The Design and Implementation of Chinese-Uyghur Printed Document Retrieval System Based on OCR

Eliyas Suleyman[1,a], Abdusalam Dawut[1], Palidan Tuerxun[1], Askar Hamdulla[2,b,]*

[1]School of software, Xinjiang University Urumqi, China

[2]Institute of Information Science and Engineering Xinjiang University Urumqi, China

[a]email: eliyas007@163.com, [b]email: askar@xju.edu.cn

Keywords: Document Retrieval; Key Word Localization; Document Segmentation

Abstract: This paper focuses on the overall framework and functions of Chinese-Uyghur printed document retrieval system based on Optical Character Recognition (OCR) system. This system mainly fulfills a Chinese and Uyghur document retrieval by inputting the keyword. The proposed system consists three steps to localize the key word in a document image. Firstly the document image is segmented to basic units which is words. Secondly the entire content of the document image is recognized by using OCR. Then the key word localization is applied. Due to the proposed system is based on OCR system, hence, the precision of localization of a key word is highly depend on the accuracy of the applied OCR system.

1. Introduction

As the society is developing rapidly, the digitized document image is most prevalent medium for people to use in work, study or daily life. With the proliferation of digital libraries and the promise of paper-less office, an increasing number of document images of different qualities are being scanned and archived[1]. Regarding the format of these doument images and make it easy to access or search it effectively, document retrieval for these digitized documents are significant problem to deal with. Unlike the text format document, the representation of a document image in computer is a 3-channel matrix. Hence, the localization of the inputted key word is not as easy as the text retrieval. To solve this problem, the implementation and design for the document retrieval system is crucial.

In this paper, a Chinese-Uyghur printed document image retrieval system based on OCR is proposed. The whole structure of the proposed system is as follows. Firstly, the content of the document image is recognized, by using an Optical Character Recognition system, as a string. Secondly, the projection based method is used to segment the whole document into basic units which is the single word image. Then, the location of the key word in image is found by using the content string and segmented basic units. At last, the rectangle is drawn to the spotted word's coordinate at the original image. The rest of the paper is organized as follows. In section 2, the methodology of the presented system is described in detail, section 3 analyzes and discusses the performance of the proposed system, In section 4 gave the brief conclusion of the proposed work. Fig.1 shows entire framework and main steps of proposed method.

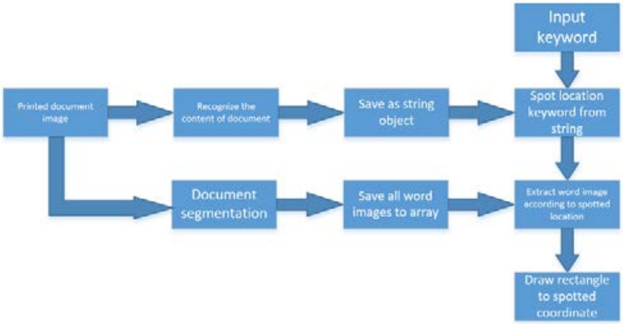

Input keyword
Printed document image
Recognize the content of document
Save as string object
Spot location keyword from string
Document segmentation
Save all word images to array
Extract word image according to spotted location
Draw rectangle to spotted coordinate

Figure 1 Framework and main steps

DOI: 10.25236/icemit.2019.032

2. Methodology

2.1 Recognition of document image

Before the system locates the key word in document image, the content of the document image should be recognized for subsequent stage's work. In order to turn the content of the Chinese and Uyghur documents into readble text format, the third party OCR system is used. After turning the content of document image to text format, the result is stored as string object. In this work, the accuracy of the OCR system is considered as 100%. Hence, the recognition error can be omitted.

