Structure Analysis of Low Resolution Fax Cover Pages

Young-Kyu Lim, Hee-Joong Kang, Chang Ahn, and Seong-Whan Lee

Center for Artificial Vision Research, Korea University
Anam-dong, Seongbuk-ku, Seoul 136-701, Korea
{yklim,hjkang,cahn,swlee}@image.korea.ac.kr

Abstract. The increase in the use of faxed documents calls for the need to handle them automatically and intelligently for efficient storage, retrieval and interpretation. A lot of work has been accomplished for page segmentation in high resolution document images. But conventional methods for page segmentation are not suitable for faxed document processing. The well-known difficulties in faxed document processing are concerned with low resolution images and non-standardized formats. In this paper, we propose an effective structure analysis method for low resolution fax cover pages, based on region segmentation and keyword recognition. The main advantages of the proposed method are its capability of accommodating various types of fax cover pages and its fast processing speed. We divide fax cover pages into three regions – header, sender/recipient information and message – to easily identify the recipient's field. The recipient's name is then extracted through the recognition of keyword. The proposed method was tested on 164 fax cover pages. The experimental results show that the proposed method works well on the various types of fax cover pages.

1 Introduction

The need for automatic faxed document management is rapidly growing as faxed document transmission increase in its volume. It is a matter of course that documents transmitted through fax machines are internally encoded in a standard fax image format. It can be easily transformed into computer readable format. So, a faxed document image can be treated in the same way as many other document images. Many studies on document image processing have been accomplished and reported [1,2]. But the conventional methods for document image processing are not suitable for faxed document processing due to the low resolution and absence of standardized formats within faxed documents images.

A fax cover page is usually accompanies the documents being transmitted. It carries the important information regarding sender, recipient, comments, date, fax and phone numbers, etc. For many fax-related applications such as the intelligent fax server, fax-to-email and fax on demand, the recipient's name needs to be correctly extracted and recognized. Figure 1 shows an application area of this research, called an automatic fax router.

S.-W. Lee and Y. Nakano (Eds.): DAS'98, LNCS 1655, pp. 99–113, 1999.

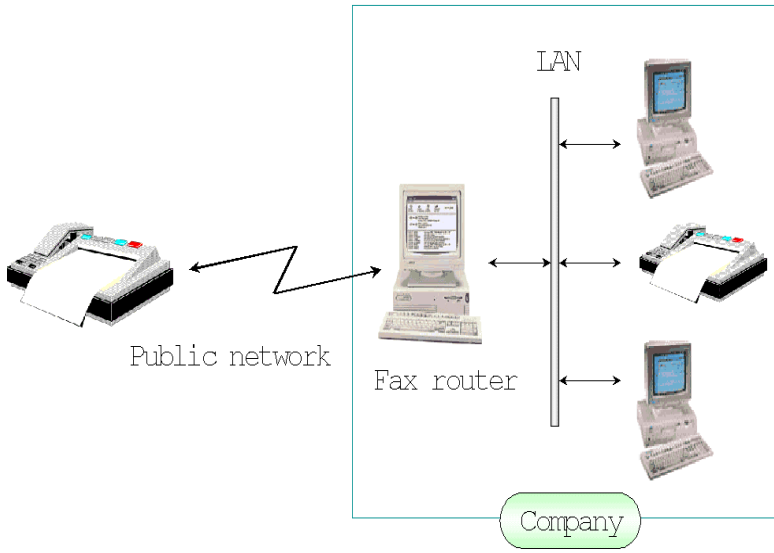


Fig. 1. Automatic fax router

Some related research has been reported in the literature. Li et al. [3] directly recognized names and addresses on fax cover pages using a scheme which strongly couples text recognition and block segmentation. Akiyama [4] performs addressee recognition on documents written in Kanji characters using address indicators like the keyword "TO:" and a double underline. Ricker et al. [5] primarily described order forms in processing faxed documents. The existing methods depend heavily on the performance of the character recognizer and are also restricted by the format characteristic of fax cover sheets such as the existence of field indicators like the double underline. But the difficulty in recognizing low resolution faxed characters demands more robust and efficient techniques.

