MVA '92 IAPR Workshop on Machine Vision Applications Dec. 7-9,1992, Tokyo

POSTPROCESSING FOR CHARACTER RECOGNITION USING KEYWORD INFORMATION

Hisao Niwa, Kazuhiro Kayashima, Yasuharu Shimeki
Intelligent Electronics Laboratory, Central Research Laboratories,
Matsushita Electric Industrial Co.,Ltd.

3-1-1, Yagumo-Nakamachi, Moriguchi
Osaka 570 Japan

ABSTRACT

We propose a new error correction method of postprocessing for character recognition which has good performance, even if the recognition accuracy is low.

In conventional postprocessing, knowledge of grammar and vocabulary is used. However, in the case of low recognition accuracy, that postprocessings can not perform good correction because there are many candidates for the character string. Our method not only makes use of the knowledge of grammar and vocabulary, but also the knowledge of content in the document. In this method first we automatically extract keywords using Zipf's law. Then we correct characters using those extracted keywords.

We experimented on postprocessing for character recognition using keyword information. We show that this extraction of keywords works very well, and we show that the recognition accuracy rises. When the recognition accuracy before postprocessing is greater than 90%, the restoration rate (the ratio of the number of corrected character to that of wrong recognized character) is higer than 70%.

INTRODUCTION

Optical Character Recognition is very useful during database input. The Japanese language has several thousand characters, and many of the characters have similar shapes. If the document image is noisy and unclear, then the accuracy of recognition processing for each character is low. In conventional postprocessing, though the knowledge of grammar and vocabulary is used, in the case of low recognition accuracy, such postprocessing can not perform good correction because there are so many candidates for the correct string..

On the other hand, humans can easily read such noisy documents. When reading documents, humans make use not only of knowledge of grammar and vocabulary, but also the contents. So, a human can choose the correct character string easily. Likewise, in our new method we use keyword information (that has automatically been extracted from the document) as its content.

FLOW OF POSTPROCESSING FOR CHARACTER RECOGNITION

Using knowledge of language, postprocessing for character recognition corrects errors that occured in the character recognition unit. The outputs of the character recognition unit have several character candidates. Postprocessing determines the best combination of characters from those character candidates.

We propose a new error correction method shown in Fig.1. The character recognition unit recognizes each character pattern one by one. It outputs N character candidates (first reliable character candidate to Nth reliable character candidate) and its reliability evaluations which show an estimate of correctness for the candidates. In our postprocessing method, correct strings are selected from this set of character candidates using the knowledge of grammar, vocabulary, and keywords which are extracted from the document. In our method there are five parts as described below.

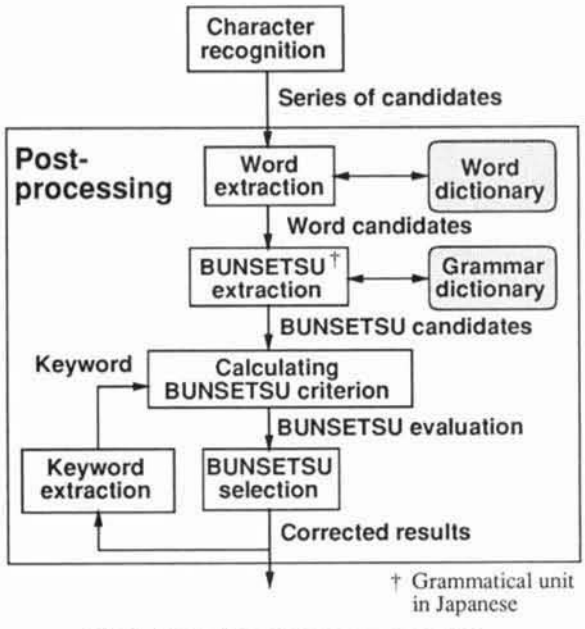

Fig.1 block diagram of postprocessing for character recognition

(1) Word extraction which determines probable words sets from character candidates by using a word dictionary. It outputs the selected word candidates.
(2) BUNSETSU extraction which extracts BUNSETSU from the word candidates by using a grammar dictionary. It outputs BUNSETSU candidates. (BUNSETSU is grammatical unit in Japanese.)
(3) Calculating BUNSETSU evaluation which evaluates BUNSETSU candidates using a character's reliability evaluation, a word's length, a word's frequency, a word's part of speech, a length of BUNSETSU, and keyword information.
(4) BUNSETSU selection which selects the most reliable BUNSETSU based on BUNSETSU evaluation.
(5) Keyword extraction automatically extracts keywords from corrected character strings.

