Mathematical Formula Identification in Printed-Chinese Documents Based on EEN Feature Function

Chunning Hou
School of Computer Science and Technology, Hebei University, Baoding, China
Email: 1220812571@qq.com

Hongyan Ma
College of Mathematics and Information Science, Hebei University, Baoding, China
Email: mahongyan@hbu.edu.cn

Bingjie Tian
Department of Economic Trade, Hebei Finance University, Baoding, China
Email: tianbingjie@sina.com

Lina Zuo, Xuedong Tian*
School of Computer Science and Technology, Hebei University, Baoding, China
Email: zuolina@cs.hbu.cn, xuedong_tian@126.com,

***Abstract*—Aiming at the problems of mathematical formula identification in printed Chinese document images, a method based on EEN feature function is proposed. First, the EEN (Edge to Edge Notation) feature function which reflects the changing situation of connected components is defined, and corresponding algorithm which can extract the function value of image features is designed. Then, the characteristics of EEN feature function that it can reflect the distributions of images in horizontal and vertical directions intuitively and adequately is utilized to realize the layout analysis on symbol level and the basic information extraction of text lines. Finally, a locating method of isolated formulae and embedded formulae in printed Chinese document images is designed by using both the layout features and the content features of mathematical formulae. The experimental results show that this method can avoid the problems of the existing methods that their location accuracy is frequently affected by the symbol parts obtained from connected areas. It has good ability in layout composition discrimination and formula identification.**

***Index Terms*—printed Chinese document images, isolated formula identification, embedded formula identification, connected components, EEN feature function**

I. Introduction

As a special information carrier, mathematical expressions are not only the important components in many scientific documents, but also an important language among scientific persons. The study about mathematical expression recognition could be traced back to 1968, Anderson [1] who first put forward the analysis and recognition problems of mathematical expressions. After that, the attention about the mathematical formula recognition has being increased.

Mathematical formula identification in printed document images or PDF documents is the first step of mathematical formula recognition. The related research works have been carried out for a long time. One of the earliest studies of formula identification was Lee and Wang [2], they used Bayes decision rule to divide isolated formulae and plain text lines according to the differences between them. In embedded expression extraction, the special mathematical symbols and structures were recognized firstly. Then, these symbols were used as the centers for finding all the embedded expressions in expand range. Fateman et al. [3], [4] proposed a formula extraction method with bottom-up strategy. First of all, the connected components were searched. Then, they were classified and merged to make up formula areas. So that the formulae could be extracted and the exact position of each formula could be obtained. Literature [5] proposed a formula locating method for Chinese documents which was composed of three steps called the extraction of Chinese characters, the extraction of embedded formulas and the identification of isolated formula. First of all, they used decision tree to distinguish Chinese characters and non-Chinese characters. Secondly, the semantic information of formula symbols, the subscript relation information among symbols and the grammatical information of formulae were all used to locate embedded formulae from the non-Chinese characters. Finally, the Gaussian mixture model was

Manuscript received October 24, 2016; revised December 7, 2016.
*Corresponding author

doi: 10.12720/jait.8.1.29-35

utilized to identify isolated formulae. In [6] – [9], the content, layout and context features of text lines and words were extracted in digital documents. The text lines and words were classified by using SVM to locate isolated formulae and embedded formulae firstly. And then they improved the method and they combined the rule-based and learning-based method together to detect both isolated formulae and embedded formulae in PDF documents. And they used the preprocessing and post-processing rules to make the extracted content more accurate. Baker et al. [10] proposed an isolated formula locating method for PDF documents. According to the ratio of the size of mathematical symbols and the number of ordinary text words in text lines, the isolated formulae were located. In literature [11] a method that could extract formulae automatically based on circular projection statistics was designed. This method collected key information through projection to record the differences between the area of Chinese characters and formulae according to the features of symbols' width, spacing and density. And those differences were utilized as a reference for formula area identification. Li et al. [12] classified text lines and segmented Chinese characters by adding new features. And they used the feature that the change of the ratio of Chinese characters' width and height in a certain range to do the second comparison, so as to determine the embedded formulae. At last they used the contour tracking method to identify the formulae. Literature [13] proposed a method which could process the document images with the mixed pictures and text. It could reduce the effect on the results of formula extraction caused by pictures and tables in documents, and improve the accuracy of formula extraction. In [14], a content-based mathematical formula extraction method of Postscript documents with Word and LaTeX was proposed. According to the features of Chinese documents themselves, literature [15] used both statistical features and fuzzy logic based decision rules to extract isolated formulae. For embedded formulae, first, the statistical feature was used to make coarse classification for Chinese characters in images, and then meticulous classification by using the content features, so as to realize the extraction of them.

