Recognition of Character Strings Printed with Large Alignment Error

Minenobu Seki, Toshikazu Takahashi, Takeshi Nagasaki, Hiroshi Shinjo, and Katsumi Marukawa
Central Research Laboratory, Hitachi, Ltd.
Kokubunji,Tokyo 185-8601,Japan
{mseki,takahash,naga-t,shinjo,marukawa}@crl.hitachi.co.jp

Abstract

Optical character reader (OCR) technology for reading documents, such as monetary transaction documents, is becoming more and more important than ever before. The position of the printed character string is sometimes largely shifted from its designated position, and there may be two or more directions in spite of one sheet of paper. There are various causes, the performance of the printer, a mistake in the printing position designed by the software, a variation in the cell positions caused by the publishers and by the publishing dates, or even a mistake of the handwriting position. We developed a recognition method for determining which character strings are to be read in such difficult situations. A method for determining the correspondence of character strings to cells with a very high success ratio was developed. This method is based on local and global rules, and effective control of these rules. It was estimated to have a 99.2% success ratio using an experiment on 11,387 character strings.

1. Introduction

In Japan, companies are legally obliged to keep documents related to monetary transactions in the physical paper form. However, because of the increasing cost of preserving such paperwork, the government is now preparing a law that will permit images of the documents and their content to be electronically preserved. Optical character reader technology for reading documents such as monetary transactions is therefore becoming more and more important than ever before. Such transaction documents usually have preprinted ruled lines forming cells (items in a table) and cell titles, and data expressed by the character strings, such as the amounts of money, addresses, and names, are printed in the cells at a later date. The structures of the documents vary are often very complicated, and the cells adjoin each other in a very narrow area. Using conventional methods for reading data, we would first analyze the documents structure and then certain cells would be extracted from the document's image [1][2]. The character strings would then be read in the cells under the assumption that they must be located at designated positions with small alignment errors.

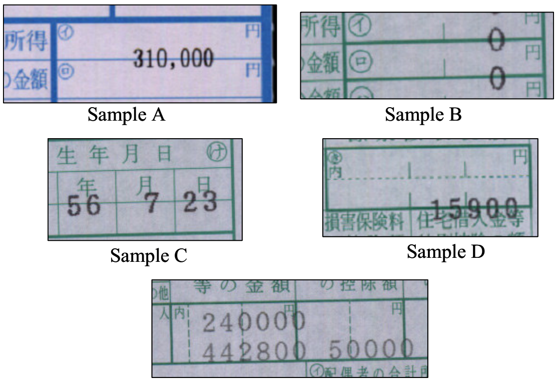
However, in actual applications, the position of the printed character string is sometimes largely shifted from the designated position, and in various directions within the document, and the character strings often cross over or touch the preprinted ruled lines or cell titles. There are various causes for this, such as the performance of the printer, a mistake in the printing position designed by the software, a variation in the cell positions in a document by the publisher or by the publishing date, and/or a mistake in the handwriting position. It may also be printed two or more times in a document. Accordingly, we developed a recognition method to determine which character string belongs to which cell to be read in such difficult situations.

2. Summary of problems

Generally speaking, to cope with the above-mentioned situations, the cell structure is analyzed and the reading area is slightly expanded according to the dimensions of the detected cell, so that the whole object (a character string) fits within the cell [3][4]. However, this expansion cannot solve the following two problems that occur when alignment errors are large and in various directions within a document.

a) When it cannot be determined whether the character string is an object to be read or is an object encroaching from another cell (samples A and B in Fig. 1).

b) When the object overlaps with the pre-printed titles and ruled lines. It is sometimes difficult to separate them in a binary image. However, if color image processing is applied to the whole document, a lot of processing time is required (samples C, D, and E).



Sample A
Sample B
Sample C
Sample D
Sample E

Fig. 1: Examples of difficult cases to read

3. Proposed method

3.1. Approach and outline

The purpose of the proposed method is to determine which character string belongs to which cell (Fig. 2). The approach is based on a decision made according to four local rules using the relation between the character string and a cell and two global rules using the consistency of the local results. Also, binary and color images are effectively processed, and processing time is kept within limits. The proposed method is executed using the following steps (Fig. 3):

1) Pre-processing: Lines are detected, and cells are detected from a binary document image.

2) Selection of character strings with alignment errors: The character strings are extracted and checked to see whether they are touching the ruled lines caused by a large alignment error. Section 3.2 will describe this in more detail.