2.2 Segmentation of document image

Before the segmentation, the basic pre-processing technique is used such as gray scaling, binarization, dilation and blurring which is noise removal and smoothing the document image. First of all, the document image is turned to gray scale image. Commonly, regular a color image contains three channels, each channel stores the 2-dimentional array which represents red, green and blue[4]. The gray scale image is determined by calculating a weighted sum of three channels components for every pixel of color image. The equation is shown below:

$$S = 0.2989 \times R + 0.587 \times G + 0.1140 \times B \quad (1)$$

where S is the sum of three weighted channel components, R, G and B represents the red, green, blue channels of color image. Then the gray image is binarized using Otsu thresholding method[2]. The equation (2) represents the Otsu method.

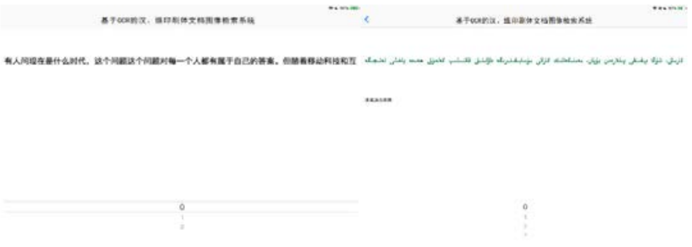
$$\sigma_\omega^2(t) = \omega_0(t)\sigma_0^2(t) + \omega_1(t)\sigma_1^2(t) \qquad (2)$$

weights ω_0 and ω_1 are probabilities of two classes, which refers text lines and the background, separated by a threshold t, and σ_0^2 and σ_1^2 and variance of these two classes. Then the binarized image is dilated.

In the purpose of localizing the key word in image, the coordinate of each single word in document image must have known by the system. Thus, raw document image should be segmented. The proposed system would segmented the document lines, then lines are segmented into words which is the basic unit of the document image. In the respect of text line segmentation, due to document style is printed, thus, the simple projection based segmentation method is used. Firstly, document image is binarized using Otsu thresholding mechanism[2]. Then the horizontal projection profile is counted using binarized image. Horizontal projection profile is a vector that contains sum black pixels of each row in binary image. Projection of a binary image is calculated as the equation (1) shown below:

$$H(j) = \sum_{i=1}^{n} p(i,j)\,, j = 1,2,m \qquad (3)$$

where H indicates the projection profile vector, where i and j refers to row and column of a binary image and p refers to currently visited pixel[6]. Then using the gap between each line to segment document image. Fig.3 shows the segmentation result Plot of projection profile is shown in Fig.2.



基于OCR的汉、维印刷体文档图像检索系统

有人问现在是什么时代，这个问题这个问题对每一个人都有属于自己的答案，但随着移动科技和互

基于OCR的汉、维印刷体文档图像检索系统

(a). Segmented Chinese text lines (b). Segmented Uyghur text lines

Figure 2 Line segmentation result

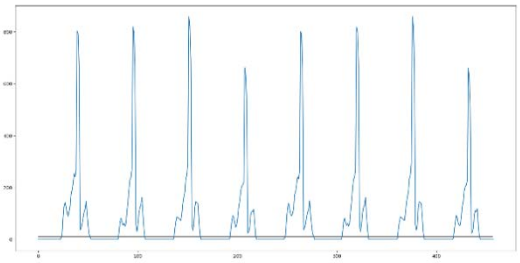
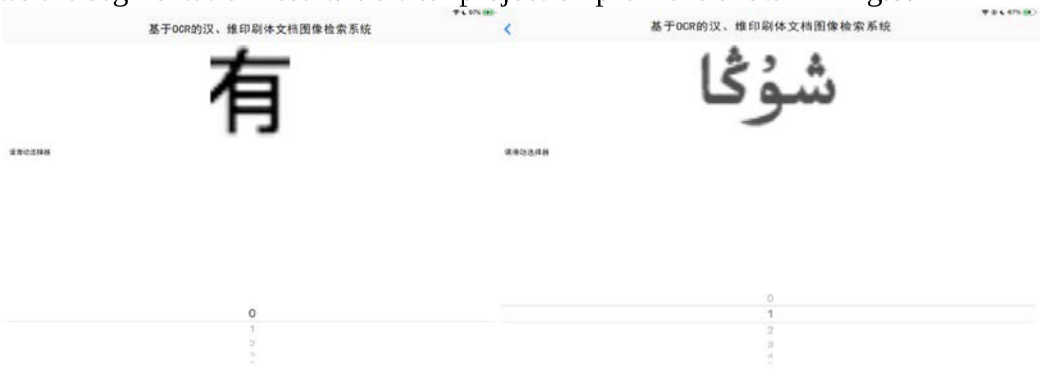


