An Improved Formula Extraction Method of Printed Chinese Layouts Based on Connected Component Run-length Feature

Fang Yang, Chunning Hou and Xuedong Tian*
School of Computer Science and Technology
Hebei University
Baoding, China
e-mail: yangfang@hbu.edu.cn; 1220812571@qq.com; *corresponding author: xuedong_tian@126.com

***Abstract*—The mathematical formula extraction is the prerequisite of formula structure analysis, recognition and retrieval. This paper studies the formula extraction method for the printed Chinese scientific and technical document images, proposes a criterion based on connected component run-length feature to estimate formulae in text lines, and then improves the formula location method based on rules. The connected component run-length's change regularity was analyzed firstly for all symbols in a text line. Then Change-rate threshold was set to estimate whether there is formula in this line. Finally, improved formula extraction method was given. The experimental results on the samples collected from printed Chinese scientific and technical documents showed that the proposed method is effective in estimate the embedded formula, and improves the accuracy of the formula location.**

Keywords-formula image; Chinese; formula location; connected component; run-length

I. INTRODUCTION

Mathematical formula is a special information carrier in scientific documents. It provides convenience for the academic exchange of different fields, and the retrieval of mathematical formula has become the focus of information retrieval research. Some scientific documents appear in the form of images, and the locations of mathematical formulae in the images become the premise of mathematical formula recognition and information extraction [1]. Mathematical formula symbols are not of uniform size, and they combined in two dimensional pattern different from normal text. In addition, a formula may appears in different position, as isolated formula which is a line, and embedded formula which is a part of a line. All those characteristics also make it difficult to locate the mathematical formula in the layout images of scientific documents.

Not only recognizing formulae but also realizing formulae retrieval need to extract formulae in scientific layout images. Lin et al [2] directly delineate the formula locations in the PDF document manually for obtaining the key words of math query in their mathematical expression retrieval system. Because of the unique characters of the mathematical formula, there is a distinct difference between the isolated formula and the ordinary text. M. Alkalai et al [3] introduced formula recognition technology to distinguish formula lines from text lines with the help of recognizing results. Chang et al [4] used the BP neural network to correct the recognition errors of text lines, and improve the recognition accuracy of formula lines. Suzuki [5] used a bottom-up locating method of formulae. Starting from basic symbols, they analyzed the sizes and locations of adjacent symbols to divide them into text areas and formula areas, and then gradually merged them so as to determine the positions of the formulae. Fateman [6] used the connected components as the base elements to locate the formulae in document layout images. Because of the disorder of the search results of connected components, the analysis of the relationship between adjacent symbol positions is quite complex, and the basic symbol detection method is gradually transformed into projection method. Wang [7] used top-down projection method to analyze the locations of formulas in many kinds of layout style. Peng [8] used a cyclic projection to map the difference between character and formula symbols. Hou et al [9] proposed the connected component run length feature function. According to the extracted Chinese document layout parameters, the location rules were set to locate the isolated and embedded formulas.

For Chinese scientific documents, the differences between Chinese characters and formula symbols are obvious, and the rule based localization methods are more concise and clear. But the proportion of the mathematical formula to the normal text in scientific documents is large, and the double column typesetting is usually used. The format parameter would be incorrect because of the lack of the characters of normal text, which leads to the decrease of the reference of the rule parameter and the accuracy of the embedded formula extraction. Therefore, this paper improves the rule based location method in the literature [9], analyzed the distribution regularity of symbol's connected component run-length in a text line, gives the estimation whether the line contains the formula or not, and then locates the mathematical formulas based on the rules.

In the second section, the principle of the mathematical formula estimation and the judging rule based on the run length feature function are introduced in detail. The third section introduces the application of the method in the extraction of isolated mathematical formulas and embedded mathematical formulas. The experiment process and the result analysis are introduced in the fourth section, and finally the conclusion is given.

II. MATHEMATICAL FORMULA ESTIMATION METHODE

A. Estimation principle

Mathematical formulae are generally made up of many symbols, each of which has different shapes and sizes, whereas Chinese character's size is almost the same. When Chinese text contains mathematical formulae, the width of the outer

978-1-5386-0612-4/17 $31.00 © 2017 IEEE
DOI 10.1109/ICVISP.2017.28

rectangle of Chinese characters and mathematical symbols is obviously different.