In conclusion, the method based on text line/word features would have problem that it would cause over-segmentation by using traditional classification strategy of text lines, which would result in the partial locating of formulae. The extraction methods based on symbol recognition frequently generate the error results that identify the characters as formula symbols, formula symbols as English characters. They were also easily influenced by the error recognition results caused by touching symbols. Some methods could not properly analyze the document layout containing graphs and tables. In order to solve these problems, an EEN feature function is defined for mathematical formula extraction which could express the distribution features of document images in horizontal and vertical directions and neglect the internal structure of characters which is useless for formula location. Based on the EEN feature function, a method of mathematical formula identification in printed Chinese documents is proposed which could analyze the documents' layout and identify the isolated formulae and embedded formulae with a better effect.

II. Definition of the EEN Feature Function

In this paper, we assume that the coordinate origin is in the upper left corner of images, and the scan order of the connected components in document images is line by line from top to bottom. The connected components are pixels in images that match a kind of connectivity rules (4 neighborhood connectivity or 8 neighborhood connectivity) and are represented by the same label [16]. The EEN feature function is defined on the basis of the pixel connected rectangles on the images of printed document layouts.

A. Optimization of the Connected Components

In the results of searching connected regions in images, the multi-component symbols would generate multi-connected areas. So, before the EEN feature function is defined, the connected areas that belong to the same symbols should be merged to reflect the distribution of the symbols.

Let P_1 , P_2 be any two adjacent connected regions whose coordinates of the upper left corner and the lower right corner are P_1: (X_{11},Y_{11}) , (X_{12},Y_{12}) ; P_2: (X_{21},Y_{21}), (X_{22},Y_{22}), respectively.

(1) Up and down structures

If

$$\begin{aligned}&(((Y_{12}+d_1)>Y_{21})\wedge\\&(((X_{11}>X_{21})\wedge(X_{12}<X_{22}))\\&\vee((X_{11}<X_{21})\wedge(X_{12}>X_{22}))\\&\vee((X_{11}>X_{21})\wedge(X_{11}<X_{22}))\\&\vee((X_{12}>X_{21})\wedge(X_{12}<X_{22})))\\&=\text{TRUE}\end{aligned} \tag{1}$$

Then merge them.

(2) Left and right structures

If

$$\begin{aligned}&(((X_{22}+d_2)>X_{11})\vee((X_{12}+d_2)>X_{21}))\wedge\\&(((Y_{11}<Y_{21})\wedge(Y_{12}>Y_{22}))\vee((Y_{11}<Y_{21})\wedge(Y_{12}>Y_{21})))\\&=\text{TRUE}\end{aligned} \tag{2}$$

Then merge them.

Where the thresholds d_1 and d_2 are the distance among characters derived from the height and width statistical histograms of initial characters in images.

B. Definition of EEN Feature Function

Definition 1: The function that describes the changes from one side of the rectangles to the other in the same direction is called EEN feature function. The rectangles are the bounding box of the connected components in the document layouts. Once an edge of the rectangle is scanned, it is recorded as an ETE. The function has two

modes, called *X*-EEN function and *Y*-EEN function, which reflect the situations of rectangle changes in horizontal and vertical directions, respectively.

The *X*-EEN function is defined as:

$$EX(t,j)=\begin{cases} TN_j & t=0 \quad TN_j \in \text{Even} \\ LT_j & 1\le t\le TN_j \quad t\in \text{Odd number} \quad 0\le j\le Height \\ RT_j & 2\le t\le TN_j \quad t\in \text{Even} \end{cases} \tag{3}$$

where TN_j is the total number of ETE on line j. It is two times of the number of connected areas on line j. LT_j stores the horizontal coordinate X_{k1}. RT_j stores the horizontal coordinate X_{k2}.