In the following sections, we show why it is useful to use automatically extracted keyword.

POSTPROCESSINGUSING AUTOMATICALLY EXTRACTED KEYWORD INFORMATION

In our proposed method, keywords (which serve as the document content) are automatically extracted, and incorrectly recognized characters are identified by using extracted keywords. The method consists of following process steps.
(1) Corrected character strings are selected from character candidates using the current set of keyword information.
(2) Additional keywords are extracted from these corrected character strings.
(3) Using this enlarged set of keyword information, BUNSETSU candidates are evaluated again and corrected strings are selected.

ZIPF'S LAW

The keyword extraction is based on Zipf's law. Zipf's law is the experimental law about a word frequency in an arbitrary document. Zipf's law consists of two laws.
(1) First law: In the case of high frequency words, there is a frequency f and rank r that are related to each other according to the following expression.

$$f \cdot r = c \qquad (1)$$

The frequency f and rank r are inversely proportional. c is constant that is fixed with the document.
(2) Second law: In the case of low frequency words, their frequency is expressed as

$$\frac{I_1}{I_f} = \frac{f(f+1)}{2} \qquad (2)$$

where I_x is the number of the words whose frequency is x in the document. Second law shows that when a word has a low frequency, there are several words which have same frequency. Fig.2 demonstrates Zipf's law. The solid line in Fig.2 is a word's frequency in a document. The left area in the figure shows the First law, and the right area inside shows the Second law.

Generally, words that are inside the boundary area (the circle area in Fig.2) lying between First law and Second law, are the keywords that we select to describe the contents of the document. The words having very high frequency appear in all documents, and these are not keywords. Also the words having low frequency are not keywords. The general frequency of words in all documents is shown by the dotted line in Fig.2. Since this general frequency doesn't include characteristic words

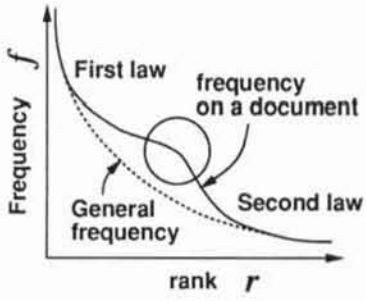

Fig.2 Zipf's law

like keywords, this line is monotony.

EXTRACTION OF KEYWORD INFORMATION

The extraction of keywords in postprocessing is very different from general keyword extraction. This is because the extraction in postprocessing determines keywords from character candidates that are ambiguous and include errors. Our proposed method is shown in Fig.3. First postprocessing without keyword using information is done. After that, the corrected character strings and the BUNSETSU of those strings are evaluated. BUNSETSU evaluation E_j is calculated using the character's reliability evaluation, the word's length, the word's frequency, the word's part of speech included in BUNSETSU j, and the length of BUNSETSU j. Since the corrected character strings still contain some errors, we evaluate the correctness of words using BUNSETU. As BUNSETSU evaluation E_j is higher, BUNSETSU j is more reliable to extract keywords. Using Zipf's law, the difference between the word's frequency in this document and the word's frequency in general is used to select the keywords.

$$K_w = \sum_{\text{(BUNSETSU } j \text{ including word } w)} \max((aE_j - F_w), 0) \qquad (3)$$

Based on these considerations, the keyword evaluation K_w is calculated as follows.
where 'a' is constant and 'F_w' is the general frequency of word 'w'.

The keyword evaluation K_w is determined by the difference between the word's frequency (that is weighted with BUNSETSU evaluation E_j) and it's general frequency F_w.