3) Separation from preprinted characters: Color image processing is executed to separate the strings from the preprinted titles and ruled lines for only the character strings found to be touching in step 2. Section 3.3 will describe this in more detail.

4) Determination of correspondence: This is most important step. The correspondence between the cell and the separated character string is determined by three local and two global rules. Section 3.4 will describe this in more detail.

5) Character string recognition: Character strings are recognized by the corresponded cell.

Step 4 is used to address the problem explained in Section 2 (a). Steps 2) and 3) are used for addressing the problems described in Section 2 (b).

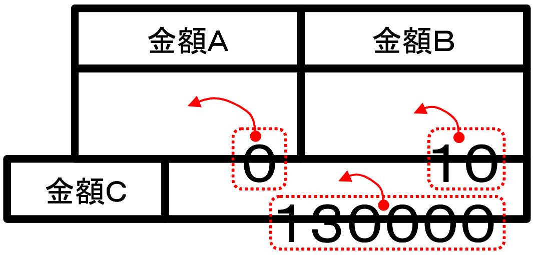


Fig. 2: Correspondence of character strings

3.2. Selection of character strings with large alignment errors

The selection is based on the positional relationship between a character string and the ruled line of a cell in question after the lines and character strings have been extracted from a binary document image. If the character string touches the line or a title, it is regarded as a candidate string to be processed by the following steps in order to determine its correspondence. If the whole character string is within the cell, the string should be directly read by character-string recognition.

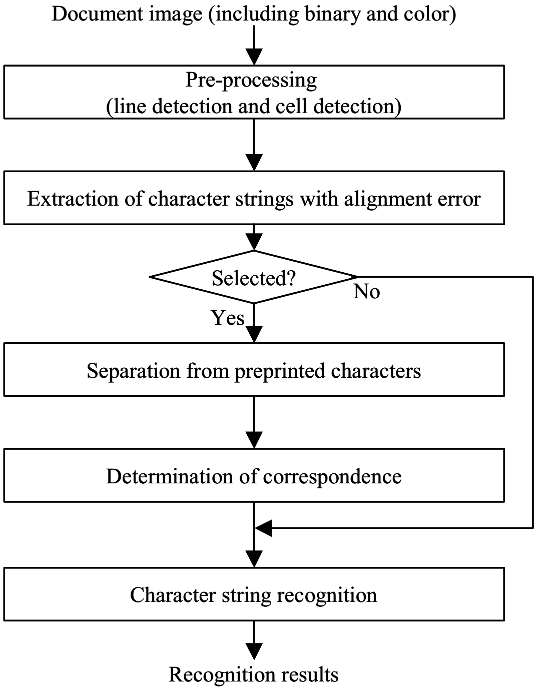

Document image (including binary and color)
Pre-processing
(line detection and cell detection)
Extraction of character strings with alignment error
Selected?
No
Yes
Separation from preprinted characters
Determination of correspondence
Character string recognition
Recognition results

Fig. 3: Flow of proposed method

3.3. Separation from preprinted characters

Color-dropout processing is based on the analysis of the color image distribution in the cell. It separates the character string from the preprinted characters and ruled lines that are normally a different color from that of the character string to be read. The color information separation is more efficient than using binary information. Moreover, the separation was additionally executed using geometric features on the binary image. The computational cost of this color processing, on only the selected strings, is much less than processing the whole document image.

3.4. Determination of correspondence

Correspondence between the cell and the separated character string is at first determined by three local rules and one global rule. Where determination is impossible using the previously stated rules, it can be done by a second global rule (based on interpolation from the results in adjacent cells) and one local rule. The rules are ordered in advance so that there are minimal determination errors. This ordering is discussed in more detail in section 3.5. Each rule is described in the following sections.

3.4.1. Presumption by global rule 1 using non-entry cells

A cell position designed for non-entry of characters has useful information for determining the correspondence to its neighboring entry cells. This is because a character string on the ruled line of a non-entry cell must be a string encroaching from the adjacent entry cell. This determined correspondence creates the simple global rule described in Fig. 4. In this figure, character (1) "0" encroaches into the non-entry cell of " 生年月日". It is easily concluded that character (1) should be in Cell A, according to the position of the ruled line touched by the character. Accordingly, character (2) should be in Cell B as a result of the previous decision made regarding character (1) placement in Cell A. In addition, it is easy to determine that character (3) should be in Cell C, character (4) in Cell D, and character (5) in Cell E. Therefore, information for determination spread from non-entry cells.