The *X*-EEN function has a strong ability to express horizontal features of document images. For example, $TN_j=0$ indicates that the line j is a blank one; otherwise, it is a non-blank one. And we can get the number of connected areas contained in this line according to the value of TN_j. The value of (RT_j-LT_j) can give the width of the connected areas. Therefore, the *X*-EEN function has a good effect on extracting information such as the width of connected areas, the height of text lines and the number of connected areas.

Let k be any connected components whose coordinates of the upper left corner and the lower right corner are (X_{k1},Y_{k1}) and (X_{k2},Y_{k2}). The extracting algorithm of *X*-EEN function is shown in algorithm 1.

Algorithm 1 *X*-EEN function extracting algorithm

```
INPUT: binary images of printed documents
OUTPUT: the value of EX(t, j)
while( j ≠ Height )
{
    t = 0;
    while( k ≠ CNum )// CNum is the total number of
//connected areas
    {
        if( j > Y_k1 && j < Y_k2 )
        {
            t++;
            EX(t, j) = X_k1;
            t++;
            EX(t, j) = X_k2;
        }
    }
    EX(0, j) = t;
}
```

Similarly, the *Y*-EEN function $EY(t,i)(0\le i\le Width)$ of document images in the *Y* direction and the corresponding extraction algorithm can be defined.

III. Mathematical Formula Identification Based on EEN Feature Function

The formulae in document images can be divided into two types called isolated formulae and embedded formulae. The isolated formulae locate on single lines. While embedded formulae coexistence with ordinary texts. In addition, the columns situation would also influence the formula extraction. The documents may be single-column layout or double-column layout. The main difference between them on formula identification is that we should identify formulae on the left area and right area respectively in the double-column document images. However, the extraction algorithms of them are consistent. So this paper takes the single-column document images as the object to analyze.

A. Layout Analysis

A document layout usually contains header, footer, table, text and other fields with the single-column layout or double-column layout, which would cause certain interference on formula locating. In order to accurately extract formulae, it is necessary to analyze the layouts of document images before formula identification.

1) Determination of the layout column

The characteristic that there are no connected areas or less connected areas in vertical midline of the layout is used to judge whether the document is a double-column one. We need to obtain the left text areas and right text areas of the document if it's a double-column document by the value of $EY(0,i)(0\le i\le Width)$. The algorithm is shown in algorithm 2.

Algorithm 2 Determination algorithm of layout column

```
INPUT: binary images of printed document
OUTPUT: the value of IsDouble
While( i ≠ Width ){
    if( EY(0,i) ≠ 0 ){
        sumETE += EY(0,i);
        colNum++;// EY(0,i) ≠ 0 is total number of column
    }
}
averETE = sumETE / colNum;
While( i ≠ Width ){
    if( EY(0,i)*3 < averETE )   BJ[i]=1;
    else {
    if( EY(0,i)*2 < averETE )   BJ[i]=1; }
    else{
    if( EY(0,i) < averETE )   BJ[i]=1; }
    else   BJ[i]=0;
}
While( i ≠ Width ){
    if( BJ[i]==1 && BJ[i]==BJ[i+1] )   n++;
    else{
        if( n ≠ 0 && n ≥ averW )// averW is average width for
//characters
        {
            temp = 1;
            Lcol = i - n + 1;
            Rcol = i;
        }
        n = 0;
    }
}
if( temp == 0 )   IsDouble = 1; //signal-column
else   IsDouble = 0; //double-column
```

2) Determination of text areas

Employ both *X*-EEN function and *Y*-EEN function to identify the text areas in layout images. The algorithm is shown in algorithm 3.

Algorithm 3 Determination algorithm of text areas

```
INPUT: binary images of printed document
OUTPUT: the text area of double-column
While( j ≠ Height ){
    if( (EX(TN_j, j) > Lcol && EX(TN_j, j) < Rcol ||
    (EX(TN_j, j) < Lcol && EX(TN_j, j) > Rcol) )//have
//connected areas
        IsEmpty[j] = 1;
    else
        IsEmpty[j] = 0;
}
While( j ≠ Height ){
    if( IsEmpty[j] == 0 && IsEmpty[j] == IsEmpty[j+1]
)
        n++;
    else{
        if( n ≠ 0 && n ≥ averH )// averH is average height
//of characters
        {
            TextTop[t++] = j − n + 1;
            TextTop[t++] = j;
        }
        n = 0;
    }
}
```

B. Formula Identification

In this paper, before to identify the formulae, we should extract the features firstly by using EEN feature function including the line features and connected components features as shown in Fig. 1.

