A Connecting Method for Disappeared Corner Patterns in Form Documents

Hiroshi SHINJO*† Kazuki NAKASHIMA* Masashi KOGA*
Katsumi MARUKAWA* Yoshihiro SHIMA* Eiichi HADANO††

Hitachi, Ltd.

Abstract

Form document structure analysis is an intrinsic technique. However, It has fundamental problems because interruptions of lines and noise cause low-quality form images. This paper focuses on connecting disappeared corner patterns. Our method has two stages for making correct connections, noise elimination, in which lines whose two end points meet no other lines are eliminated and object line selection, where only frame lines of forms are selected. Experiments with form images demonstrated the feasibility of this method.

1 Introduction

In offices, forms - application forms, order forms, etc., are encountered frequently. Form structures and characters on forms should be recognizable automatically. Because positions of characters or items are defined by the structure of a form, automatic form processing necessarily uses form structure analysis. Such analysis is generally based on the arrangement of lines of which forms consist[1,2,3]. Interruption of lines and noiselike lines, however, cause incorrect analysis, so an error-free analysis technique is necessary.

This paper describes a method for connecting corner patterns in which portions of the horizontal and vertical lines are not visible, referred to as "disappeared corner patterns." This method eliminates noise and connects lines in order to recreate disappeared corner patterns correctly.

2 Difficulties of Connecting Disappeared Corner Patterns

Two kinds of conventional line connecting methods cannot connect disappeared corner patterns. The first basic method changes the white pixels between two lines to black pixels in order to expand and connect black pixel area, and change black pixels to white to shrink black pixel area[4]. This method, however, can only connect small gaps, because it cause incorrect connection between lines and other components that are characters or noise if expansion parameter is large. The second conventional method connects interruptions between both end points of a line, because it is based on the straight arrangement of two lines. It can also connect a "disappeared" intersection, because it connects both the horizontal and vertical lines that make up the intersection individually. It cannot connect, however, a disappeared corner because the two lines are arranged at a right angle.

Moreover, there is the possibility of ambiguity of connection for disappeared corners, as shown in Figure 1. The black circles indicate correct connections, and the white circles indicate incorrect connection. Form documents consist of characters, lines, and noise that may be recognized as lines, and thus incorrect connection of disappeared corners or intersections.

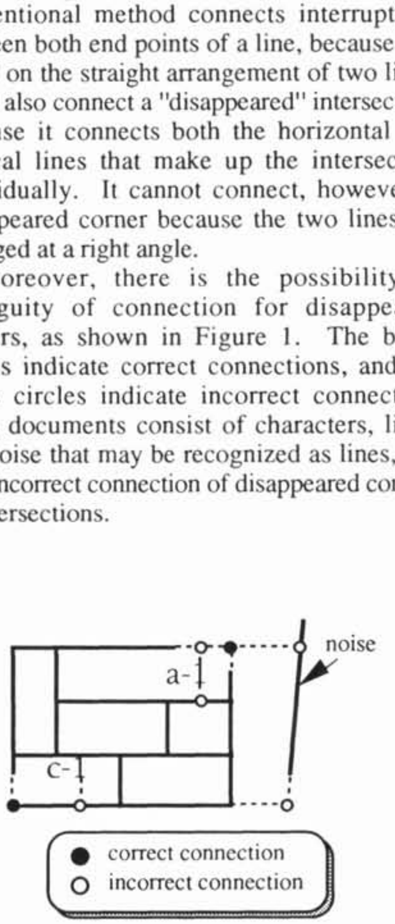


Figure 1. Ambiguity of connection for disappeared corner patterns

[*] Address: 1-280, HIgashi-koigakubo, Kokubunji, Tokyo 185 Japan.

[†] E-mail : shinjo@crl.hitachi.co.jp

[††] Address: 322-2, Nakazato, Odawara, Kanagawa 250 Japan.

3 Approach

3.1 Limitation of Corner Pattern

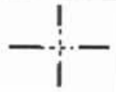
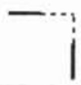
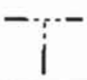
Because of much ambiguities of connection, we limit an available disappeared corner pattern in this paper. Table 1 shows three types of disappeared intersection patterns. Out method is limited to connecting the "L-junction" disappeared corner pattern, shown in Table 1. There are two reasons for this limitation. The first is that the "cross-intersection" described above can be connected easily by conventional methods. The second is to avoid connection errors. Connecting "T-junctions" might cause incorrect connections because image processing cannot correctly discriminate all strokes in characters and noises from lines.

