Robust Math Formula Recognition in Degraded Chinese Document Images

Ning Liu*, Dongxiang Zhang†, Xing Xu†, Long Guo‡, Lijiang Chen§, Wenju Liu* and Dengfeng Ke*
*National Lab. of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China
Email: liuning19880928@gmail.com, {wenju.liu, dengfeng.ke}@ia.ac.cn
†School of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu, China
Email: {zhando, xing.xu}@uestc.edu.cn
‡School of Electronics Engineering and Computer Science, Peking University, Beijing, China
Email: guolong@pku.edu.cn
§Lejent technology Co., Ltd, Beijing, China
Email: lijiang.chen@lejent.com

***Abstract*—In this paper, we study the problem of math formula recognition (MFR) in degraded Chinese document images. Compared to traditional optical character recognition (OCR), the MFR problem brings new challenges in terms of character segmentation and structural analysis, especially in degraded images. To tackle these issues, we propose an over-segmentation strategy to split and recognize adhesive formula elements based on convolutional neural network (CNN). In addition, we propose a hierarchical framework for formula structure analysis that constructs the formula in a top-down manner to iteratively split the regions into recognizable units. Due to the lack of degraded Chinese document images with math formulas in the community, we also harvest a diverse ground-truth dataset containing 100 images submitted from our system users. Extended experiments demonstrate the effectiveness and robustness of our proposed method in comparison with state-of-the-art methods.**

***Keywords*-Math Formula Recognition, Chinese Document Image, Convolutional Neural Network.**

I. Introduction

Optical Character Recognition (OCR) has been extensively studied for decades, and there have been some commercial OCR softwares [1], [17] available on the market. Nevertheless, there still retain many challenges for smartphone-captured document images, such as uneven illumination, low contrast, various noise and layout analysis. In particular, for math formulas, structural analysis is also a great challenge.

MFR has been well solved on high-quality document images [12], [13], [15] or on online handwritten systems [16]. While for degraded Chinese document images, this subject faces greater challenges for the effects of uneven illumination, low contrast and various noise. Math formula recognition in degraded Chinese document images is a challenging and important task in real applications. As a startup company for K12 online education, we provide services for students to find similar homework or exercises for more practice on their weak area. Especially, MFR in degraded Chinese document images is critical for us. To use our smartphone app, users simply need to take a picture of an exercise and issue the query to our system. Once receiving the image, we conduct OCR and use the extracted text information to query our knowledge database.

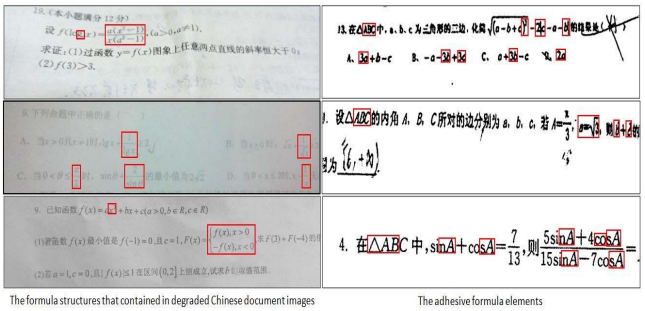

Figure 1. Examples to show the two challenges that we face when solving math formula recognition problem in degraded Chinese document images.

In this procedure, the accuracy of OCR plays a crucial role to find relevant results. In the millions of queries we receive everyday, math problems occupy the largest share of all subjects. The OCR results of math formulas are the most critical information to find relevant results for math problems. This is the motivation that we focus on this issue.

Most MFR systems [12], [13] utilize the connected components (CC) on the binary images to locate characters and conduct structure analysis for math formulas. The CC are not accurate on the binary results of low quality images, which adds two challenges in MFR: the analysis of formula structure that mixed in Chinese text and the recognition of adhesive formula elements. The two challenges are shown in Fig. 1, in which, the regions within red rectangles on the left column contain formula structures, and the regions within red rectangles on the right column contain adhesive formula elements.

In this paper, we focus on the OCR for math formula in degraded Chinese document images. We propose an end-to-end OCR framework to deal with this issue, which is shown in Fig. 2. This framework can be divided into two modules: 1) preprocess and character based recognition module; 2) formula analysis module.