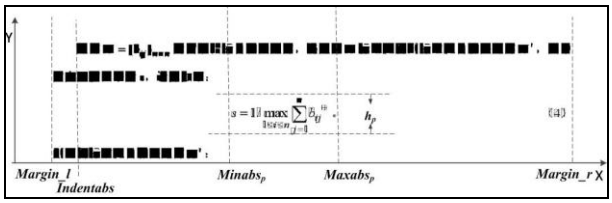


Figure 1. Line feature extraction.

In Fig. 1, the black areas represent the bounding boxes of the connected components of Chinese characters, the same below in Fig. 2 and Fig. 3.

The line features are defined as follows in EEN mode.

(1) The height of a row

$$h_p = j_{ep}\,|_{j_{ep}\in[EX(0,j_{ep})\neq 0\wedge EX(0,j_{ep}+1)=0]} - j_{sp}\,|_{j_{sp}\in[EX(0,j_{sp})\neq 0\wedge EX(0,j_{sp})=0]} \quad (4)$$

where h_p be the height, j_{ep} be the terminate ordinate and j_{sp} be the starting ordinate of the line p.

(2) The width of characters

If $j_{sp} \le EY(TN_{i_{ek}}, i_{ek}) = EY(TN_{i_{ek}+1}, i_{ek}+1) = EY(TN_{i_{sk}}, i_{sk}) = EY(TN_{i_{sk}-1}, i_{sk}-1) \le j_{ep}$.

Then

$$W_k = i_{ek}\,|_{i_{ek}\in[EY(TN_{i_{ek}}, i_{ek})\neq 0\wedge EY(TN_{i_{ek}+1}, i_{ek}+1)=0]} - i_{sk}\,|_{i_{sk}\in[EY(TN_{i_{sk}}, i_{sk})\neq 0\wedge EY(TN_{i_{sk}-1}, i_{sk}-1)=0]} \quad (5)$$

where W_k be the character width, i_{sk} be the starting abscissa and i_{ek} be the terminate abscissa of character k. $TN_{i_{sk}}$, $TN_{i_{sk-1}}$, $TN_{i_{ek}}$ and $TN_{i_{ek+1}}$ are all odd number.

(3) The min/max horizontal coordinates of the main content of a layout

Let j be any line, then

$$\begin{cases} \text{Margin\_}l = \min(EX(1, j)) \\ \text{Margin\_}r = \max(EX(TN_j, j)) \;\; TN_j \in Even \end{cases} \quad (6)$$

(4) The min/max horizontal coordinates of text line

Let $j_{sp} \le t \le j_{ep}$, then

$$\begin{cases} Minabs_p = \min(EX(1, t)) \\ Maxabs_p = \max(EX(TN_t, t)) \;\; TN_t \in Even \end{cases} \quad (7)$$

(5) The min horizontal coordinates of text-indent

Let $AverH$ be the average height of the document. If

$$(Minabs_p - Margin\_l) \in ((3*AverH)/2, 3*AverH) \quad (8)$$

Then

$$sum+ = Minabs_p \quad (9)$$

$$n++ \quad (10)$$

Then

$$Indentabs = sum / n \quad (11)$$

1) Isolated formula identification

Isolated formulae have many differences in geometric characteristics compared with the other components on document layouts. So it is relatively easy to locate isolated formulae by using EEN feature function. The isolated formulae are divided into several types in this paper.

(1) Indent type

Take the isolated formula shown in Fig. 2 as an example. It has an important feature that the minimum horizontal ordinate of the formula lines is the biggest one in all lines or between $Margin\text{-}l$ and $Indentabs$. We can use it to locate the isolated formulae. However, the titles of tables also have the same feature. So it's necessary to use the average height of rows $AverH$ to further determine them.

