High Performance Chinese/English Mixed OCR with Character Level Language Identification

Kai Wang
Institute of Machine Intelligence, Nankai University, Tianjin 300071, P. R. China
wangkai_nkimi@yahoo.com.cn

Jianming Jin
HP Labs China, Beijing, 100084, P. R. China
jian-ming.jin@hp.com

Qingren Wang
Institute of Machine Intelligence, Nankai University, Tianjin 300071, P. R. China
qrwang@expervison.com.cn

Abstract

Currently, there have been several high performance OCR products for Chinese or for English. However, no one OCR technique can be simultaneously fit for both the English and the Chinese due to the large differences between Chinese and English. On the other hand, Chinese/English mixed document increases drastically with the globalization, so it is rather important to study the Chinese/English mixed document processing. Obviously, the key problem to resolve is how to split the mixed document into two parts: Chinese part and English part, so that the different OCR techniques can be applied to different parts. To further improve the previous system performance, a novel Chinese/English split algorithm based on global information is proposed and a rule for language identification is achieved by Bayesian formula. Experiment shows, the system error rate drops from 1.52% to 0.87% on magazine samples and from 1.32% to 0.75% on book samples, more than 2/5 of errors are excluded, which provides an experimental support for our research work.

1. Introduction

Up to present, optical character recognition (OCR) has been a fully-fledged academic subject, and the commercial OCR products have shown high performance in practical application. In current market, the main Western OCR products include ScanSoft OmniPage OCR, ABBYY FineReader OCR, Expervision OCR and IRIS OCR, while the main Chinese OCR products include Tsinghua TH-OCR, HanWang OCR and Beixin OCR, all of which show high performance either on the English document image processing or on the Chinese document image processing. However, no one OCR technique can be simultaneously fit for both the English and the Chinese due to the following differences between Chinese and English:

1) The difference on the character segmentation. Most of Chinese characters consist of two or more connected components, while the English characters only include one component except “i” and “j”. On the other hand, it is a common phenomenon that the adjacent English characters are touched, while a gap usually exists between the adjacent Chinese characters. Therefore, the key problem considered during the Chinese character segmentation is how to merge multiple components into one character, while that during the English character segmentation is how to split multiple touched characters into a few single characters.

2) The difference on the size of the font set and the character set. The size of the font set is small but the size of the character set is large for Chinese, while it is just on the contrary for English. Therefore, the key problem considered during the Chinese character recognition is how to distinguish so large number of different characters. But for English, the key is to distinguish different fonts and make use of font information to improve the character recognition accuracy.

3) The difference on the complexity of the character structure. The structure of Chinese characters is generally much more complicated than that of English characters, so that the large difference certainly exists between the recognition technique for Chinese and that for English.

On the other hand, Chinese/English mixed document increases drastically with the globalization, so it is rather important to study the Chinese/English mixed document processing. As mentioned above, there have been several high performance OCR

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.14

IEEE computer society

products for Chinese or for English. In such a case, the key problem is how to split the mixed document into two parts: Chinese part and English part, so that the different OCR techniques can be applied to different parts.

Currently, there have been several studies concerned on this topic (see, for example literature [1]-[5]), and all of which employed similar algorithms which consist of the following two steps:

1) Initial segmentation is conducted by connected component analysis or vertical projection. As a result, some blocks are achieved, each of which is a Chinese character, a component of a Chinese character, an English character or several touched English characters.

2) Merge or split initial blocks by the structure analysis or the recognition confidence.

The algorithm adopted by previous works is straightforward, simple, practical and high-performance, which is also the foundation of our work. The main difference between our work and previous works is that the global information is employed by us to further improve the performance, which will be presented in detail in the following sections.

The rest of this paper is organized as follows: The related works are introduced in Section 2. Section 3 presents a novel algorithm for the Chinese/English split. The simulation is reported in Section 4, which provides an experimental support for our work. Summary and conclusion are given in Section 5.

2. The flow of the proposed system

The flow of our system is shown in Figure 1, which consists of three modules: pre-processing, bilingual OCR and post-processing. Each module in which will be described in the following subsections.

2.1. Pre-processing

The following operations are applied in this module: a) Load image into memory; b) Binarize for converting color image to B/W image; c) Smooth the image to filter noise; c) Deskew to straighten the image; d) Auto-rotation to correct the image orientation; e) Locate to separate the image into text region, image region, table region, etc; f) Split the text region into several lines; g) Do the coarse segmentation by vertical projection and connected component analysis mixed method.

After the pre-processing, some blocks are acquired, each of which is:

a) A Chinese character;

b) Or a component of a Chinese character;

c) Or a English character;

d) Or several touched English characters.