The process flow of the first module is: binary the original image, conduct CC analysis to locate characters on the binary image, and then recognize the characters with the trained classifier, finally analyze the layout and text lines. The process flow of the first module is basically the same

2379-2140/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDAR.2017.27

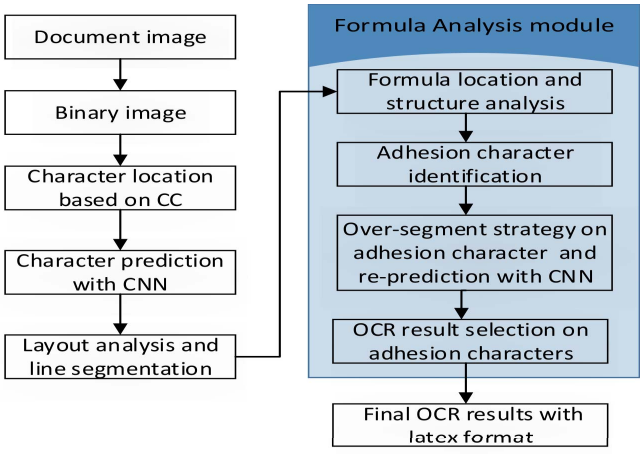

Figure 2. Flowchart of the proposed end-to-end OCR system for smartphone-captured Chinese document images containing math formulas.

to existed OCR frameworks. Our contributions are mainly reflected in the second module.

Our contributions in this paper are three-fold:

- We propose an end-to-end OCR system for smartphone-captured Chinese document images containing mathematical formulas. The proposed system can get accurate OCR results with structures for math formulas.
- We propose an over-segmentation strategy based on CC of the binary image to locate characters more precise, which improves the OCR results when the binary results of low-quality document images are not precise.
- We harvest a new benchmark database and make it publicly accessible. The dataset contains one hundred smartphone-captured Chinese document images containing math formulas, named MFR100. We type out their ground truth text with and without latex format manually. Through experiments, we have verified the effectiveness of the method. The F-measure of our method on MFR100 is up to 0.85, which is significantly better than several state-of-the-art methods.

The rest of the paper is organized as follows. Section II provides an overview of OCR related works, including the binarization, the location and recognition of text lines and characters, math formula structural analysis. Section III describes the proposed method, in which, the formula structural analysis and over-segmentation strategy on adhesive characters are the mainly introduced components in Section III-A and Section III-B separately. Section IV presents the experimental results. Finally, Section V concludes the paper.

II. Related Work

In general, a complete OCR system consists of two basic components: preprocess and recognition. Preprocess is composed of image enhancement and binarization [2], layout analysis [3], text line and character location [4]. Image enhancement and binarization eliminate the effects of uneven illumination, low contrast, various noise. Layout analysis locates text areas. Through analysing the CC on binary image, we get the text line and character locations.

There are mainly two kinds of methods in OCR systems: character recognition based methods [4], [5] and text line recognition based methods [6], [7]. Character recognition based methods locate characters by CC, and recognize each character individually. Various classifiers [8] are utilized to accomplish this task. Among these classifiers, CNN performs best. Character recognition based methods are accurate and fast when locating the characters correctly, but they highly rely on the performance of preprocess result. When the binary result is not accurate on low-quality document images, the corresponding OCR result is not satisfied.

To eliminate the dependency on preprocess and binarization results, text line recognition based methods [6], [7] are proposed. These methods locate text lines instead of characters. The text lines are considered as the recognition units to predict OCR results. LSTM based classifiers are the state-of-the-art methods on text line recognition. These methods perform better on low-quality document images than the character recognition based ones. While for structured text, such as math formulas, they cannot handle well. Their recognition speed is relatively slow than the character recognition based ones.