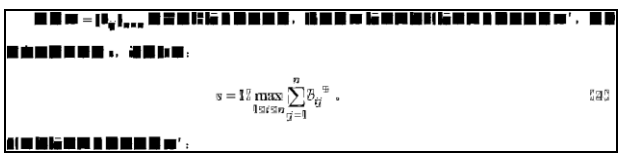


Figure 2. Isolated formula of indent type.

When the minimum horizontal ordinate of the line p is meet the (12).

$$\begin{aligned}&(((Minabs_p > Indentabs) \wedge (h_p > AverH)\\&\wedge(h_p < 6*AverH)) \vee ((Minabs_p < Indentabs)\\&\wedge(Minabs_p > Margin\_l) \wedge (h_p > AverH)\\&\wedge(h_p < 6*AverH))) = \text{TRUE}\end{aligned} \quad (12)$$

Then all the characters between j_{sp} and j_{ep} are isolated formula characters and we should recorded them into the array $Formula\_is$.

(2) Aligned type

Take the isolated formula shown in Fig. 3 as an example. The isolated formulae in this type has an important feacture that the minimum horizontal ordinate of formulae is equal with $Margin\text{-}l$ or $Indentabs$.

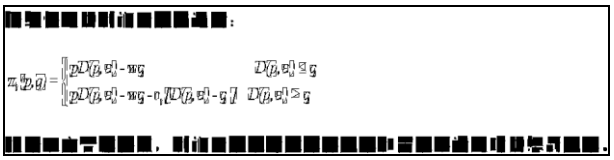


Figure 3. Isolated formula of aligned type.

A). Aligned with $Margin\_l$

When the minimum horizontal ordinate of the line p is meet the (13).

$$\begin{aligned}&((Minabs_p = Margin\_l)\\&\wedge(h_p > AverH)\\&\wedge(h_p < 6*AverH))\\&= \text{TRUE}\end{aligned} \quad (13)$$

Then all the characters between j_{sp} and j_{ep} are isolated formula characters and we should recorded them into the array $Formula\_is$.

B). Aligned with $Indentabs$

When the minimum horizontal ordinate of the line p is meet the (14).

$$\begin{aligned}&((Minabs_p = Indentabs)\\&\wedge(h_p > AverH)\\&\wedge(h_p < 6*AverH))\\&= \text{TRUE}\end{aligned} \quad (14)$$

Then there are two other features of formula will be introduced.

a) Located at the middle of the layout

If

$$(Minabs_p - Margin\_l)/(Margin\_r - Maxabs_p) = 1 \quad (15)$$

Then all the characters between j_{sp} and j_{ep} are isolated formula characters and we should recorded them into the array $Formula\_is$.

b) Contain the series number at the end of the line

If

$$\begin{aligned}&((EX(T-1,t) - EX(T-2,t)) > AverH/2)\\&\vee((EX(T-3,t) - EX(T-4,t)) > AverH/2)\\&\cdots\cdots\\&\vee((EX(T-11,t) - EX(T-12,t)) > AverH/2)\\&= \text{TRUE}\end{aligned} \quad (16)$$

Then all the characters between j_{sp} and j_{ep} are isolated formula characters and we should recorded them into the array $Formula\_is$. In which $t = j_{sp} + h_p/2$ and $T = EX(0,t)$.

(3) Multi-lines

Take the isolated formula shown in Figure 4 as an example. The feature of these formulae is that it contains the upper part or lower part, and they are all alone in a line. It is easy to be lost during positing due to the upper part or lower part belongs to a small component. So it is necessary to make a re-determination.

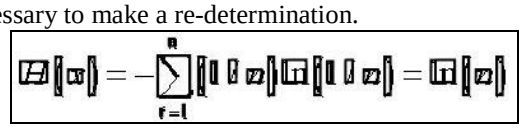


Figure 4. The isolated formula of multi-lines type.

Let line p has been marked as a text line for isolated formula. If

$$Minabs_{p\pm1} > Minabs_p \quad (17)$$

When

$$\begin{aligned}&(j_{sp} - j_{e(p-1)}) < averLineSpace\\&\vee(j_{s(p+1)} - j_{ep}) < averLineSpace\\&= \text{TRUE}\end{aligned} \quad (18)$$

Then all the characters between $j_{s(p-1)}$ and $j_{e(p-1)}$ or $j_{s(p+1)}$ and $j_{e(p+1)}$ are recorded into the array $Formula\_is$. And $averLineSpace$ is the average spacing when $EX(0,j) = 0 (0 \le j \le Height)$.