Figure 3 Horizontal projection profile

Then word segmentation is performed to segmented text line images[5]. The mechanism is almost same to the text line segmentation process, due to the high regularity of gaps that is between each neighboring words. Hence, projection profile is calculated by same equation (3), but calculation direction is vertical which is different from line segmentation. After counting the vertical projection profile, we calculate the maximum gap between each connected components in line image, then define a threshold to segment these words in line image. Through the entire process of segmentation, ever time the system segments a word, the segmented word image would be stored in an array. Then, coordinate of each segmented word would be stored in a object that has 4 properties which is x, y, width and height, which means the coordinate at original image and size of segmented word image. Fig.4 shows the segmentation result. Vertical projection profile is shown in Fig.5.


基于OCR的汉、维印刷体文档图像检索系统
有
基于OCR的汉、维印刷体文档图像检索系统
شۇڭا
0
1
2
0
1
2

(a). Segmented Chinese word

(b). Segmented Uughur segmented word

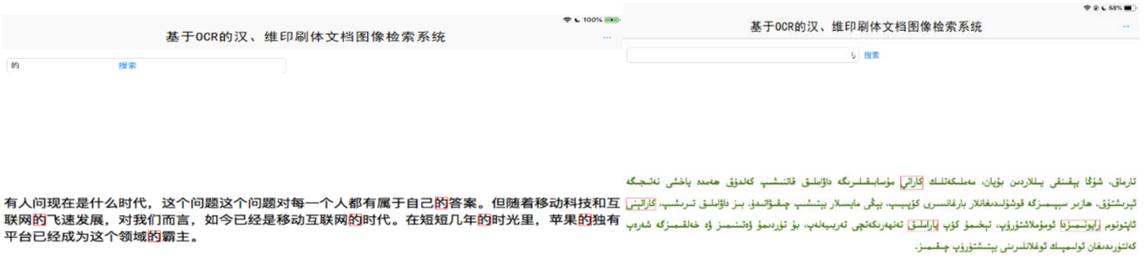
Figure 4 Word segmentation result

Figure 5 Vertical projection profile

After the process of segment document to basic units, the object that contains word images is stored in one dimension array which would be used for locating the key word in original document image.

2.3 Matching and Locating

Proposed system would use the previously segmented word images and the OCR result string to spot the location of key word in document image. At the beginning, keyword's location is searched in string and spotted location is stored. Then, word image object is extracted according to stored location. Finally, use extracted word object's coordinate information and use it to draw a rectangle to original image. Localization result is shown in Fig.6.


100%
基于OCR的汉、维印刷体文档图像检索系统
的
搜索
68%
基于OCR的汉、维印刷体文档图像检索系统
搜索

有人问现在是什么时代，这个问题这个问题对每一个人都有属于自己的答案。但随着移动科技和互联网的飞速发展，对我们而言，如今已经是移动互联网的时代。在短短几年的时光里，苹果的独有平台已经成为这个领域的霸主。

