Electrical Engineering
Vol.12, No.6, June 2014, pp. 4579 ~ 4588
DOI: 10.11591/telkomnika.v12i6.5440

A Preprocessing and Analyzing Method of Images in PDF Documents for Mathematical Expression Retrieval

Xuedong Tian*, Botao Yu, Jing Sun
College of Mathematics and Computer, Hebei University, Baoding, Hebei, China
*Corresponding author, e-mail: txdinfo@yahoo.com

Abstract

PDF documents are the important information resources for a mathematical expression retrieval system. As a major component of PDF documents, the image objects must be converted to coded form with the help of character recognition and document analysis technology firstly for content based searching. Therefore, the quality of these images becomes the key factor which decides the correctness in this conversion process. Considering the characteristics of PDF images and mathematical expressions, a preprocessing and analyzing method was proposed which includes the modules of PDF image extraction, graying, binarization, denoising, skew correction and layout parameter detection. The features of mathematical expressions were adequately considered to avoid the information loss in image converting process and the adverse interference both to the analysis and correction process resulted from formulas. The experimental results show that the method is effective in improving the accuracy and efficiency of document image recognition, analysis and retrieval.

Keywords*: PDF image preprocessing analysis resolution*

1. Introduction

Nowadays, more and more PDF documents [1] are widely used in various fields, such as information storage, transmission, exchange, and so on. How to use PDF documents efficiently and conveniently becomes a hot point in the field of document analysis and process. Although a variety of techniques about PDF document analyzing, processing and applying, such as content editing and information extraction, have been developed and applied, there exist many unsolved problems in related topics because of the complexity and diversity of PDF structure.

In a PDF document, information exists in many forms including codes and font boxes of characters, images and graphs etc, which indicates the clues of extracting information from PDF documents. Totally, the methods of extracting information from PDF documents could be divided into three categories [2-5]. The first kinds of methods are only based on the contents including the codes and font boxes of PDF files. Through extracting the character codes, character fonts and their boxes, the content of PDF documents could be obtained. The limitation of this strategy lies in that the characters' geometric information contained in PDF documents is gross rather than exact coordinates of each character's bounding box, which results in the difficulties of the information extraction of PDF documents based on the geometrical features. The second strategy is to obtain information from the corresponding images converted from PDF documents. The effects of these methods depend on the correctness of layout analysis and character recognition with which the contents and their relationships could be known. Undoubtedly, the generally accepted strategies of PDF information extraction are employing a hybrid method combining both the character content features and the corresponding image features to ensure the accuracy and efficiency of obtained information. Baker et al. [2] proposed a method of extracting mathematical expressions from PDF documents. The characteristic of their method is to integrate the content feature of characters and the geometrical feature of corresponding images of the layouts in the process of locating the mathematical symbols in documents. The content of PDF documents including the font bounding boxes and the geometric positions of characters is extracted from the PDF source firstly. Then the characters' actual bounding boxes are obtained based on the images of the corresponding font boxes. The method could obtain both the content information and the geometrical information of PDF

Received December 28, 2013; Revised February 27, 2014; Accepted March 12, 2014

documents. In literature [3], Baker et al. further discussed the topic of extracting and analysis of mathematical formulas in PDF documents with the help of syntax features. Firstly, the bounding boxes (PDFBBs) information of PDF documents are obtained through decompressing the PDF documents with the open source Java software called multivalent and the true glyph bounding boxes (GBBs) are located with the help of corresponding images. Then, Anderson's algorithm is employed and improved to turn the two dimensional mathematical formula into linear representation. Finally, the linear expression is analyzed and the result is represented as a formula syntax tree with which the formula that is consistent with the original expression in Latex format could be easily generated. Lin et al. [4] proposed a method of identifying mathematical expressions in PDF documents. The features used in formula extraction are divided into three layers called geometric layout, character and context content. In the first step, which is called preprocessing, mathematical expression elements are obtained from the original PDF symbols through a series of processing based on the knowledge of PDF symbols and the text lines are detected. Then, the isolated formulas are extracted with the help of a hybrid method of combining rule method and SVM classifier based on geometric, character and context features. Finally, the embedded formulas are located with a rule-based algorithm and character features. In literature [5], the extraction of embedded formulas from PDF documents is further discussed. The method segments text lines into words. Different from the literature [4], these words are then divided into two classes called formulas or texts with an SVM classifier instead of previous rule-based method. The words identified as the components of formulas are merged into complete formulas.