In this paper, we propose an improved method to extract the recipient's name in low resolution faxed cover pages. The main advantages of the proposed methods are its capability of accommodating various types of fax cover pages and its fast processing speed. At first, the whole image is divided into three regions, which are defined as header, sender/recipient information and message. After the image has been divided, the recipient's field is detected by seeking only through the sender/recipient information region, which contains the target. The keyword is extracted and recognized. Multi-layer backpropagation neural networks are used for character recognition. The recipient's name is identified by comparing the result from character recognition with candidate names in the lexicon.

This paper is organized as follows. Section 2 describes in detail the proposed structure analysis method we adopted for low resolution fax cover pages. Section 3 shows experimental results and analysis for the fax cover pages we collected. Finally, section 4 gives a brief conclusion and directions for future work.

2 Proposed Structure Analysis Method

The images of fax cover pages are inherently of low resolution and the quality is degraded. The approaches that depend heavily on character recognition are not expected to produce good results. In case of Hangul which is Korean, it is necessary to have a more powerful character recognizer. This is not a feasible suggestion in terms of overall system performance. So, we propose an efficient and fast structure analysis method for fax cover pages, based on region segmentation and keyword recognition. This method is able to reduce the dependency on the character recognition result. The whole schematic of the proposed method is shown in Figure 2.

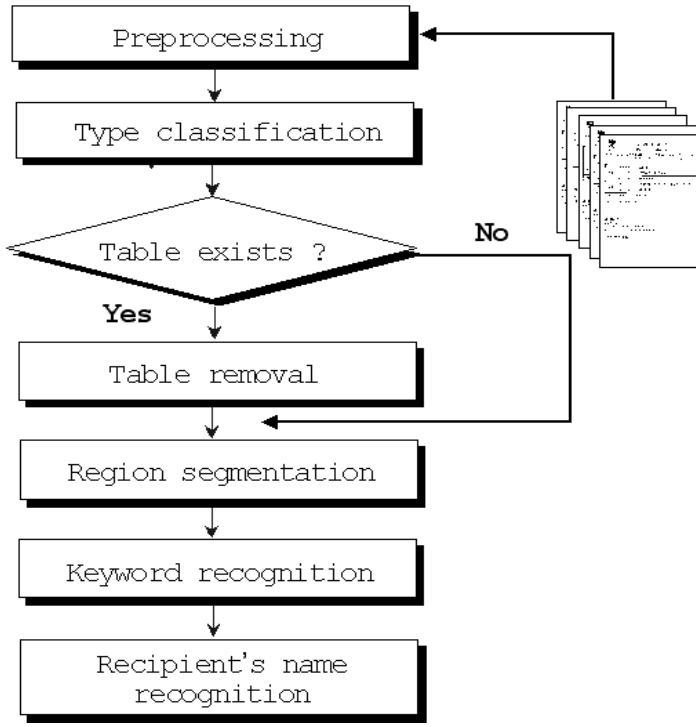


Fig. 2. Schematic of the proposed method

In order to extract information from the fax cover pages, connected components should first be generated. A connected component is a region where black pixels are connected by 8-connectivity [6]. The bounding boxes of connected components are the basic feature used in this system.

2.1 Preprocessing

Preprocessing is done in order to remove noise in images and restore degraded images [7,8,9]. In particular, skew correction is one of the most important tasks in preprocessing. Most existing structure analysis methods do not handle fax cover pages with skew. Applying a Hough transform to one point of the connected components, we designed a computationally efficient skew detection algorithm.