2) Embedded formula identification

The embedded formulae exist in ordinary text lines. There are many differences in size and ordinate between embedded formula symbols and Chinese characters. Therefore, we must to find these differences for realizing embedded formulae location.

Let the vertical ordinate of the text line p is j_{sp}, the vertical ordinate of k is $EY(TN_i, i)$ $(0 \le i \le Width, TN_i \in \text{Odd number})$ and $j_{sp} \le EY(TN_i, i) \le j_{ep}$. The features of the embedded formula extraction are as follows.

(1) Baseline features

The Chinese characters have an important feature that they share the same baseline, while the embedded formula characters are not. We could find that the left-upper corner of connected components of embedded formulae is bigger than those of Chinese characters, but

smaller than those of the punctuation marks on the same line.

If

$$((EY(TN_i,i) > j_{sp}) \wedge (EY(TN_i,i) < Symord)) = \text{TRUE} \quad (19)$$

Then the character k is recorded into the array $Formula\_em$. In which $Symbol = \min(EY(TN_i,i))$.

(2) Area features

Generally, the area of the embedded formulae is smaller than Chinese characters.

Let the area of character k is $Area_k = ((EY(TN_{i_{sk}}+1,i_{sk}) - EY(TN_{i_{sk}},i_{sk}))*W_k$.

If

$$((Area_k < averA*4/5) \wedge (Area_k > Minarea)) \vee (Area_k > averA*3/2) = \text{TRUE} \quad (20)$$

Then the character k is recorded into the array $Formula\_em$. In which $averA = \sum_{k=1}^{CNum} Area_k \Big/ CNum$ and $Minarea = \min(Area_k)$。

C. Experiment Results and Analysis

A text layout analysis and formula identification system based on EEN feature function are implemented in this paper.

Table I shows the result of isolated formulae location on the test set of 203 single-column and double-column document images, and embedded formulae location on the test set of 102 single-column and double-column document images based on the method of EEN feature function.

TABLE I. THE RESULT OF FORMULA EXTRACTION BASED ON EEN FEATURE FUNCTION

Types	Number of formulae	Number of identify	Correct identify of the formulae
Isolated formulae	932	943	827
Embedded formulae	1007	1189	773

For isolated formulae, the method can ignore the internal structure of characters. It's easy to get the result directly by making use of layout and content of the formulae with EEN feature function. When the height of isolated formulae is too small or the minimum horizontal coordinate of formulae is equal to $Margin-l$ or $Indentabs$ but the height is smaller than $AverH$, the locating errors will occur. This is because the embedded formulae exist in text lines, their location mainly use the size features. And the punctuation and the upper right corner will also produce some influences for it.

IV. CONCLUSION

In this paper, we describe a mathematical formula identification method based on EEN feature function for printed Chinese documents. This method can make use of the layout features of documents and the geometry features of mathematical formulae to achieve identification work. The value of EEN feature function can intuitively reflect the layout characteristics of document images. It is a useful tool for the analysis of layout columns, extraction of the information of characters and text information. With the help of it, the layout analysis and formulae identification can be realized. The main advantage of isolated formulae identification based on EEN feature function is that it can avoid the influence caused by character components. The accuracy of embedded formulae depends on whether the characters can be merged correctly. The errors in merging characters would easily produce the locating mistakes of character areas which would result in the embedded formula be separated in two-parts. For these shortcomings, we will improve the parameters and identification algorithms to raise the accuracy of formula extraction in the future work.

ACKNOWLEDGMENT

This work is supported by the National Natural Science Foundation of China (Grant No. 61375075), the Natural Science Foundation of Hebei Province (Grant No. F2012201020; F2013201134) and the Project of Human Resources and Social Security of Hebei Province (Grant No. JRS-2016-1090).

REFERENCES

[1] R. H. Anderson, "Syntax-directed recognition of hand-printed two-dimensional mathematics," in *Interactive Systems for Experimental Applied Mathematics*, Academic Press, New York, 1968, pp. 436-459.

[2] H. J. Lee and J. S. Wang. "Design of a mathematical expression recognition system," in *Proc. 3rd International Conference on Document analysis and Recognition*, Montréal, Canada, 1995, pp. 464-468.