In a mathematical retrieval system, the information of PDF documents, whether existing in coded state or in image form, should be treated as content flows. So, the image objects contained in PDF documents (simply called PDF images) must be recognized and analyzed with OCR (Optical Character Recognition) technology firstly, which lays a foundation of the following indexing and retrieval process. Nevertheless, the processing objects of ordinary OCR systems are generally the document images with high quality which are generated by scanners with a default resolution such as 300dpi in character recognition or 600dpi in mathematical formula recognition. When the OCR systems are used to recognize PDF images, the recognition rate would decrease because the quality, for example, the resolutions of these images, is varied and uncertain. Especially, the correctness of a formula recognition and analysis system will be influenced by the variation in image resolution more seriously. This is because formulas may contain some special symbols which consist of very small or little strokes, and formulas frequently express calculating meaning with the spatial arrangements of symbols implicitly. The characteristics make the performance of a formula recognition and analysis system to be sensitive to the quality of document images. Therefore, it is necessary to process the images to a higher quality and analyze them to obtain some parameters for guiding the following modules before recognition.

Although all of the methods and algorithms of document image processing, analysis and recognition could be employed for treating the images converted from PDF documents theoretically, many unsolved problems exist in every step of a PDF image processing system because of the characteristics of PDF images and mathematical expressions.

In the aspect of image extraction from PDF documents, Chen et al. [6] proposed a method to extract the images in PDF documents. On the basis of the introduction of PDF format, a scheme of obtaining the image data and decoding them into normal data is designed. The obtained images are saved as jpg files. Wang [7] designed an extracting algorithm of the images in PDF documents. This method specially considered the requirements on the image format of OCR systems. Li and Liu [8] put forward a PDF reader with a strategy of positioning key information and ignoring secondary message. Experimental result shows the proposed method could extract and display the information in PDF files accurately.

In the field of document analysis and recognition, many research works have been done in document image preprocessing and analyzing [9-20].

The methods mentioned above lay the foundation of our work. In this paper, a preprocessing and analyzing method of PDF images is designed to improve the quality of PDF images and the performance of the following recognition and retrieval process. It includes the modules of PDF image extraction, graying, binarization, denoising, skew correction and layout parameter detection that consider the characteristics of PDF images. In the process, the features of mathematical expressions are adequately considered to avoid the information loss in

image converting process and the adverse interference both to the analysis and correction process resulted from formulas. The experimental results show that the method is effective in improving the accuracy and efficiency of document image recognition, analysis and retrieval.

The organization of the paper is as follows. In the second part, the procedure of the PDF image preprocessing and analyzing method is given. The research and design of modules of the system is discussed in the third part. The last part is the results of experiments and the conclusions of the entire paper.

2. Description of the Preprocessing and Analyzing Method of PDF Images

The images are extracted from PDF documents through a PDF document analysis module firstly. The quality of them is irregularity. Many degraded phenomenons such as noise, skew and low resolution exist in extracted images because they come from different collecting ways. Therefore, these images need to be preprocessed and analyzed before recognition and retrieval. The structure diagram of the process is shown in Figure 1.

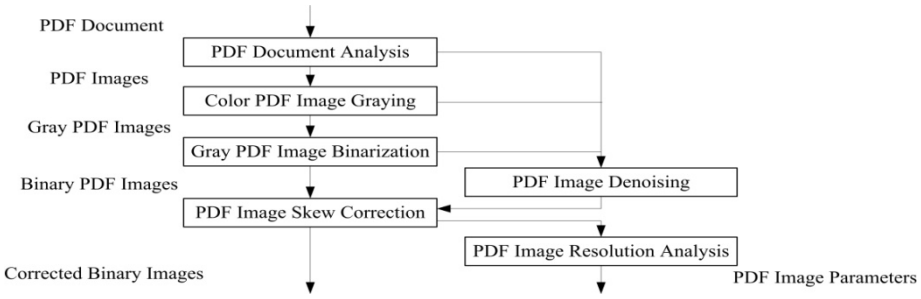

Figure 1. Structure Diagram of PDF Image Preprocessing

The images in PDF documents might exist in various modes such as colour images, gray images and binary images. Considering the efficiency of recognition, analysis and retrieval, all extracted PDF images are transformed into binary images.