Please refer to literature [1]-[5] for more detail.

2.2. Post-processing

The following operations are applied in this module:

a) Performance evaluation: To evaluate whether the recognition result is acceptable. The bilingual OCR processing will be redone if the result is not good enough. This is a **high-level** control to improve the performance of the whole system.

b) Layout restoring: To make the processed layout be consistent with the original layout shown in the document image.

c) Result output: To export the processing result to a file with the user-specified format.

2.3. Bilingual OCR

The module is the kernel of the Chinese/English mixed OCR system, which consists of the following steps (as is shown in Figure 1):

a) Chinese/English split: To split each text line into two parts: Chinese part and English part.

b) Chinese/English character segmentation: To acquire the image of each Chinese/English character.

c) Chinese/English OCR engine: To achieve the recognition result for each character. Two mature OCR techniques are adopted in this step: Expervision OCR for English recognition and Beixin OCR for Chinese recognition.

d) English segmentation evaluation: To evaluate the quality of English character segmentation according to the recognition result. This is an **internal control** for the segmentation module to improve the performance of the English character segmentation.

3. The algorithm for Chinese/English split

3.1. The center-equidistance property based region partition

As is shown in Figure 2, when center point of each Chinese character is projected onto a horizontal line, it can be seen that the distance between any two adjacent points is equivalent, which is called *the center-equidistance property*, and the corresponding distance is called *period*.

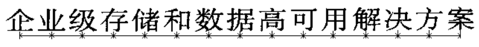

Figure 2. The center-equidistance property of Chinese characters

On the other hand, the center-equidistance property also exists in some English fonts, which are called mono-space fonts, such as Courier New font shown in Figure 3.

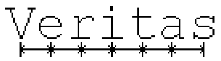

Figure 3. The center-equidistance property of English characters

Chinese/English split method proposed in this paper includes the following three steps.

a) The extraction of the center-equidistance property.

One or two period (if English characters with mono-space font exist in the text line) is achieved, and the text line is divided into several regions according to the period (as shown in Figure 4(b)).

b) Region partition.

For the situation of the number of the continuous Chinese characters being less than three, the center-equidistance property cannot be extracted (such as "在" and "还是" in Figure 4(b)). In such a case, a period based Chinese search algorithm is employed to find back the lost Chinese characters. Several regions are generated in this step, each of which is either Chinese or English (as shown in Figure 4(c)).

c) Region based language identification.

For each region generated in (b), a language identification rule is applied to distinguish whether a region is Chinese one or English one.

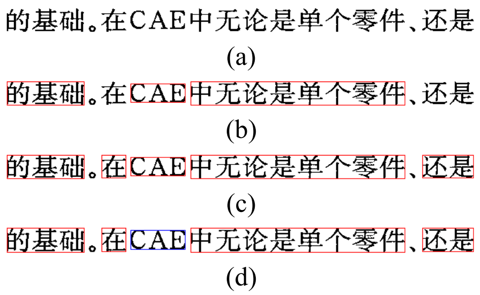
的基础。在CAE中无论是单个零件、还是
(a)
的基础。在CAE中无论是单个零件、还是
(b)
的基础。在CAE中无论是单个零件、还是
(c)
的基础。在CAE中无论是单个零件、还是
(d)

Figure 4. An example for Chinese/English split (a) Original image(b) The extraction of the center-equidistance property (c) Region partition (d) Region based language identification

The first two steps have been described in literature [6]. Here we only discuss how to conduct the region based language identification.

3.2. The rule for region based language identification

A region based language identification rule is inducted in this subsection, which shows a higher theoretical performance than single character based method.

Without loss of generality, the method based on the fisher classifier proposed in literature [7] is used for the language identification on single character. For the sake of mathematical definition, let us introduce some symbols first.

c the accuracy of the language identification on single Chinese character

e the accuracy of the language identification on single English character

n the number of characters in a to-be-identified region

m the number of characters be identified as Chinese(that is, the number of characters be identified as English is $n - m$)

CR a Chinese region

ER an English region

$P(.\mid m)$ the probability of the region be CR or ER when the number of characters be identified as Chinese is m

$P(m\mid .)$ the probability of m characters be identified as Chinese when the region is CR or ER

$P(.)$ the probability of a region be CR or ER

a the probability of a region be CR (that is, the probability of a region be ER is $1 - a$)