Text line and character location [10] are used for text line recognition methods and character recognition methods respectively. For several years, character location and recognition is a common practice in OCR, and CNN based methods are regraded as the state-of-the-art. In recent years, LSTM based text line recognition methods [7], [9], [11] perform better than character recognition methods, especially in handwritten OCR scene. The advantage of text line recognition methods lies in that they do not need rely on precise character location result, which is the bottleneck of character recognition methods on smartphone-captured low-quality document images. The shortcoming of text line recognition methods lies in that they cannot handle structural analysis of mathematical formulas, which is important for our application. Hence we choose a character recognition method based on CNN. We propose an over-segmentation strategy to split adhesive characters to improve the precision of character location result, which solves the bottleneck of character recognition based OCR methods.

There are some methods [12], [13] that embed a math formula structural analysis module into a OCR system. Some methods [13] distinguish formula lines from text lines, and conduct formula structural analysis in the selected formula lines. While in our application, formulas may be mixed with text in the same line, which adds difficulty in formula location and structural analysis. Some methods [12] solved the structural analysis of formulas that mixed with text in the same line. They perform well on the high quality binary images. While for the degraded document images, the

binary images of them are often with low qualities (split and adhesion phenomenons on the CC of binary images), which adds difficulty in MFR. We solved this issue by combining a hierarchical formula structure analysis method with the over-segmentation strategy talked above, which outperforms the state-of-the-art methods in MFR.

III. Proposed Method

In this section, we present our end-to-end OCR system for smartphone-captured Chinese document images containing mathematical formulas, as depicted in Fig. 2. The process flow on the left part in Fig. 2 is basically the same to other existed OCR systems. For most OCR systems, binarization algorithm is the first process step. Then the CC on the binary image are analyzed to locate characters. The characters are sent to the CNN classifier to get their labels and confidences. The CNN structure follows LeNet-5 proposed in [14]. We set the output unit number to 7401 and each unit represents a character. These outputs include English letters, numbers, common Chinese letters and operators. Based on the information of characters and some other features of document images, we conduct layout analysis and line segmentation. All the characters are sorted by the region, line and coordinate information. Then the OCR results are output by the order of the sorted characters.

Different from existed OCR systems, our system has two advantages: first, we can handle mathematical formula structures that mixed in Chinese text, and output the OCR results with latex format for mathematical formulas; second, we propose an over-segmentation based serialized CNN algorithm to solve the recognition problem of adhesive characters, which improves the precision of character location result. We will discuss the two advantages in detail in the following Section III-A and Section III-B.

A. Formula Structural Analysis

We discuss math formula structural analysis in this section. We concentrate our energy on elementary math formulas, whose special structures mainly include equation sets, fractions, superscripts and subscripts. The formulas are mixed in Chinese text, as shown in Fig. 1. The regions that framed in red rectangles on the left column of Fig. 1 are in special formula structures.

We utilize the feature vector (tr, lc, w, h, pl, pc) of the characters to analyze formula structures. The six features represent the top row number, the left column number, the width, the height, the predicted label and the predicted confidence of the character in turn. The mean height w_m, mean width h_m and mean predicted confidence pc_m of all characters are also utilized to analyze formula structures.

1) Analysis of Fractions: Fraction structures are detected by the following rules: identify candidate fractional lines in characters that with high aspect ratio. Retain the fractional lines that have characters above and below them in the same segment line simultaneously, and treat them as a fractional structure. The characters above the fractional line are regarded as numerators, and the characters below the fractional line are regarded as denominators.

2) Analysis of Superscripts and Subscripts: Superscripts and subscripts are detected by the vertical coordinates of each character and its nearest left neighbour.

Character c is judged as a superscript when it satisfies:

$$\frac{tr_c + h_c - tr_p}{h_p} < 0.5 \text{ and } \frac{h_c}{h_m} < 0.5.$$

Character c is judged as a subscript when it satisfies:

$$\frac{tr_c - (tr_p + h_p)}{h_p} < 0.5 \text{ and } \frac{h_c}{h_m} < 0.5.$$

In which, tr_c, h_c represent the top row number and height of character c, tr_p, h_p represent the top row number and height of the nearest left neighbour of character c.

3) Analysis of Equation Sets: The characters that be recognized as braces and whose heights are twice larger than h_m are regraded as the beginnings of equation sets. On the right side of the satisfied braces, a character that located near the central row, or a Chinese character, is regarded as the ending of an equation set. The characters whose vertical coordinates are within the vertical range between the beginning and the ending characters are regarded as the elements in the equation set. Line segmentation, fractional structure analysis, superscripts and subscripts analysis are conducted in the range of equation set. These methods are the same to the methods introduced in former sections.