PDF images might come from different sources. Many of them contain a lot of noise pixels. If these noise pixels are wrongly considered as normal image pixels, the spatial relationships of characters would be disrupted and error results will be produced. These situations occur more frequently in symbol recognition and structural analysis modules of a mathematical formula recognition system. Because formula symbols come from several character sets, there exist more similar symbols which are hardly distinct from each other than normal text. When noise pixels lie in a key area, the error results are inevitable. Therefore, it is necessary to design an algorithm to delete the noise pixels from PDF images and avoid the normal pixels being wrongly erased with the help of the relative spatial, syntax and semantic knowledge.

When a document image mainly consists of text contents, the skew of it will be very harmful because the aberrance of the distribution features of characters on the layout might occur, which results in the error result of the logical structures. Therefore, with the help of layout knowledge, we could detect the skew angle of a document image and rotate the image a corresponding angle to obtain a corrected image.

Different from the images in OCR system, the quality of the images in PDF documents varies in a large extent. Especially, the resolutions of them are not unified to a default value because of the difference of scanning operations, which will influence the correctness of the whole system. Although various kinds of interpolation algorithms in digital image processing could improve the visual effect of the images, the recognition rate could not be increased essentially. Therefore, it is necessary to measure the resolution value of images roughly to direct the following steps of document recognition and analysis to adjust the related parameters to fit the variations. This could be considered as a special segment of mathematical formula recognition used in mathematical expression retrieval.

3. Implement of the PDF Image Preprocessing and Analyzing Method

Aiming at the need of mathematical expression recognition and retrieval, a solution of PDF image preprocessing and analyzing is designed considering the characteristics of the images in PDF documents and the features of formulas. It need pay attention to that the preprocessing algorithm should take targeted measures to process PDF images according to the actual states of them for meeting the needs of recognition and retrieval rather than apply all modules to the images aimlessly.

3.1. Extraction of Images from PDF Documents

The images are saved as objects in PDF documents. They should be extracted firstly according to the PDF document specification [1], [6-8]:

Step 1. Obtain the offset of the cross-reference table and the object number of Catalog at the trailer of PDF file.

Step 2. Get the content of cross-reference table.

Step 3. Get the content of Catalog object and all Page objects through searching the Page tree. Save all the Page objects in a stack.

Step 4. If the stack is empty, end; otherwise, obtain the information of XObject through reading the Page object in the top of stack. If XObject object exists, go to step 5; otherwise, go to step 4.

Step 5. If the subtype of the XObject is Image, go to Step 6; else, go to Step 4.

Step 6. Obtain the related information through reading the content of XObject and extract image. Decode the pixel data in the image to generate a normal image.

Step 7. Convert the image in different forms into BMP format with the same bit depth. Go to Step 4.

3.2. Binarization and Denoising of PDF Images

When a PDF image is a colourful one, it should be transformed into a gray image. The graying algorithm of colour images is fully discussed in related literatures, so this paper will not discuss it.

Binarization [9, 10] of gray images is to transmit an image in which a pixel is saved with multi binary bits into another one whose pixel has only two values called black and white saved with only one binary bit to express its gray level. Assume that $f(i, j)$ is the pixel value of image in point (i, j), then:

$$b(i,j) = \begin{cases} 0 & f(i,j) < \theta \\ 1 & f(i,j) \geq \theta \end{cases}$$

Where $b(i, j)$ is the pixel value of point (i, j) in the binarized image and θ is the binarying threshold value varying from the minimum value to the maximum value of gray values of the image.