According to Bayesian formula, we have

$$P(CR\mid m)=\frac{P(m\mid CR)\cdot P(CR)}{P(m\mid CR)\cdot P(CR)+P(m\mid ER)\cdot P(ER)} \quad (1)$$

where

$$P(m\mid CR)=C_n^m\cdot c^m\cdot(1-c)^{(n-m)} \quad (2)$$

$$P(m\mid ER)=C_n^m\cdot(1-e)^m\cdot e^{(n-m)} \quad (3)$$

By formula (1), (2) and (3), it follows that

$$P(CR\mid m)=\frac{c^m\cdot(1-c)^{(n-m)}\cdot a}{c^m\cdot(1-c)^{(n-m)}\cdot a+(1-e)^m\cdot e^{(n-m)}\cdot(1-a)} \quad (4)$$

According to the result given in literature [7], c is 0.971 and e is 0.966. On the other hand, the probability of an unknown region being Chinese or English should be equal, thus we have

$$P(CR\mid m)=\frac{1}{1+(\frac{0.034}{0.971})^m\cdot(\frac{0.966}{0.029})^{(n-m)}} \quad (5)$$

$$\geq\frac{1}{1+(\frac{0.034}{0.971})^m\cdot(\frac{0.971}{0.017})^{(n-m)}}$$

$$=\frac{1}{1+(\frac{0.034}{0.971})^{(2m-n)}\cdot 2^{(n-m)}}$$

Under the condition of $\begin{cases} n-m \le 2m-n \\ 2m-n \ge 3 \end{cases}$, we have

$$P(CR \mid m) \ge \frac{1}{1+(\frac{0.068}{0.971})^{(2m-n)}} \ge 99.96\% \tag{6}$$

Similarly,

$$P(ER \mid m) = \frac{1}{1+(\frac{0.029}{0.966})^{(n-m)} \cdot (\frac{0.971}{0.034})^{m}}$$

$$\ge MAX(\frac{1}{1+(\frac{0.034}{0.971})^{(n-m)} \cdot (\frac{0.971}{0.034})^{m}}, \frac{1}{1+(\frac{0.029}{0.966})^{(n-m)} \cdot (\frac{0.966}{0.029})^{m}})$$

$$= \frac{1}{1+(\frac{0.029}{0.966})^{n-2m}} \tag{7}$$

Under the condition of $n-2m \ge 2$, we have

$$P(ER \mid m) \ge \frac{1}{1+(\frac{0.029}{0.966})^{2}} \ge 99.91\% \tag{8}$$

By formula (6) and (8), the following language identification rule can be achieved.

a) Under the condition of $\begin{cases} 3m \ge 2n \\ 2m-n \ge 3 \end{cases}$ (that is, $m \ge \max\left(\frac{2n}{3}, \frac{n}{2}+\frac{3}{2}\right)$), the region should be identified as Chinese.

b) Under the condition of $n-2m \ge 2$ (that is, $m \le \frac{n}{2}-1$), the region should be identified as English.

c) Otherwise, it is probably a mixed region. In such a case, the language identification on single character (just like what have been done in literature [7]) is applied.

4. Experimental study

To verify the validity of our research work, the following conditions were considered during the generation of testing samples.

a) Magazine samples and book samples were considered.

b) The scan resolution was set to be 300DPI or 400DPI.

c) The ratio of English characters is from 10% to 90%.

Two systems were implemented for the experimental study.

System I: A Chinese/English OCR system constructed by the language identification method proposed in this paper.

System II: A Chinese/English OCR system constructed by the language identification method proposed in literature [1]-[5].

The only difference between system I and system II is the difference of the language identification methods.

System accuracy testing results are shown in Table 1 and Table 2, where Error I and Error II represent the recognition error number of Chinese and English respectively. Experiment shows, the system error rate drops from 1.52% to 0.87% on magazine samples and from 1.32% to 0.75% on book samples, more than 2/5 of errors are excluded, which provides an experimental support for our research work.

Table 1. System accuracy on magazine samples

System	Character Number	Error I	Error II	Recognition Error Rate
I	108239	534	402	0.87%
II	108239	677	962	1.52%

Table 2. System accuracy on book samples

System	Character Number	Error I	Error II	Recognition Error Rate
I	135825	478	530	0.75%
II	135825	568	1215	1.32%

5. Conclusion

The Chinese/English mixed document processing is studied in this paper. Under the case of Chinese/English OCR being mature, the key problem is how to split the document into two parts: Chinese part and English part, so that the different OCR techniques can be applied to different parts. To further improve the previous system performance, a novel Chinese/English split algorithm based on global information is proposed and a rule for language identification is achieved by Bayesian formula. Experiment shows, the system error rate drops from 1.52% to 0.87% on magazine samples and from 1.32% to 0.75% on book samples, more than 2/5 of errors are excluded, which provides an experimental support for our research work.

Only Chinese/English mixed document processing is studied in this paper. But in fact, the system flow can also be extended to the processing of other Western language/Eastern language mixed document. On the other hand, the language identification method proposed in this paper can also be applied to the handwritten document.