B. Over-Segment Strategy on Adhesive Characters

For character recognition based OCR methods, the precision of character location is the bottleneck. When the document images are affected by low contrast, uneven illumination and various kinds of noise, the adjacent mathematical formula characters are easy to merged into one connected component. The characters in red rectangles in Fig. 1 are adhesive characters.

The Identification of Adhesive Characters

Three features of a character c are utilized for the identification of adhesive characters in mathematical formulas:

(1). The character types of the adjacent characters of character c. The characters can be divided into two groups according to their CNN prediction results: formula elements (English letters, numbers and mathematical operators) and non-formula elements (Chinese characters). We mark this formula element rate feature as $\alpha_1(c)$, which is calculated as

$$\alpha_1(c) = \frac{n_c}{N},$$

in which, N represents the neighbour number of a character, and it is a predefined parameter, n_c represents the formula element number in its neighbours. The higher value of $\alpha_1(c)$, the more likely that this character is a formula element.

(2). The confidence value of CNN prediction of character c. The confidence values of CNN prediction of adhesive

characters are tend to be lower than that of normal characters. The absolute confidence value is not a reliable feature used to distinguish adhesive characters from normal characters, for that the confidence values of normal characters in low-quality images may be lower than that of adhesive characters in high-quality images. To eliminate the effect of different qualities of images, we utilize relative confidence value as the second feature $\alpha_2(c)$, which is defined as

$$\alpha_2(c) = \frac{p_c}{p_{mean}},$$

in which, p_c represents the CNN prediction confidence of character c, and p_{mean} represents the mean confidence value of all characters in the document image. The lower value of $\alpha_2(c)$, the more likely that this character is a adjacent character in mathematical formula.

(3). The aspect ratio of character c. From Fig. 1 we get that since the CC of adhesive characters contain two or more characters, hence their aspect ratios tend to be higher than that of normal characters. We define the aspect ratio of character c as the third feature $\alpha_3(c)$, which is defined as

$$\alpha_3(c) = \frac{w_c}{h_c},$$

in which, w_c and h_c represent the width and height of the character c respectively.

We utilize $(\alpha_1(c), \alpha_2(c), \alpha_3(c))$ as the feature vector of character c, which is used for the identification of adhesive characters in mathematical formulas. A two class SVM classifier based on this feature vector is trained to distinguish adhesive characters and normal characters.

The Over-Segment Strategy

The over-segment strategy on adhesive characters is based on the candidate segment lines. The candidate segment lines are selected by vertical projection on the binary image of the adhesive character. The segment lines are selected by the following procedure:

1) Calculate the vertical projection vector on the binary image of the adhesive character;
2) Find all peaks (local maximum point) in the projection vector. For the adjacent peaks whose distance is less than one-tenth of the width of the character, retain the maximal peak, and filter out others;
3) Take the maximal value of remainder peaks as p_{max}, and filter out the peaks whose values are less than the half of p_{max};
4) Find the valleys (local minimum point) between adjacent peaks, and filter out the valleys whose values are greater than the half of the mean value of the two adjacent peaks;
5) The locations of the remainder valleys are taken as the locations of segment lines.

Fig. 3 shows the segment lines on adhesive characters. The blue triangles on figures of the bottom row in Fig. 3 show the filtered peaks, and the red dotted lines show the filtered segment lines. From the segment lines in right figure of Fig. 3, we can see that the true segment lines are selected alone with some redundant segment lines, which is called over-segmentation.

Algorithm 1 The over-segmentation based serialized CNN algorithm.