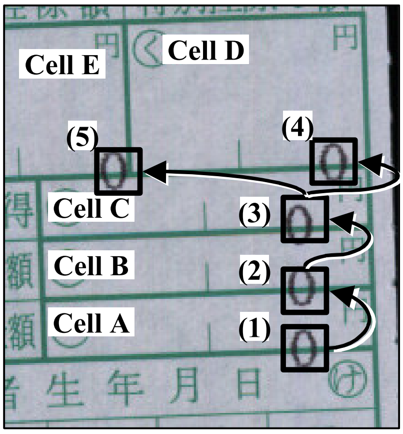


Fig. 4: Presumption using non-entry cells

3.4.2. Determination of correspondence between cells and character strings by three local rules

(1) Local rule 1: Rectangular size of a character string and a cell

There must be correlation between the size of a cell and the size of a character string in that cell. For example, a large string generally requires a large cell. If the size of the processing character string is larger than the size of the processing cell, it can be concluded that the string should be in the nearest adjacent cell and does not correspond to the processing cell. The height of a cell is Hc, and the width is Wc; the height of a string is Wst, and the width is Wst. If $Wc \le Wst$ or $Hc \le Hst$, the string is determined not to correspond to the processing cell.

(2) Local rule 2: Layout of characters lying on two adjacent horizontal cells

When a character touches a vertical ruled line of a cell, it must be determined whether the character belongs to the left or right cell. In this determination process, the positional relationship between the vertical ruled line and characters of two adjacent cells is analyzed according to the five cases shown in Fig. 5. $Rlap$ is the touching character, $Rarit$ is the right end of characters in cell A (left cell), $Rblft$ is the left end of characters of in cell B (right cell), Dma is the distance between the left end of $Rlap$ and the right end of $Rarit$, Dmb is the distance between the right end of $Rlap$ and the left end of $Rblft$, Dcb is the distance between the right end of $Rlap$ and the center of Cell B, Dla is the width of $Rlap$ in Cell A, Dlb is the width of $Rlap$ in Cell B, Wb is the distance between the boundary line and the left end of characters in cell B, and Wpb is the distance between the left end of $Rlap$ and the right end of Wb.

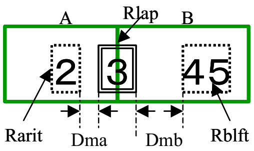
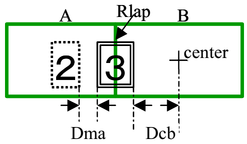
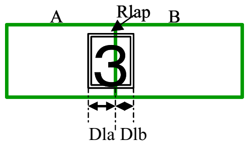
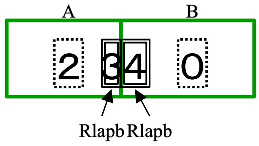
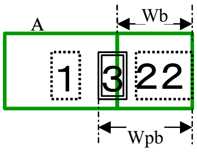
C1:Both $Rarit$ and $Rblft$ exist: if $Dma \le Dmb$,then $Rlap$ belongs in Cell A. if $Dma > Dmb$ $Rlap$ belongs in Cell B.	 A Rlap B 2 3 45 Rarit Dma Dmb Rblft
C2: Either $Rarit$ or $Rblft$ exists: if $Dma \le Dcb$,then $Rlap$ belongs in Cell A. if $Dma > Dcb$,then $Rlap$ belongs in Cell B.	
C3: Neither $Rarit$ nor $Rblft$ exists: if $Dla \le Dlb$,then $Rlap$ belongs in Cell A. if $Dma > Dcb$,then $Rlap$ belongs in Cell B.	
C4: Two $Rlap$'s exist: $Rlapa$ belongs in Cell A. $Rlapb$ belongs in Cell B.	
C5: Special case of C1 and C2: if $Wpb \le Wb$,then $Rlap$ belongs in Cell A.	

Fig. 5: Determination based on character lying on two adjacent horizontal cells

(3) Local rule 3: Occupation degree of a character string lying on two adjacent vertical cells

In general, when the occupation area of the character in the processing cell is large, the possibility that the character belongs to that cell is higher. We set a threshold for the ratio $(R = Din / (Din + Dut))$ of the occupation area in the cell against the total character area. A threshold of 0.9 was set as local rule 3.