Assume that a line of text contains M characters, $S=(s_1,s_2,...,s_M)$, the kth character's width is $w(k)$, $1\le k\le M$. Then the width variable quantity of the two adjacent characters is $\Delta w(k)$.

$$\Delta w(k)=\begin{cases}|w(k)-w(k-1)| & 2\le k\le M\\ w(k) & k=1\end{cases} \quad (1)$$

Normalized the $\Delta w(k)$, then the change rate is $\Delta'(k)$.

$$\Delta'(k)=\frac{\Delta w(k)}{\Delta w(k)+w(k)}\times 100\% \quad (2)$$

When the adjacent characters are Chinese characters, the width changes little, so the width change rate is approximately 0. In other cases, the greater the difference in width between adjacent characters, the greater the rate of width change. Then, set width change rate threshold T_Δ, and the rule to estimate formula symbol is,

- Rule 1 if $\Delta'(k)\ge T_\Delta$, then $S_k\in F$, where F is the formula symbol set.

When a text line contains a mathematical formula, the formula set will has a string of elements which have serial indexes. Set length threshold T_L, then the rule to estimate formula is

- Rule 2 if $(s_t,s_{t+1},...,s_{t+l})\subset F$ and $l\ge T_L$, then this string is formula.

B. Realization of estimation rules

The width change rate is calculated based on the character width. When a connected component is used to represent character, the width of the character is the distance from the left border to the right boundary of the connected component, that is, the run-length.

Assume that there are M connected components on the jth line of image whose unit is pixel, the connected component run-length feature function is defined as reference [9]:

$$EX(t,j)=\begin{cases}TN_j & t=0\\ LT_j & t\ is\ odd\\ RT_j & t\ is\ even\end{cases} \quad (3)$$

Where t is variable, $0\le t\le 2M$; The LT_j represents the left boundary coordinate of the connected component, the RT_j represents the right boundary coordinate; TN_j is the number of boundary coordinates of the line, so $TN_j=2M$.

When t is even, the $\left(t/2\right)$th connected component's width is

$$w(t/2)=EX(t,j)-EX(t-1,j) \quad (4)$$

For a text line, supposed j is the middle-line of this string of connected components, then the algorithm of estimation rules is shown as Table 1.

TABLE I. ALGORITHM OF ESTIMATION RULES

Input: $EX(t,j)$, T_Δ, T_L

Output: *start*, *L*

```
1   j=mid-line;
2   for (t=2; t<EX(0,j); t+=2)
3   {
4       w[t/2]=EX(t,j) - EX(t-1,j);
5   }
6   w[0]=w[1];  label=false;  L=0;
7   for (i=1; i<=EX(0,j)/2; i++)
8   {
9       calculate Delta[i];
10      if (Delta[i]> T_Δ)
11      {
12          L++;
13          if ( label==false)
14              {start=i; label=true; }
15      }
16      else
17      {
18          if (label==true)
19              {
20                  label=false;
21                  if ( L> T_L )  return ( start, L);
22                  else  {start=0; L=0;}
23              }
24      }
25  }
```

III. FORMULA EXTRACTION METHOD

The flow chart for formula extraction is shown in figure 1. There are three set of rules in this method, rule set for isolated formula extraction, formula estimation and embedded formula extraction. All those rules are used thresholds or parameters which are calculated by connected component run-length feature function.

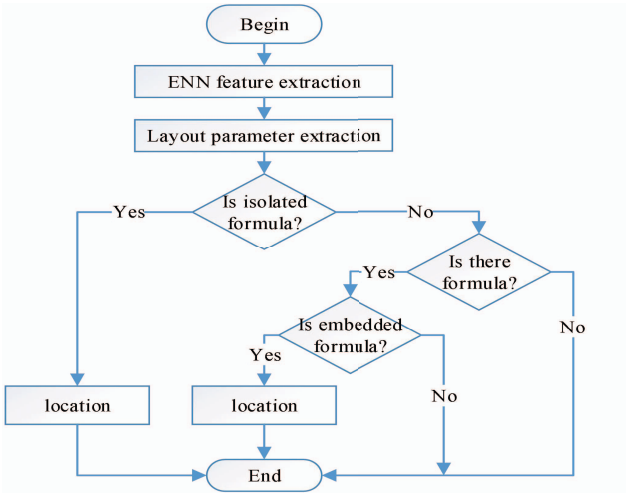

Figure 1. Flow chart for formula location method.

The isolated and embedded formulas are identified through employing the rules proposed in reference [9].

A. Isolated formula identification rules

- $(Minabs > Indentabs) \cap (1.2 \times AverH < h_p < 6 \times AverH)$ (5)

- $(MarL < Minabs < Indentabs) \cap (1.2 \times AverH < h_p < 6 \times AverH)$ (6)

Minabs is the left boundary of text line, h_p is the height of this line, *Indentabs* is the text indent, *AverH* is the average line height, *MarL* is the left boundary of text block. The line is identified as isolated formula as long as it satisfies equation (5) or (6).