Input: The original gray value matrix of adhesive character c; the locations of candidate segment lines on the character $L_i, i \in 1, 2, \ldots, n$; the trained CNN classifier that takes formula elements (English letters, numbers and mathematical operators) as its outputs;

Output: The OCR result r of the adhesive character;

Initialize: set the start column s and end column e to 0, set r to empty, and set an empty sub-character vector v that contains the column range of the sub-characters and their predicted character labels and confidences of CNN;
for each $s \in \{0, L_i\}, i \in 1, 2, \ldots, n$ **do**
    Clear the sub-character vector v.
    for each $e \in \{L_j\}, L_j > s, j \in 1, 2, \ldots, n$ **do**
        Intercept the part of c that in the column range $[s, e]$;
        send it to CNN classifier, and record its predicted character label and confidence;
        push the information of the sub-character into vector v.
        if the predicted confidence is lower than the confidence of the tail element in v **then**
            break the inner loop;
        end if
    end for
    Find the element that with the maximum confidence in v, add its predicted label into r, and set the start column s to its right column.
end for
return OCR result r.

We propose an over-segmentation based serialized CNN algorithm to solve the recognition problem of adhesive characters. The algorithm flow is listed in Algorithm 1. In Algorithm 1, we get sub-patches based on the candidate segment lines. The recognition process starts from the first sub-patch on the left of the original patch, and continuously merge the sub-patches on the right side. A new combined patch is sent to CNN classifier and get the recognition result and confidence when a sub-patch is merged. We finish the merging process when the confidence of recognition result begins to decline, and take the combined patch that has the maximum confidence as one confirmed character. The above serialized recognition process is conducted on the rest of the unconfirmed part of the adhesive character till all the sub-patches are confirmed.

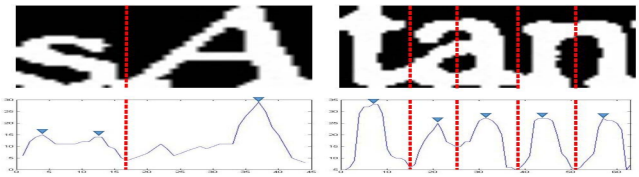

Figure 3. Examples to show the candidate segment lines in adhesive characters.

IV. Experimental Result

A. MFR100 Dataset

We harvested a diversified ground-truth dataset, which is named as MFR100 and can be downloaded at: https://github.com/zhendejianzheng1988/MFR100. There are three folders in MFR100: the folder named "image" contains 100 degraded Chinese document images that contain math formulas, the folder named "txt_without_math_structure" contains the ground truth text that without math structure information for each document image in folder "image" and the folder named "txt_with_math_structure" contains the ground truth text that with math structure information. The representations of math formula structures in MFR100 are shown in Table I. These degraded document images are collected from the daily queries of our users. The degradations include: background noise, camera-shake blur, tilt, uneven illumination, low contrast and handwritten pollution. Examples are shown in Fig. 4.

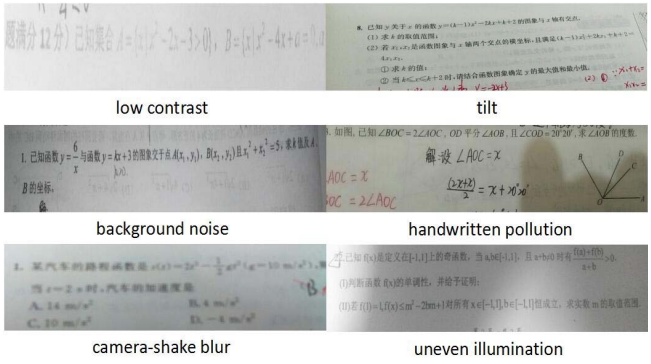

Figure 4. Examples to show the degradations of the document images in our MFR100 dataset.

Table I
THE REPRESENTATIONS OF MATH FORMULA STRUCTURES IN MFR100.

Formula structure	Text without structure	Text with structure
$\frac{a}{b}$	a b	frac_begin a frac_tag b frac_end
x^2	x 2	x super_begin 2 super_end
x_2	x 2	x sub_begin 2 sub_end
$\sqrt{a}$	a	sqrt_num a sqrt_end
$F(x)=\begin{cases} f(x) \\ g(x) \end{cases}$	F(x) f(x) g(x)	F(x)= equ_begin f(x) equ_split g(x) equ_end

B. The Process of Formula Structural Analysis

Fig. 5 shows the process of formula structural analysis discussed in Section III-A. In the image of the upper row in Fig. 5, the characters within red rectangles are the candidate fractional lines that are selected by our principles described in Section III-A1, and the characters that framed in yellow rectangles are the candidate numerators and denominators. These selected fractional lines and their numerators and denominators compose the fractional structures.