The binarization method of gray images could be divided into two categories: global threshold methods and local thresholds methods. The former define a unique threshold value θ for the binarization of whole image. This algorithm has the advantage of being implemented simply and running fast. However, it could not process the images with inhomogeneous distribution of gray value. The robustness of the later is better because it could employ different threshold value θ in binarization process according to the situation of the current pixel. However, its calculating complexity is higher than the former.

Considering the diversity of PDF images, we employ the local threshold value method as the binarying strategy as discussed in literature [11].

The denoising of layout images which contain mathematical expressions is more complex than those of normal text. Some special symbols must be considered to avoid wrong deleting operations of components of mathematical symbols.

Step 1. Obtain all connected components on the layout image using a connected components searching algorithm.

Step 2. Identify the candidate noise components according to the geometric features obtained from statistic histogram of components size.

Step 3. Remove the seeming formula symbols' components such as "." of symbol "*i*" and "*j*" from the candidate noise components with the help of the composition knowledge of formula symbols in geometric level [12].

Step 4. Delete the candidate noise components.

3.3. Skew Detection and Correction of Document Images in PDF Files

The kernel technology of skew correction of document images is the skew detection of layout images.

The strategies of skew detection could also be divided into two classes called top-down method and bottom-up method [13-15].

The typical method of top-down method is projection algorithm. Through multiple projection operation to an image with different angles, the state of pixel distribution could be obtained and the angle in which the maximum space of white pixels occurs on the projection histogram is identified as the skew angle of the layout image. This method is simple to be implemented. But it is only suitable for the skew detection of simple layout images.

The bottom-up method searches the layout components, such as connected areas, firstly. Then, it detects the skew angle of the layout in the combining process of components. For example, in a Hough transform based skew detection method [13-15], the layout components are used as the sampling points of Hough transform to obtain the skew angle of the layout. This method could process any layout images as long as the layouts contain a certain number of characters used for Hough transform. The disadvantage of this method lies in that it might be disturbed easily when a layout image contains some linear elements such as mathematical symbols which would be wrongly extracted by Hough transform as the skew detection objects.

In this paper, the Hough transform based skew detection method [13-15] is employed and improved to detect the skew angle of document images in PDF files, in which the mathematical symbols are especially analyzed to avoid the occurrences of wrong skew detection.

Step 1. Obtain all connected components on the layout image using a connected components searching algorithm.

Step 2. The connected components are combined into character boxes according to the distance threshold values between character components, characters, text lines and paragraphs acquired with the statistic histogram of the distance of connected components on the layout.

Step 3. Select effective sample points of Hough transform with the help of symbol features in mathematical formulas and character features in normal text.

Step 4. Calculate the sample points of Hough transform in every symbol box according to the sample rules. Fulfil Hough transform to get the skew angle.

Step 5. Rotate the image according to the got angle.

3.4. Detection and Estimation of the Layout Parameters of Document Images in PDF Files

Let $P(L, S, N)$ be a three-tuple of the layout parameters of document images in PDF files, in which L is the language kind of layouts (limited to Chinese or English), S is the statistic size of characters in layout which is correlated to the resolution of layout images, and N is the number used for obtaining the parameter S. $P(L, S, N)$ will be transferred to the symbol recognition and structure analysis modules as the reference of parameter selection.

Parameter L is the precondition of the resolution analysis of PDF images because different language kind of layouts has different features in character size. In this paper, character strokes' running-number feature based method is explored to identify the language kind of layouts [16-17].

The character size S is related to the language kind of layouts. Table 1 and Table 2 show the average value of actual size of printed Chinese and English characters in different type sizes measured by a connected component searching algorithm, from which we can see the difference in size between two kinds of layouts.

Therefore, S should be identified by means of statistic method and consider not only the language kind of layouts but also the influence coming from non-text symbols. In our method, the histogram method [18-20] is employed to obtain the estimated value of character size. From Figure 2 and Figure 3 we can see that the parameter S could be detected by the proper calculation of the data that comes from the corresponding histogram.