The Hough transform is one of the most popular voting methods for skew detection, line detection, circle fitting problems, etc. [10]. A set of points in x-y axes is mapped into Hough space using the following equation. Figure 3 shows the graphical representation of the Hough transform.

$$\rho = x \sin\theta + y \cos\theta \tag{1}$$

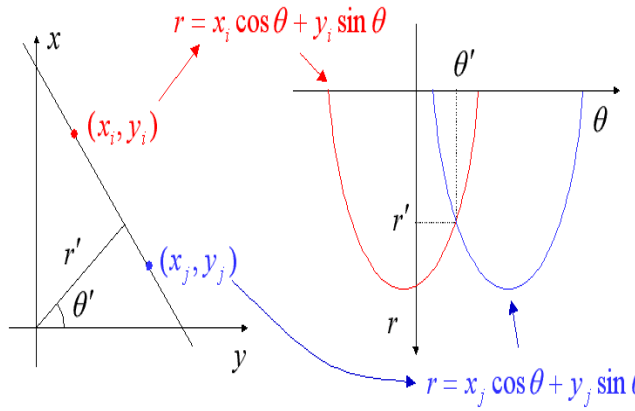


Fig. 3. Graphical representation of Hough transform

We represent the parameter space as an array of accumulators that stand for discrete parameter values. Each of the points in the image votes for several parameters, according to the transform equation. To find the parameters that characterize the line, we should detect peaks in the parameter space.

Since the computation time of Hough transform is proportional to the number of the input pixels, we use only the bottom-right point of the connected components to reduce its cost [11].

2.2 Type Classification of Fax Cover Pages

In this section, we describe the method for classifying the fax cover pages. It is important to develop adequate and fast processing methods for the various types.

We classify fax cover pages into two types. Type I includes unstructured fax cover pages without tables and type II are structured ones with tables. Figure 4 illustrates the types of fax cover pages discussed in this paper. Figure 4(a) shows examples of type I and (b) shows those of type II. The existence of a table is the key feature for type classification on the given fax cover pages. Existing methods do not handle fax cover pages with tables, but the proposed method overcomes this restriction.

There are many objects in fax cover pages – text, table, isolated lines, images, etc. The bounding box of connected components should be investigated for object classification [12]. Two thresholds, width-threshold (T_w) and height-threshold (T_h), are used for object classification. In the process of object classification, we need to check through the connected components in order to find out if there is any table object among them. When we investigated a large of fax cover pages in practice, there were seldom complex forms of table in them. It enabled us to use a simple algorithm for type classification of fax cover pages.

Figure 5 shows the geometric definition of connected components. This definition will be used in the remainder of this paper. The condition for extracting table candidates is given in Eq. (2). By investigating the connected components to see whether or not they satisfy Eq. (2), images or tables will be extracted as candidates.

$$TableCandidates(CC) = \begin{cases} True, & if \quad W > T_w \ \texttt{and} \ H > T_h \\ False, & otherwise. \end{cases} \tag{2}$$

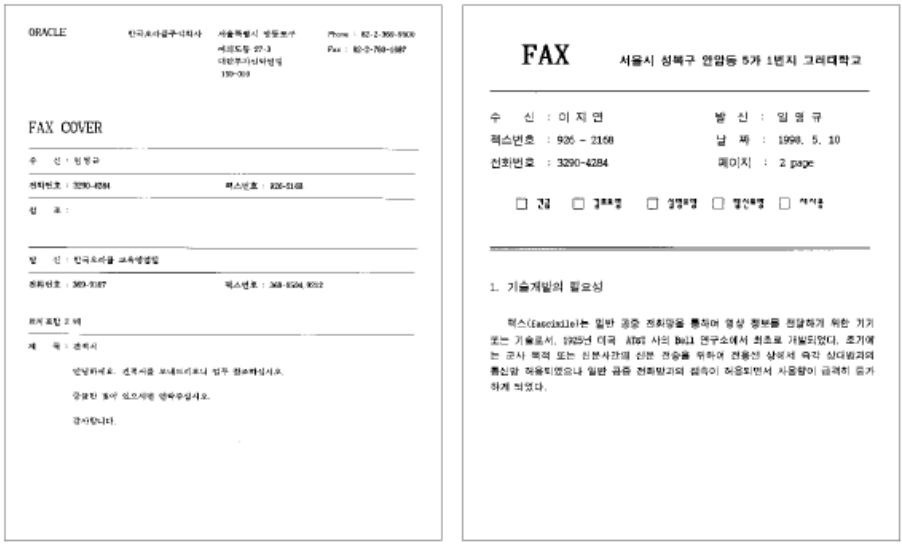
The most identifiable feature of tables in an image is the existence of lines. The ratio of black pixels to a connected component is also used for identification of tables. A table consists of inner and outer lines, as shown in Figure 6. The lines can be detected by histogram value of vertical and horizontal direction[13]. The lines may include noise that distorts their original shape, making it difficult to find them, so run-length smoothing and the number of continuous black pixels with appropriate threshold are used for accurate detection.