3.2 Key Stages for Connection

There are two key stages to this method for connecting disappeared corner patterns without making incorrect connections. The first eliminates noise that might be incorrectly recognized as lines. This noise elimination is based on the assumption that on a form both end points of a line must meet another line. Therefore, if neither end point of a line meets any other lines, it is ignored as noise.

The second stage selects only the outermost frame lines of forms as the object lines of connection. This is because there are many noiselike lines inside the frame, such as vertical and horizontal strokes in characters, and marks written by hand. Moreover, "L-junctions" generally occur in the outermost frame lines of tables.

Table 1. The suitability of our method for disappeared corner pattern

broken corner patterns	cross- intersection	L-junction	T-junction
			
conventional method	available	not available	not available
Our method	not available	available	not available

4 Algorithm

The algorithm for this corner connecting method (shown in Figure 2) is as follows.

(1) line extraction
Only horizontal and vertical lines that are longer than a threshold are extracted from the object image. Forms rarely have any diagonal lines.

(2) noise elimination
Lines whose two end points do not meet other lines are eliminated as noise.

(3) frame line extraction
Only frame lines are extracted as the objects of corner connection.

(4) disappeared corner pattern estimation
If an end point of a frame line does not meet another line, it is assumed that there is a disappeared corner. If the end points of all frame lines meet other lines, the connecting processing ends.

(5) end point connection
If an end point of a frame line does not meet another line, it is connected to the end

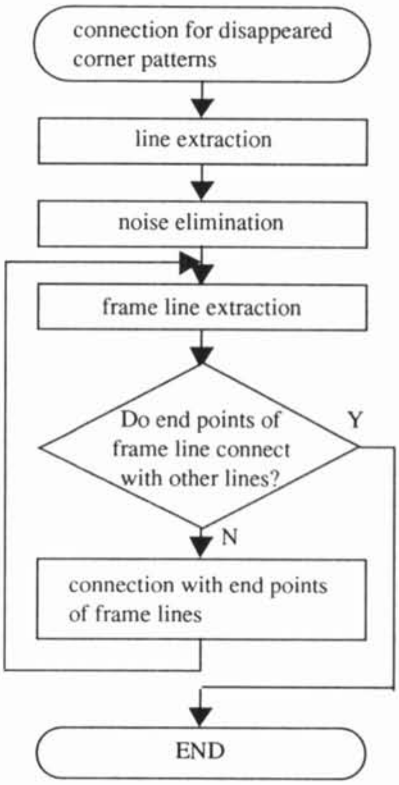

connection for disappeared corner patterns
line extraction
noise elimination
frame line extraction
Do end points of frame line connect with other lines?
Y
N
connection with end points of frame lines
END

Figure 2. Flow of this method

point of another frame line. The outermost frame line has first priority in cases where there are multiple frame lines that cross no other lines at their end point. This is because it is possible that the inner lines are not frame lines but noise.

The flow of this corner connecting method is shown in Figure 3. Figure 3(a) shows an original image that has noise and characters. Figure 3(b) shows the result of line extraction and noise elimination. The vertical stroke in "a-1" is eliminated because neither end points meets a line, but the one in "c-1" is not eliminated because its upper end point meets another line.

Figure 3(c) shows the result of extracting the frame lines of the form. The thick lines are extracted as the outermost lines, and the thin lines are extracted as candidates for frame lines in case the form shape is steplike. Small gray arrows indicate which side a frame line is of a form. The upward arrows, for example, indicate the upper edges of frames.

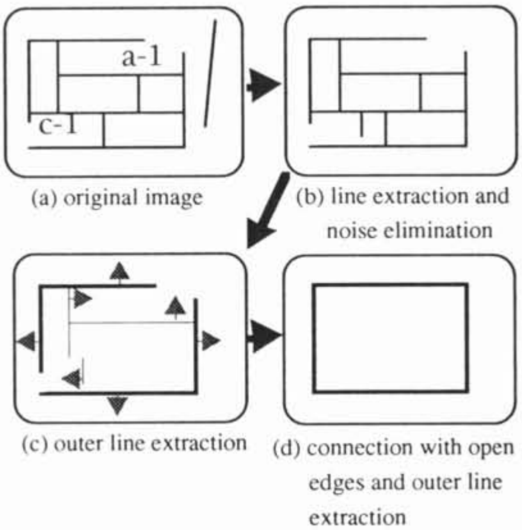
The corner connection process (between Figure 3(c) and (d)) connects two disappeared corner patterns, i.e., the top right and bottom left corners of the form. The explanation for connection is the case of the latter one. The two frame lines on the left side of the form, the long thick one and the short thin one, are extracted. Neither meets another line at its lower end point. However, because the long line is the outer one it has first priority for the connection. Therefore, the bottom left corner consists of the downward extended line of the outermost left line connected to the lowest line. The top right corner connection takes place in the same way.