Table 1. The Average Height and Width of Printed Chinese Characters in Various Font Sizes

No.	Type size	Character number	Actual height (pixels)	Actual width (pixels)
1	0	638	127	119
2	2	1837	70	63
3	4	4591	46	43
4	6	4354	26	24
5	8	3864	18	18

Table 2. The Average Height and Width of Printed English Characters in Various Font Sizes

No.	Type size	Character number	Actual height (pixels)	Actual width (pixels)
1	42	750	88	71
2	22	2497	48	39
3	14	6196	31	23
4	7.5	5905	18	13
5	5	3875	13	14

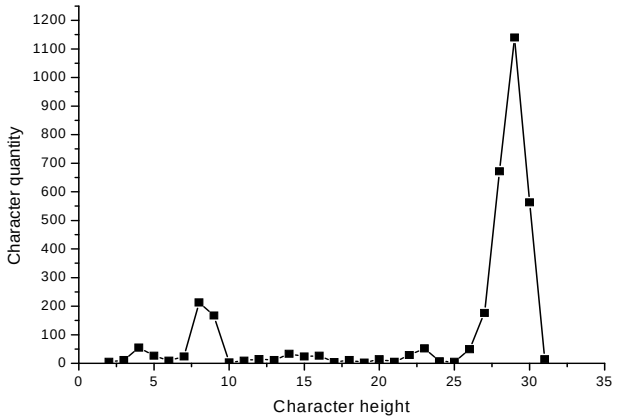

Figure 2. Histogram of a Page of Printed Chinese Characters (*Song*, 6 *Hao*)

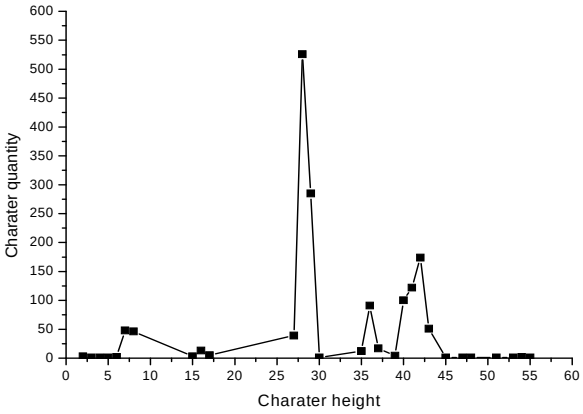

Figure 3. Histogram of a Page of Printed English Characters (Times New Roman, 14 Pound)

Step 1. Obtain all connected components on the layout image using a connected components searching algorithm.

Step 2. The connected components are combined into character boxes according to the distance threshold values between character components, characters, text lines and paragraphs acquired with the statistic histogram of the distance of connected components on the layout.

Step 3. Establish the statistic histogram of size of connected components in layout image.

Step 4. Analyze the statistic histogram of connected components in layout image and obtain the parameter *S*.

For Chinese characters, assume that *P* is the corresponding pixel value of character height to the max value in the statistic histogram. The significant character heights *H* in Chinese layout is denoted as:

$$H = (h_1,\ h_2,\ ...,\ h_n) \qquad (h_{i+1} > h_i,\ h_1 = \lfloor P\theta + 0.5 \rfloor,\ h_n = P + \Delta) \tag{1}$$

Where θ and Δ are parameters decided by experiment.

English characters could be classified into two types according to character height. Type 1 is the letters with the smaller height such a, c and e. Type 2 is the taller letters of b, A, and f. Let P_1 be the corresponding pixel value of character height to the max value in the statistic histogram for Type 1 and P_2 for Type 2. From Figure 4, we can know P_1=28 and P_2=42.