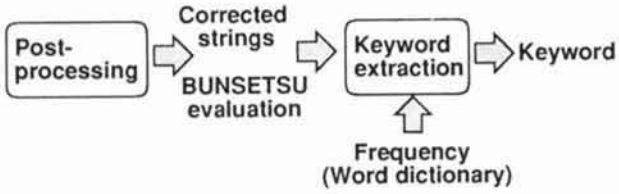

Fig.3 Keyword extraction

POSTPROCESSING USING KEYWORD INFORMATION

Using the automatically extracted keyword information, BUNSETSU evaluation is re-calculated. The keywords reflect the document's content. Therefore, the keyword evaluation should be integrated into the word frequency evaluation. So the evaluation of word frequency E_{jh} is the sum of general frequency F_w and the keyword evaluation K_w. If the word's frequency is larger, the difference of frequency is less important. This relationship is realized by the square root function. The evaluation E_{jh} for BUNSETSU j is shown.

$$E_{jh} = \frac{\sum_{(\text{word } w \text{ included BUNSETSU } j)} \sqrt{bK_w + F_w}}{(\text{the number of word in the BUNSETSU})} \quad (4)$$

where b is constant and it is the parameter that shows how much you should consider the keyword evaluation K_w versus the general frequencey of the word 'w'. Equation (4) shows that as the keyword evaluation K_w and as the general frequency of word F_w becomes larger, the correctness of BUNSETSU is higher. The postprocessing is executed again using this BUNSETSU evaluation rather than the previous evaluation.

Moreover, if the candidate character set does not include the correct character, inferred characters using the keyword information are added to the candidate set of characters. The inferrence rule that we use to add these characters is as follows.

(1)Search the character strings which partially coincide with the keywords.

(2)If no candidate characters coincide with the keywords, then add characters of keywords with the lowest reliability evaluation to the candidate set.

EXPERIMENTAL RESULTS

This section discusses the experimental results of our proposed postprocessing for character recognition. The input data to the postprocessing consists of the characters that OCR(OCR is a program that runs on Panasonic personal computers). OCR outputs 10 candidate characters per input character. The postprocessing for character recognition is executed on SparcStation 2 (28.5MIPS,4.2MFLOPS)workstation. The word dictionary contains about 60,000 words, including the discription of a word string(its part of speech and frequency). Recognized documents include 20 printed data sheets(about 13,000 characters) of scientific and general publications. The recognition accuracy without postprocessing is 77.5% ~ 97.5% (average 89.1%).

EFFECT OF USING KEYWORD INFORMATION

Table 1 shows an example of extracted keywords. Table 1 shows keywords form a document on "neuro control air conditioner". They are listed in order of the keyword evaluation K_w. In this table keywords which have large evaluation value K_w represent the document content. Equation(3) causes the keyword evaluation K_w of "that" and "way" to be 0 since the general frequency F_w of these words is high.

Table 1 Example of keyword

keyword	keyword evaluation K_w
快適 (comfortable)	1893
エアコン (air-conditioner)	1849
認識 (recognition)	1624
ニューラル (neural)	1535
⋮	⋮
もの (which)	0
こと (what)	0

Next, Fig.4 shows several examples of correction using our keyword information. In this example the first postprocessing (1-pass recognition) without keyword information was unable to correct the strings. However, keywords ("character", "neuro", "learning", "recognition" and "language") were extracted from another part of the document. The second postprocessing (2-pass recognition) which used the keyword information corrected the character strings. It could correct "学省(no meaning)" → "学習(learning)", "文学(literature)" →"文字(character)", and "前証(no meaning)言語(language)" → "前記(above)言語" by using Equation (4) and by adding the character strings from keywords when there were no correct characters in the candidate set.

In Fig.4 the space between characters shows the boundary of BUNSETSU. Even if there are plausible but wrong BUNSETSU, the postprocessing using keyword information can still correct such strings.