6. References

[1] Hong Guo, Xiaoqing Ding, Zhong Zhang, and Fanxia Guo, “Realization of a high-performance bilingual Chinese-English OCR system”, In: Mary Kavanaugh, Penny Storms eds. ICDAR’95: Third International Conference on Document Analysis and Recognition. Los Alamitos, California: IEEE Computer Society Press, 1995, pp. 978-981.

[2] Zhidan Feng, and Qiang Huo, “Confidence guided progressive search and fast match techniques for high performance Chinese/English OCR”, In: R. Kasturi, D. Laurendeau, C. Suen eds. ICPR’02: 16th International Conference on Pattern Recognition. Los Alamitos, California: IEEE Computer Society Press, 2002, vol. III, pp. 89-92.

[3] Qiang Huo, and Zhidan Feng, “Improving Chinese/English OCR performance by using MCE-based character-pair modeling and negative training”, In: A. Antonacopoulos ed. ICDAR’03: Seventh International Conference on Document Analysis and Recognition. Los Alamitos, California: IEEE Computer Society Press, 2003, pp. 364-368.

[4] Jianming Jin, and Qingren Wang, “Research on multi-language OCR systems integration”, Journal of Software, 2002, vol. 13(supplement), pp. 225-230(in Chinese).

[5] Wumo Pan, Jianming Jin, Guangshun Shi, and Qingren Wang, “A system for automatic Chinese business card recognition”, In: A. Antonacopoulos ed. ICDAR’01: Seventh International Conference on Document Analysis and Recognition. Los Alamitos, California: IEEE Computer Society Press, 2001, pp. 1138-1141.

[6] Kai Wang, Jianming Jin, Wumo Pan, Guangshun Shi, and Qingren Wang, “Mixed Chinese/English document auto-processing based on the periodicity”, In: Proceedings of 2004 International Conference on Machine Learning and Cybernetics. Shanghai, China: IEEE Computer Society Press, 2004, vol. 6, pp. 3616-3619.

[7] Yefeng Zheng, Changsong Liu, and Xiaoqing Ding, “Single character type identification”, In: Paul B. Kantor, Tapas Kanungo, Jiangying Zhou eds. Proceedings of SPIE Document Recognition and Retrieval IX. Bellingham, Washington, USA: SPIE – the International Society for Optical Engineering, 2002, vol. 4670, pp. 49-56.

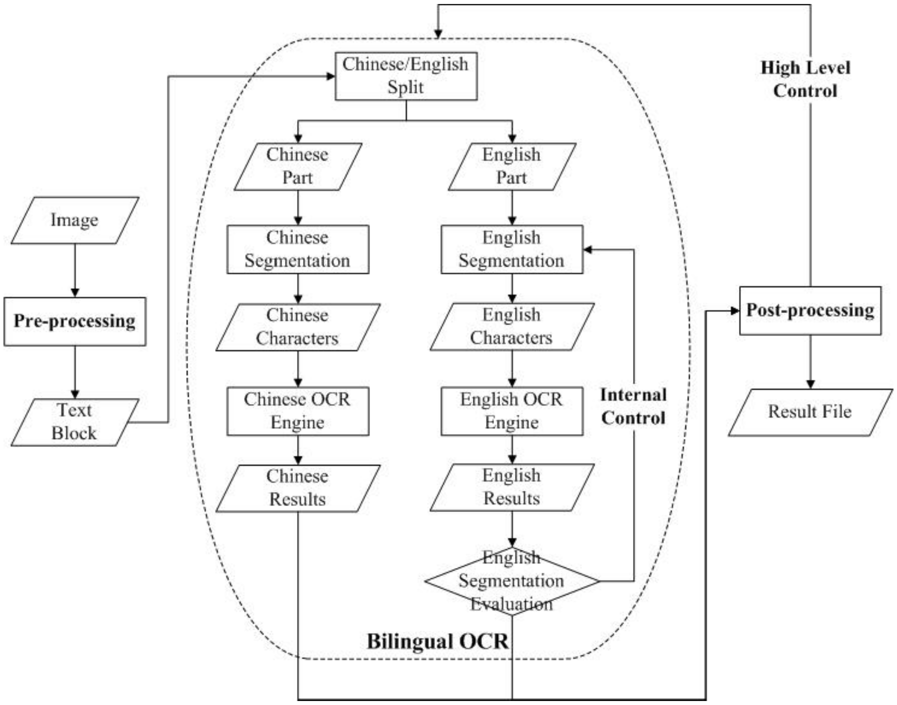

Figure 1. The flow of Chinese/English mixed OCR system