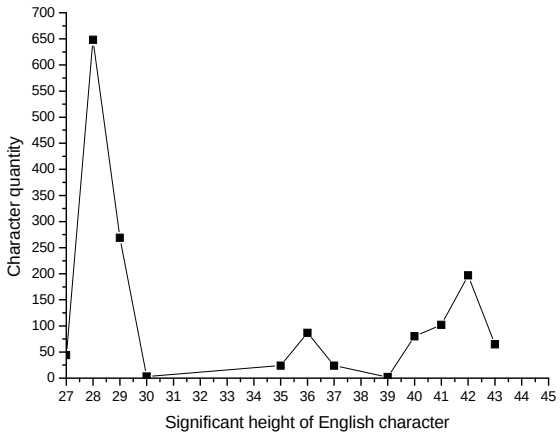

Figure 4. Histogram of Character Heights of an English Layout

The significant character heights *H* in English layouts is denoted as:

$$H = (h_1,\ h_2,\ ...,\ h_n) \qquad (h_{i+1} > h_i,\ h_1 = P_1 - \Delta_1,\ h_n = P_2 + \Delta_2) \tag{2}$$

Where Δ_1 and Δ_2 are parameters decided by experiment.

Other parameter definitions for Chinese and English characters are the same as follows.

The character numbers in *H* are defined as C_h in Equation (3).

$$C_h = (c_1,\ c_2,\ ...,\ c_n) \tag{3}$$

The pixel value of character widths corresponding to characters in *H* is defined as *W* in Equation (4).

$$W = (w_1,\ w_2,\ ...,\ w_n) \tag{4}$$

The character quantity corresponding to characters in *W* is defined as C_w in Equation (5).

$$C_w = (c_1,\ c_2,\ ...,\ c_n) \tag{5}$$

We have character average height H_{avg} and average width W_{avg} as Equation (6) and Equation (7).

$$H_{avg} = \sum HC_h / \sum C_h = \sum_{i=1}^{n} h_i c_i \Big/ \sum_{i=i}^{n} c_i \quad (6)$$

$$W_{avg} = \sum WC_w / \sum C_w = \sum_{i=1}^{n} w_i c_i \Big/ \sum_{i=i}^{n} c_i \quad (7)$$

Here we have S as Equation (8).

$$S = (H_{avg}, W_{avg}) \quad (8)$$

4. Experimental Results and Analysis

We developed a PDF image preprocessing and analyzing system with the proposed method in Visual C++ developing work bench. The sample PDF documents come from network in Chinese and English language. The images extracted from PDF documents in Chinese and English are shown in Figure 5 and Figure 6.

Figure 5. Extracted Image of Chinese Layout

Figure 6. Extracted Image of English Layout

The histograms of significant character heights of the images in Figure 5 and Figure 6 are shown in Figure 7 and Figure 8 respectively.

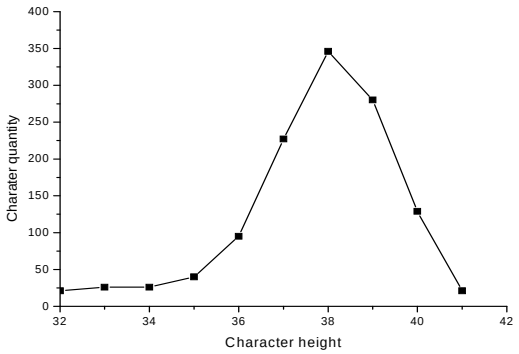

Figure 7. Histogram of Image in Figure 5

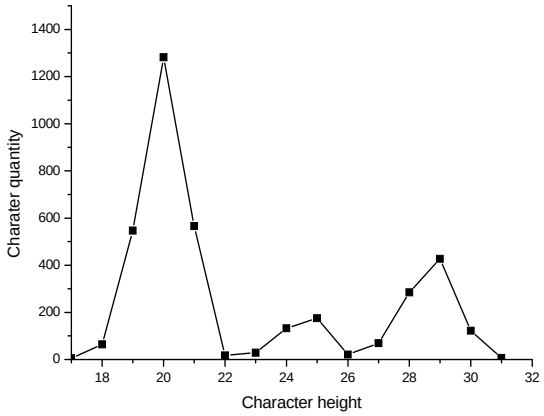

Figure 8. Histogram of Image in Figure 6

The estimation result of *S* is shown in Table 3.