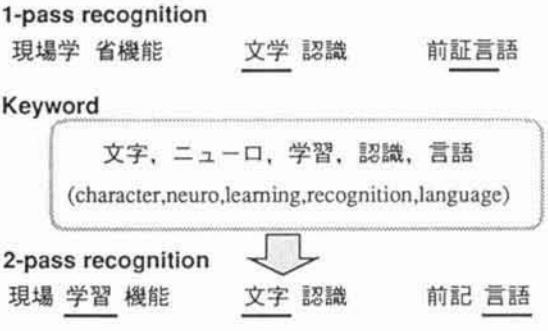

Fig.4 Postprocessing using keyword information

Table 2 shows the performance results of our

Table 2 Recognition rate after postprocessing

document	first candidate recognition rate	1-pass recognition rate	2-pass recognition rate	candidates recognition rate	restoration rate	number of character	execution time (sec)	processing speed (char/s)
1	97.5	99.8	99.8	99.8	100.0	473	6.8	69.6
2	94.8	98.3	98.5	99.9	72.2	848	13.1	64.8
3	88.9	93.8	94.0	95.8	73.9	836	9.7	86.2
4	85.6	90.3	91.2	94.7	61.5	548	4.7	116.6
5	77.7	85.9	86.5	92.6	59.0	631	9.5	66.8

postprocessing. The restoration rate is the evaluation of postprocessing as defined below.

$$\text{restroration rate} = \frac{\text{(2-pass recognition rate - first candidate recognition rate)}}{\text{(candidates recognition rate - first candidate recognition rate)}}$$

where "2-pass recognition rate" is the one after the postprocessing using keyword information. "First candidate recognition rate" is the one before postprocessing. "Candidates recognition rate" is the ratio that the canditates include the correct character. Table 2 and Fig.5 show the relation between the first candidate recognition rate and the 2-pass recognition rate for 20 documents. Fig.5 and Table 2 show that the restoration rate is higer than 70% when the first recognition rate before postprocessing is higher than 90%.

Furthermore, by adding character strings of the keywords to a candidate set, we confirm that the recognition rate increases about 1% for documents with low first candidate recognition rates(<80%), and that the recognition rate does not decrease for the document with high first canditate recognition rates(>90%). This additional method contributes positively to recognition, especially when a document is noisy and unclear.

We conclude the following pointsconserning this metod of adding character strings.

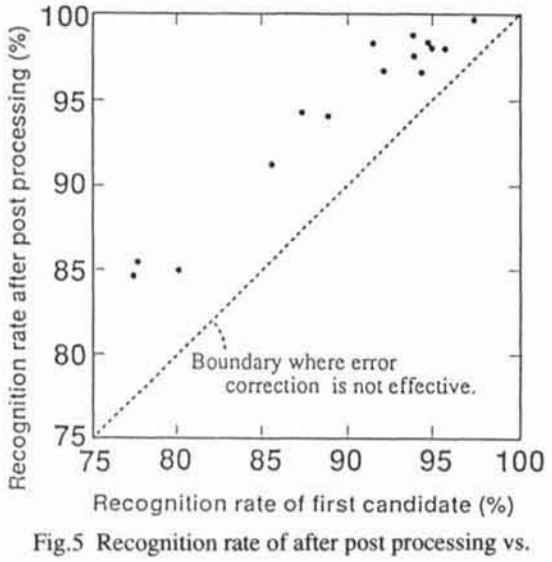

Fig.5 Recognition rate of after post processing vs. recognition rate of first candidate

(1) Even when the first recognition rate before postprocessing is low, the result of the postprocessing is not degraded.

(2) Even when the first recognition rate before postprocessing is high, the addition of character candidates does not harm the performance.

To say briefly, the postprocessing extracts keywords successfully, and the postprocessing using keyword information raises the recognition accuracy.

CONCLUSION

We propose a new postprocessing method using automatically extracted keyword information, and show our experimental results of this method.

Keywords are selected which represent the information about a document's content. We propose an extraction method for keywords from character candidates that include errors. By our experimental results of postprocessing using keyword information, we showed that use of keyword information raises the recognition accuracy. These keywords are also useful for seaching a document quickly. To raise recognition accuracy more, we need to study how to extract a higher level of meaning from a document and a postprocessing method that uses this knowledge.

Acknowledgments

We would like to thank Mr. Yokoe at Computer Division of Matsushita Electric Industrial Co.,Ltd. for his support in the character recognition unit

References

(1) Sugimura T.:"Error Correction Method for Character Recognition Based on Confusion Matrix and Morphological Analysis", Trans. IEICE **J72-D-II**, pp.993-1000(1990).

(2) Niwa H., Kayashima K., and Shimeki Y.:"A Postprocessing for Character Recognition and an Analysis of Error Correction in Postprocessing --- A Use of Keyword Information", IEICE Technical Report, **PRU91**-135(1992).