The structural analysis of equation set is shown in the image of the bottom row in Fig. 5. The character within the red rectangle is the detected brace of the equation set. The adjacent characters within the green rectangles are the detected elements in the equation set. These elements are divided into two lines according to their vertical coordinates, and each line is regraded as an equation. Additional formula structural analysis is conducted on each equation.

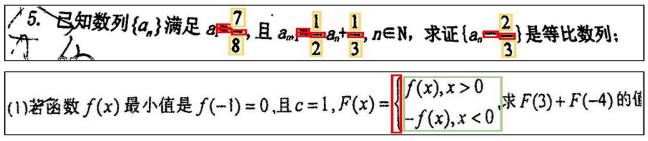

Figure 5. Examples to show mathematical structural analysis.

C. The Process of Serialized CNN Algorithm

Fig. 6 shows the process of serialized CNN algorithm to solve the recognition of adhesive characters. In this figure, the sub-patches that framed in red rectangles are unconfirmed sub-patches. Their confidences of recognition results are listed blow them. We choose the sub-patch that with the maximum confidence as one confirmed character. The confirmed character are framed in green rectangles. The recognition process finishes when all parts of the adhesive character are confirmed. We combine the results of the confirmed characters as the recognition results of the adhesive character. In step 4 of Fig. 6, the adhesive character that contains "tan" is split into three independent characters correctly.

D. Comparation with Other Methods

We compare our method with a traditional CNN based OCR method, a traditional LSTM based method and two well-known OCR softwares: Tesseract [17] and ABBYY FineReader [18]. The traditional CNN and LSTM based methods are implemented by us with our preprocessing algorithm. We report the precision, recall and F1-measure in Table II and Table III. Table II shows the values of indicators

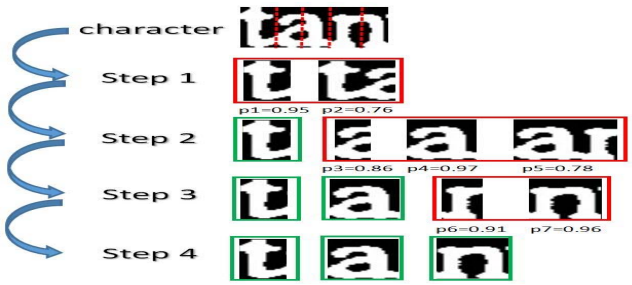

Figure 6. Examples to show the candidate segment lines in adhesive characters.

of all methods on MFR100. The values are calculated for the OCR results and ground truth text without formula structure. The proposed method performs best among these methods in the indicators listed in Table II. Although Tesseract [17] and ABBYY FineReader [18] are well-known OCR softwares, the preprocess algorithm of them cannot handle these degradations well simultaneously, so the values of indicators of them are relatively low. The traditional CNN and LSTM methods share the same preprocess algorithm with proposed method. The traditional CNN method performs bad when the character location results are not accurate. The traditional LSTM method does not rely on the accuracy of character location results, but it cannot handle formula structures well. So the results of them are not as well as the proposed method.

Since only CNN and our proposed method can handle MFR with formula structure, we report their results in Table III. Benefited from the over-segmentation strategy, our proposed method achieves better performance at character locating and hence the recognition results are more accurate.

Table II
THE EXPERIMENTAL RESULTS ON MFR100 FOR THE TEXT WITHOUT FORMULA STRUCTURE.

	Precision	Recall	F-measure
Traditional CNN	77.57	85.37	80.66
Traditional LSTM	76.18	82.19	78.62
Tesseract [17]	68.42	79.23	74.50
ABBYY FineReader [18]	75.81	81.92	76.18
Proposed	**81.95**	**85.95**	**83.44**

Table III
THE EXPERIMENTAL RESULTS ON MFR100 FOR THE TEXT WITH FORMULA STRUCTURE.