B. Embedded formula identification rules

- $|Cy - baseline| < T_{line}$ (7)

- $AreaC < AverCha$ (8)

Cy is the component's left-upper corner *Y* axis coordinate, baseline is the Chinese characters top line, T_{line} is threshold. *AreaC* is the area of this component, and *AverCha* is the average area of all the Chinese characters. The connected component is identified as a formula symbol as long as it satisfies both equation (7) and (8).

IV. Experiments and Analysis

The proposed method is experimentally implemented on 309 randomly selected scientific and technical document images, and all those images are divided into 807 text blocks. Firstly, the connected component run-length feature function of each text image blocks is extracted, then the parameters are calculated according to the reference [9].

Chinese text lines, isolated formula lines and text lines containing embedded formulas, and their component widths and corresponding connected component width change rates are shown from Figure 2 to Figure 7 respectively.

Figure 2. Sample of Chinese character line

Figure 3. Width change rate of Chinese character line

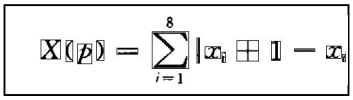

Figure 4. Sample of isolated formula

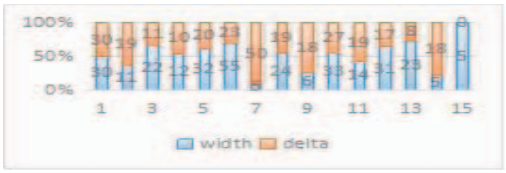

Figure 5. Width change rate of isolated formula line

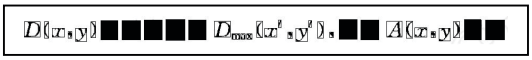

Figure 6. Sample of text line containing embedded formulas

Figure 7. Width change rate of text line containing embedded formulas

Finally, formula extraction method is tested in those samples and compared with reference [9]'s method. Table II shows the results, in which we set $T_{\Delta} = 10\%$, $T_L = 5$.

The improved method's correct identification rate for isolated formula is 89.27% , and the rule-based method is 88.74%. The effe ct of the new method does not change much with the original method in the isolated formula identification, because the same rules are used with little different coefficient.

For the embedded formula, the new method's correct identification rate is 69.25%, as where the rule-based is 65.72%.

TABLE II. Results of Formula Extruction Algorithm

	#formula	#identified		#correct identified	
		Rule-based	New	Rule-based	New
Isolated formula	941	961	950	835	840
Embedded formula	1132	1321	1153	744	784

V. Conculation

In this paper, we improve the rule-based mathematical formula location method by add formula estimation rules. The connected component run-length feature is used to calculate the width change rate, which can represent the change from Chinese character to formula symbols. Connected components in a line are classified by width change rate threshold, then we can estimate if there is formula in this text line. This estimation helps formula location method to get embedded formula easily. The experiment results shows that the estimation rule based on connected component run-length feature is effective to embedded formula identification.

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)，the Key Project of the Science and Technology Research Program in University of Hebei Province of China (Grant No. ZD2017208) and the Project of "One Province One University".

REFERENCES

[1] R. Zanibbi, D. Blostein, "Recognition and retrieval of mathematical expressions," International Journal on Document Analysis and Recognition. vol.15, No. 4, pp. 331-357, 2012.

[2] X.Y. Lin, L.C. Gao, and Z. Tang, "Research on mathematical formula indentification in digital Chinese documents,"Acta Scientiarum Naturalium Universitatis Pekinensis, vol. 50, No. 1, pp. 14-24,Jan. 2014.

[3] M. Alkalai, J.B. Baker, V. Sorge, et al," Improving formula analysis with line and mathematics identification," Proceedings of 12th International Conference on Document Analysis and Recognition. Washington, DC, USA, pp. 334-338, Aug, 2013.

[4] X.F. Chang, J. Cui, X.W. Liu, et al. "Research on mathematical formulas extraction from printed document based on neural network," Application Research of Computers, vol. 25, No. 11, pp. 3483-3485,2008.

[5] M. Suzuki, F. Tamari, R. Fukuda, et al. "INFTY: an integrated OCR system for mathematical documents," Proceedings of the ACM Symposium on Document Engineering, Grenoble, France, pp.95-104,Nov,2003.

[6] R.J. Fateman. "How to find mathematical on a scanned page," Proceedings of SPIe, No.3967, pp. 98-109,1999.

[7] Y. Wang. "Research and implementation of mathematical formula location and recognition algorithm in text," Beijing: Beijing JiaoTong University, 2016.

[8] X.Y. Peng, J.P. Mao,"Mathematical fromula automatic location method based on circular projection statistics," Journal of Image and Signal Processing. Vol. 2, pp.37-41,2013.

[9] C. Hou, H. Ma, B. Tian, et al， "Mathematical formula identification in printed Chinese documents based on EEN feature function," Jorunal of Advances in Information Technology,vol. 8, No. 1, pp.29-35 ,February 2017.