Table 3. Estimation Results of *S* in Figure 5 and Figure 6

Layout type	Character number	Character height (pixels)	Character width (pixels)
Chinese	1211	38	38
English	3745	23	17

5. Conclusion

In this paper, a scheme of analyzing and preprocessing PDF images for mathematical expression retrieval is put forward. It includes the modules of PDF image extraction, graying, binarization, denoising, skew correction and layout parameter detection. Especially, some special strategies are designed for fitting the characteristics of mathematical components. The experimental result shows the effectiveness of the proposed method.

Although the proposed method could process the normally printed images in PDF documents, it has many shortages because of the variation in layout styles and image collecting modes. The further work is to improve the robustness of this scheme to process more kinds of layout images.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (Grant No. 61375075) and the Natural Science Foundation of Hebei Province (Grant No. F2012201020).

References

[1] Adobe Systems Incorporated. PDF Reference Version 1.7. 2006.
[2] Baker J, Sexton AP, Sorge V. *Extracting Precise Data on the Mathematics Content of PDF Documents*. DML 2008: Towards Digital Mathematics Library. Birmingham. 2008: 75-79.
[3] Baker J, Sexton AP, Sorge V. *A Linear Grammar Approach to Mathematical Formula Recognition from PDF*. ICM 2009: Intelligent Computer Mathematics. Berlin. 2009: 201-216.
[4] Lin XY, Gao C, Tang Z. *Mathematical Formula Identification in PDF Documents*. Proceeding of International Conference on Document Analysis and Recognition. Beijing. 2011: 1419-1423.
[5] Lin XY, Gao LC, Tang Z. *Identification of Embedded Mathematical Formulas in PDF Documents Using SVM*. Document Recognition and Retrieval. San Francisco. 2012; 8297 0D 1-8.
[6] Chen Y, Liu LZ, Ye H. Automatically Extracting Images of JPEG Format from PDF Documents. *Journal of Information Engineering University*. 2007; 8(2): 213-216.
[7] Wang JT, Kang XD, Li M, et al. Extraction of Recognizable Images from PDF File. *Computer Engineering and Design.* 2006; 27(9): 1539-1541.

[8] Li Q, Liu SJ. Design and Implementation of PDF Reader. *Computer Engineering and Design*. 2010; 31(7): 1635-1638.
[9] Zhang XZ. Chinese Character Recognition Technology. Beijing: Tsinghua University Press. 1992.
[10] Hu JZ. Computer Character Recognition Technology. Beijing: China Meteorological Press. 1994.
[11] Tian DZ. Recognition Preprocessing of Visual Document Image. PhD Thesis. Baoding: Hebei University; 2007.
[12] Liu S L. PDF Images Pre-processing for Formula Recognition. Master Disertation. Baoding: Hebei University; 2012.
[13] *Hinds SC, Fisher JL, D'Amato DP. A Document Skew Detection Method Using Run-length Encoding and the Hough Transform. Proceedings of the 10th International Conference on Pattern Recognition(ICPR). Atlantic City. 1990: 464~468.*
[14] *Le DX, Thoma GR, Wechsler H. Automated Page Orientation and Skew Angle Detection for Binary Document Images. Pattern Recognition. 1994; 27(10): 1325-1344.*
[15] Tian XD, Guo BL. The Method for Chinese Document Layout Analysis Based on Comprehensive Features. *Journal of Chinese Information Processing*. 1999; 13(4): 22-28.
[16] Lu XC, Yi BZ, Ping XJ, et al. English and Chinese Scripts Identification of Noised Document Image. *Computer Engineering and design.* 2007; 28(21): 5150-5152.
[17] Liang X. An Extraction Method for Mathematical Expressions in English and Chinese Printed Documents. Master Disertation. Baoding: Hebei University; 2010.
[18] Guo L, Ping X J, Zhou L. Script Identification of Document Image Based on Stroke Direction Histogram. *Journal of Information Engineering University*. 2011; 12(2): 231-237.
[19] Li C, Ding XQ, Wu YS. An Algorithm for Text Location in Images Based on Histogram Features and Ada Boost. *Journal of Image and Graphics*. 2006; 11(3): 325-331.
[20] Liang HW. Direct Determination of Threshold from Bimodal Histogram. *Pattern Recognition and Artificial Intelligence*. 2002; 15(2): 253-256.