	Precision	Recall	F-measure
Traditional CNN	79.19	86.67	82.48
Proposed	**82.46**	**87.85**	**85.37**

V. CONCLUSION

In this paper, we addressed the problem of MFR in degraded Chinese document images. Specifically, we proposed an end-to-end OCR method to deal with this issue. This method is a character recognition based method using CNN, which may fail to predict accurate character locations under various degradations. To overcome this drawback, we proposed an over-segmentation strategy to split and recognize adhesive formula elements based on convolutional neural network (CNN). In addition, we take a top-down manner to iteratively split the regions into recognizable units to analyze formula structures. We conduct extensive experiments on our MFR100 dataset with various degraded Chinese document images. The experimental results validated the effectiveness and robustness of our proposed solution in comparison with state-of-the-art methods.

VI. ACKNOWLEDGMENT

This research is supported by the National Nature Science Foundation of China (No. 61602087, No. 61602089, No. 61573357, No. 61503382, No. 61403370, No. 61273267 and No. 91120303), the Fundamental Research Funds for the Central Universities under grants No. ZYGX2016J080, and China Postdoctoral Science Foundation (Grant No. 2017M610019).

REFERENCES

[1] Suzuki M, Tamari F, Fukuda R, et al. INFTY: an integrated OCR system for mathematical documents[C], Proceedings of ACM symposium on Document engineering, 2003: 95-104.

[2] Pratikakis I, Zagoris K, Barlas G, et al. ICFHR2016 Handwritten Document Image Binarization Contest (H-DIBCO 2016)[C], in ICFHR, 2016: 619-623.

[3] Grana C, Serra G, Manfredi M, et al. Layout analysis and content enrichment of digitized books[J], Multimedia Tools and Applications, 2016, 75(7): 3879-3900.

[4] Doermann D S, Tombre K. Handbook of Document Image Processing and Recognition[M], Heidelberg: Springer, 2014.

[5] Patel C, Patel A, Patel D. Optical character recognition by open source OCR tool tesseract: A case study[J], International Journal of Computer Applications, 2012, 55(10).

[6] Breuel T M, Ul-Hasan A, Al-Azawi M A, et al. High-performance OCR for printed English and Fraktur using LSTM networks[C], in ICDAR, 2013: 683-687.

[7] Asad F, Ul-Hasan A, Shafait F, et al. High Performance OCR for Camera-Captured Blurred Documents with LSTM Networks[C], Document Analysis Systems (DAS), 2016 12th IAPR Workshop on. IEEE, 2016: 7-12.

[8] Fedorovici L O, Precup R E, David R C. GSACBased Training of Convolutional Neural Networks for OCR Applications[M], Computational Intelligence Systems in Industrial Engineering. Atlantis Press, 2012: 481-504.

[9] Messina R, Louradour J. Segmentation-free handwritten Chinese text recognition with LSTM-RNN[C], in ICDAR, 2015: 171-175.

[10] Nguyen V D, Chow Y W, Susilo W. A CAPTCHA scheme based on the identification of character locations[C], International Conference on Information Security Practice and Experience. Springer International Publishing, 2014: 60-74.

[11] Yousefi M R, Soheili M R, Breuel T M, et al. Binarization-free OCR for historical documents using LSTM networks[C], in ICDAR, 2015: 1121-1125.

[12] Furukori F, Yamazaki S, Miyagishi T, et al. An OCR System with OCRopus for Scientific Documents Containing Mathematical Formulas[C], in ICDAR, 2013: 1175-1179.

[13] Yamazaki S, Furukori F, Zhao Q, et al. Embedding a mathematical OCR module into OCRopus[C], in ICDAR, 2011: 880-884.

[14] LeCun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE, 1998, 86(11): 2278-2324.

[15] Lin X, Gao L, Tang Z, et al. Mathematical formula identification in PDF documents[C], in ICDAR, 2011: 1419-1423.

[16] Mouchre H, Viard-Gaudin C, Zanibbi R, et al. ICFHR 2016 CROHME: Competition on Recognition of Online Handwritten Mathematical Expressions[C], in ICFHR. 2016.

[17] Tesseract OCR (2016). https://github.com/tesseract-ocr.

[18] Abbyy OCR (2016). https://www.abbyy.com/.