[3] R. J. Fateman, "How to find mathematical on a scanned page," in *Proc. SPIE-The International Society for Optical Engineering*, 1997, pp. 98-109.

[4] R. Fateman, T. Tokuyasu, B. P. Berman, and N. Mitchell, "Optical character recognition and parsing of typeset mathematics," *Journal of Visual Communication and Image Representation*, vol. 7, no. 1, pp.2-15, March 1996.

[5] Y. S. Guo, N. T. Tan, Ch. P. Liu, and L. Huang, "An identification method for mathematical expression in scanned Chinese document," *Journal of Chinese Information Processing*, vol. 22, no. 4, pp. 83-87, July 2008.

[6] X. Y. Lin, L. C. Gao, Z. Tang, X. F. Lin, and X. Hu, "Mathematical formula identification in PDF documents, " in *Proc. 2011 International Conference on Document Analysis and Recognition*, Beijing, 2011, pp. 1419−1423.

[7] X. Y. Lin, L. C. Gao, Z. Tang, X. Hu, and X. Y. Lin, "Identification of embedded mathematical formulas in PDF documents using SVM," in *Proc. International Conference on Document Recognition and Retrieval XIX*, San Francisco, USA, 2012, pp. 1−8.

[8] X. Y. Lin, L. C. Gao, and Z. Tang, "Research on mathematical formula identification in digital Chinese documents," *Acta Scientiarum Naturalium Universitatis Pekinensis*, vol. 50, no. 1, pp. 17-24, Jan. 2014.

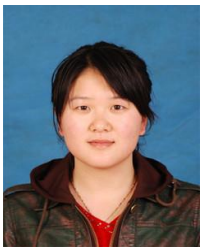
[9] X. Y. Lin, L. C. Gao, Z. Tang, J. Baker, and V. Sorge, "Mathematical formula identification and performance evaluation in PDF documents," *International Journal on Document Analysis and Recognition*, vol. 17, no. 3, pp. 239-255, 2014.

[10] J. Baker, A. P. Sexton, and V. Sorge, "Towards reverse engineering of PDF documents," in *Proc. the 4th Workshop*, Bertinoro, 2011, pp. 65-75.

[11] X. Y. Peng and J. P. Mao, "Mathematical formula automatic location method based on circular projection statistics," *Journal of Image and Signal Processing*, vol.2, pp. 37-41, 2013.
[12] D. R. Li and T. D. Xu, "Research on an extraction method for mathematical formulas embedded in printed documents," *Computer Application and Software*, vol. 31, no. 4, pp. 102-110, 2014.
[13] F. Li, "Extraction, recognition and reconstruction of mathematics formulas in English scientific document," Dalian University of Technology, Dalian, China, 2007.
[14] Z. W. Zhang, "Research on digitization of mathematical expressions," Ph. D. dissertation, University of Science and Technology of China, Anhui, China, 2007.
[15] L. P. Zhang, "The study on the method of printed mathematical formula extraction," M. S. thesis, Hebei University, Baoding, China, 2007.
[16] H. B. Gao and W. X. Wang, "New connected component labeling algorithm for binary image," *Computer Applications*, vol. 27, no.11, pp. 2776-2785, 2007.

Chuning Hou was born in Hebei, China, 1990. She received the bachelor degree in Industrial and Commercial College, Hebei University, Baoding, Hebei, China, 2014. And she is currently pursuing the M. S. degree in the school of computer science and technology, Hebei University, Baoding, Hebei, China. Her research interests are image processing and pattern recognition.

Hongyan Ma was born in China, 1980. M. S.. She is a lecturer at Hebei University, Baoding China. Her research interests is intelligent information processing.



Lina Zuo was born in China, 1983. M. S.. She is a lecturer at Hebei University, Baoding China. Her research interests is intelligent information processing.

Bingjie Tian was born in China, 1988. M. S.. She is a teaching assistant at Hebei Finance University, Baoding China. Her research interests are are computational linguistics.

Xuedong Tian was born in China, 1963. Ph. D. He is a professor at Hebei University, Baoding China. His research interests include pattern recognition and image processing, information retrieval. He is the corresponding author of this paper.