Figure 3(d) shows the result of the disappeared corner pattern connection and frame line extraction. After the outermost frame lines are connected, the thin lines in Figure 3(c) are determined to be inner lines not frame lines. Therefore, the corner connection process ends because there are no frame lines whose end points do not meet other lines.

5 Experiments and Results

We applied this method to 32 form document images with disappeared corner patterns selected from 1872 samples in our form document database. Of the 32 samples, 25 had disappeared corners and seven had corners with the paper folded up (Figure 4). There is no intrinsic difference between the disappeared corner patterns and the folded corners.

The experimental results showed there were 30 correct and two incorrect connections. With the two incorrect images, the lines could not be extracted correctly because the images had too much noise, as shown in Figure 5. This noise is caused by light in the small space between the paper and the scanner that occurs because the folded corner is thicker than the rest of the paper. However, of the samples in which the lines were extracted correctly, all disappeared corner patterns were connected in



(a) original image

(b) line extraction and noise elimination

(c) outer line extraction

(d) connection with open edges and outer line extraction

Figure 3. Method for disappeared corner connection

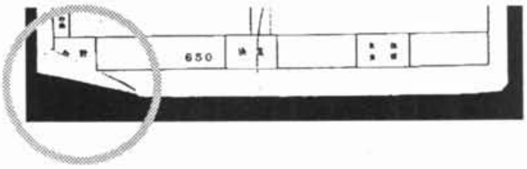


Figure 4. Example of folded up paper

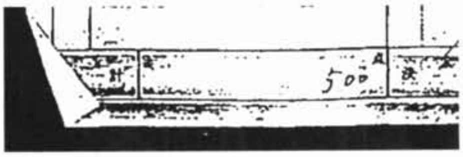


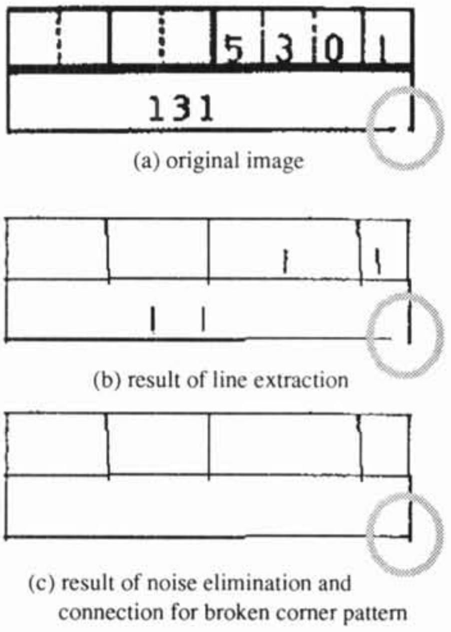
Figure 5. Example of noisy image by folding up on the bottom left corner

the experiments.

Figure 6 shows part of a sample image. The bottom right corner, which is marked by gray circles, is the disappeared corner. The original image is shown in Figure 6(a), the result of line extraction is shown in Figure 6(b), and the result of noise elimination and connection of the disappeared corner pattern is shown in Figure 6(c). The disappeared corner is connected, and the vertical strokes in the characters are eliminated as noise.

6 Conclusion

We proposed a method for connecting disappeared corner patterns, "L-junctions". This method has two key processes: noise elimination, in which lines whose two end points meet no other lines are eliminated, and object line selection, where only frame lines of forms are selected. In experiments with form images, the 30 samples in which the lines were extracted correctly, the corners connected.



(a) original image

(b) result of line extraction

(c) result of noise elimination and connection for broken corner pattern

Figure 6. Example of experimental results

References

[1] T. Watanabe, "Layout Recognition of Multi-Kinds of Table-Form Documents," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 17, No. 4, pp. 432-445, April 1995.

[2] J. Liu, X. Ding, and Y. Wu, "Description and Recognition of Form and Automated Form Data Entry," Proc. of 3rd ICDAR, pp. 579-582, 1995.

[3] M. D. Garris, "Correlated Run length Algorithm (CURL) for Detecting Form Structure within Digitized Documents," Third Annual Symposium on Document Analysis and Information Retrieval, pp. 413 - 424, 1994.

[4] Y. Nakagawa and A. Rosenfeld, "A note on the use of local min and max operations in digital picture processing," IEEE Trans. System, Man, and Cybernetics, Vol. SMC-8, No. 8, pp. 632-635, 1978.