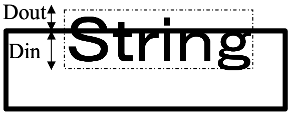


Fig 6: Determination based on character lying on two adjacent vertical cells

3.4.3. Determination by global estimation

The alignment error will be continuous in a document in many cases. If the correspondence of the character string to the cell cannot be determined by the above four rules (global rule 1 and local rules 1 to 3), an interpolation based on the results in several adjacent strings is executed to estimate the correspondence. The number of results of the adjacent strings at distance β (if β is large, it will be in the whole document) from the processing string is counted. Here, *UpNum, LowNum, LftNum,* and *RitNum* mean the correspondence number to the upper, lower, left, and right cells, respectively. The processing string corresponds to the same direction as *GlobalDir,* which is the global direction of correspondence. The rule of decision for *GlobalDir* is formulated as follows:

$$if \quad UpNum \geq LowNum + \alpha \Rightarrow GlobalDir = Upper\_cell$$
$$if \quad LowNum \geq UpNum + \alpha \Rightarrow GlobalDir = Lower\_cell$$
$$if \quad RitNum \geq LftNum + \alpha \Rightarrow GlobalDir = Left\_cell$$
$$if \quad LftNum \geq RitNum + \alpha \Rightarrow Global = Right\_cell$$

The value α is contained because global direction is inappropriate when difference of correspondence numbers or themselves is few. The values of α and β are parameters that can be tuned by heuristics.

3.5 Control of local rules and global rules

Each rule for cell determination described above sometimes gives different results, so the order of determination is important, so the determination should be done in the order of the rule having the higher decision reliability. Table 1 summarizes the reliability of each determination rule. It ranks them in order of high reliability from top to bottom. We can expect a minimal number of correspondence errors by applying the determination rules in this order. In addition, local rule 3' (local rule 3 with a threshold 0.5) is prepared for the remaining undetermined cases.

Table 1: Accuracy of discrimination rules

	Determination rules	Reliability
1	Local rule 1: Rectangular size of a character string and a cell	Very High
2	Local rule 2: Layout of a character string lying on two adjacent horizontal cells	Very High
3	Global rule 1: Presumption using non-entry cells	High
4	Local rule 3: Occupation degree of a character lying on two adjacent vertical cells (Threshold=0.9)	High
5	Global rule 2: global estimation	Low
6	Local rule 3': Occupation degree of a character lying on two adjacent vertical cells (Threshold=0.5)	Low

4. Experimental results

An experiment used 193 test document images related to monetary transactions. The size of the documents was B5. The images were individual binary (400 dpi) and color images (200dpi). The color of the preprinted ruled lines and cell titles was red, blue, or green. The documents included 11,387 character strings (the average number of strings in a document was 59). Among the strings, there were 2,261 strings with alignment errors. The correspondence results from the proposed method are listed in Table 2. The proposed method gave 2,244 correct correspondences for the strings, i.e., a success ratio of 99.2%. Moreover, there are no incorrect correspondences to other, normal character strings. A few non-correspondences existed (0.8%), when the global correspondence direction was not determined and value R of local rule 1 equaled just 0.5, the line detection was mistaken, and the string separation from the preprint was mistaken. Also, the 160 msec. processing time for the document images is practical and within limits.

Table 2: Correspondence results

	Number	%
Success	2,244	99.2%
Failure	17	0.8%
Sum	2,261	100.0%

5. Summary

A method for determining the correspondence of character strings to cells, with a very high success ratio, was developed. This method is based on local and global rules, and the effective control of these rules. Also, binary and color images were effectively processed, and the processing time was kept within limits. The success ratio was estimated at 99.2% using an experiment on 11,387 character strings.

Acknowledgment

The authors would like to thank Dr. Hiroshi Sako for his help and support.

References

[1] H. Shinjo, E. Hadano, K. Marukawa, Y. Shima, and H. Sako, “A Recursive Analysis for Form Cell Recognition,” Proc. of ICDAR'01, pp. 694-698, 2001.

[2] H. Hirano, O. Yasuhiro, and Y. Fumio, “Field Extraction Method from Existing Forms Transmitted by Facsimile,” Proc. of ICDAR'01, pp. 738-742, 2001.

[3] K. Lee, H. Byun, and Y. Lee, “Robust Reconstruction of Damaged Character Images on the Form Documents,” Proc.of GREC’97, pp. 149-162, 1997

[4] D. Nishiwaki and K. Yamada, “An Improvement of Numeral String Recognition Performance on Black Ruled Line Forms using Character Touching Type Verification,” Proc. of GREC’99, 1999.