تارماق، شۇڭا يېقىنقى يىللاردىن بۇيان، مەملىكەتلىك كاراتې مۇسابىقىلىرىگە داۋاملىق قاتنىشىپ كەلدۇق ھەمدە ياخشى نەتىجىگە
ئېرىشتۇق. ھازىر سېپىمىزگە قوشۇلىدىغانلار بارغانسىرى كۆپىيىپ، يېڭى مابىسلار يېتىشىپ چىقىۋاتىدۇ. بىز داۋاملىق تىرىشىپ، كاراتېنى
ئاپتونوم رايونىمىزدا ئومۇملاشتۇرۇپ، ئىخچام كۆپ پارامىلىق تەنھەرىكەتچى تەربىيەلەپ، بۇ تۈردىمۇ ۋەتىنىمىز ۋە خەلقىمىزگە شەرەپ
كەلتۈرىدىغان ئولىمپىك ئوغلانلىرىنى يېتىشتۈرۈپ چىقىمىز.

(a). Localization of Chinese document (b). Localization of Uyghur document

Figure 6 Localization result

3. Analysis

According to proposed system is based on OCR, thus, the accuracy of applied OCR system would influence the precision of the information retrieval quality directly. In this work, we assume that the correctness of utilized OCR system's accuracy is 100% which mean it can recognize all content in the document image. Therefore, if the recognition system which is OCR make error or is not able to recognize full content of document image, then it would directly cause the error for whole system.

When all the content in document image is correctly recognized by the OCR system (the recognition rate is 100%), recognition result in the document image would be stored in a string object. Then, system would use this string object to search keyword. For example, a plain text Chinese printed document image has 20 characters, which may contain a word or punctuation. We predict the recognition system will correctly recognize the 20 characters. Then, document image would be segment to 20 pieces. This means that content string's length and word image array's length is equal. This makes sure that the process of locating or spotting keyword perform precisely.

When the text recognizer is not able to recognize the full content of document image, that is to say, the correct rate of optical character recognizer is not 100%. This would cause the positioning of the keyword will be wrong, which means if the OCR system makes a slight error then it would make whole document retrieval system malfunction. For instance, there are 20 characters in the original document image, but when the OCR recognition system outputs a result that is missing or redundant information, the keyword positioning will be inaccurate. Even though system is able to find the keyword from result string, due to the inaccurately located position of keyword may draw a rectangle at a wrong place in original document image.

In summary, there are advantages and disadvantages in the proposed document retrieval system. On the one hand, if the OCR system is promising and strong enough to detect and recognize full content of given document image, then the proposed image is able to be retrieved. On the other hand, if OCR system could not make sure the correctness of recognition result, the system's performance would be influence dramatically.

4. Conclusion

This paper presents an easy-to-implement Chinese-Uyghur document retrieval system based on Optical Character Recognition system. The proposed system shows that if the recognition rate of OCR system is strong enough, system's performance is promising. However, if the OCR system's accuracy is not acceptable, the proposed system's performance would decline accordingly. Hence, our future work will be focused on OCR-free document retrieval system.

Acknowledgements

This work is supported by Natural Science Foundation of China (NSFC) under Grant numbers of

61461049 and 61662076.

References

[1] Lu S, Li L, Tan CL. Document image retrieval through word shape coding[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2008, 30(11):1913.

[2] Ohtsu N. A Threshold Selection Method from Gray-Level Histograms[J]. IEEE Trans.syst.man.cybern, 1979, 9(1):62-66.

[3] A Abliz, W Simayi, K Moydin, A Hamdulla. A Survey on Methods for Basic Unit Segmentation in Off-line Handwritten Text Recognition [J]. International Journal of Future Generation Communication and Networking. 2016(9):137-152

[4] Ptak, R., Żygadło, B., & Unold, O. Projection–Based Text Line Segmentation with a Variable Threshold[J], International Journal

[5] Al-Dmour, Ayman, and R. A. Zitar. "Word Extraction from Arabic Handwritten Documents Based on Statistical Measures." International Review on Computers & Software 11.5(2016).

[6] Papavassiliou, Vassilis, et al. Handwritten document image segmentation into text lines and words. Pattern Recognition43.1 (2010):369-377.