The necessary features for the type classification of fax cover pages are as follows.

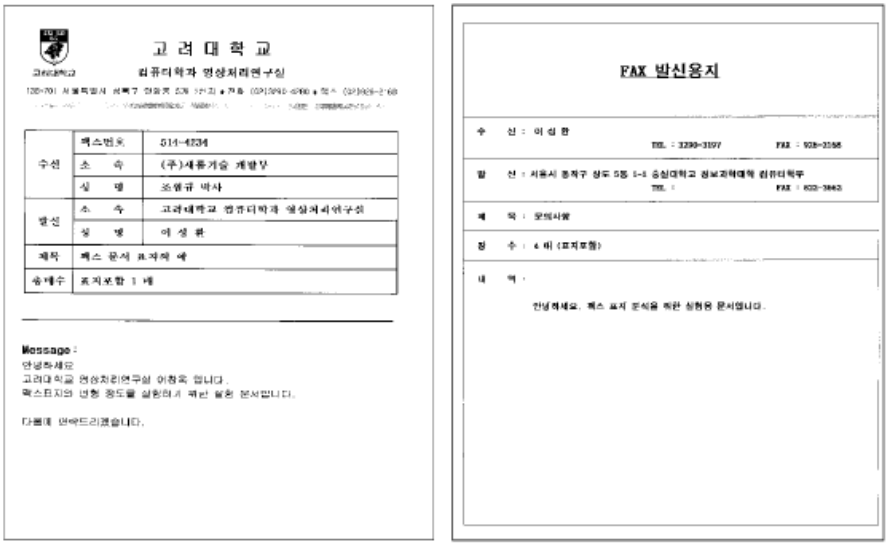
– Histogram

  The histogram feature of a connected component is defined by projecting the black pixels of the skeleton into x- and y-axes and then accumulating the number of black pixels in each point of the x and y axes. This feature is used for detecting lines with similar length.

$$X = \sum_{i=1}^{W} N_i, Y = \sum_{j=1}^{H} N_j \tag{3}$$



(a)



(b)

Fig. 4. Examples of fax cover pages. a) Unstructured fax cover pages. b) Structured fax cover pages.

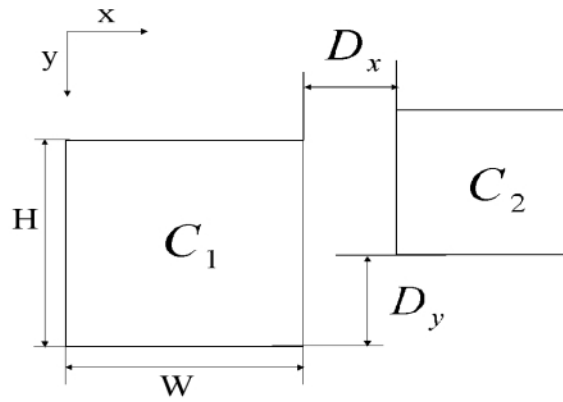


Fig. 5. Geometric definition of the connected components

where W is the height of a connected component and H is the width of a connected component.

– Density

The density of a connected component is defined by the summation of all the black pixels within the connected component.

$$D = \sum_{i,j=1}^{M} C_{ij} \tag{4}$$

where if a pixel point (i, j) is a black pixel, $C_{ij} = 1$; otherwise, $C_{ij} = 0$.

– Ratio

The ratio of a connected component is the number of black pixels to the area.

$$R = \frac{D}{H * W} \tag{5}$$

If the ratio of black pixels to a connected component is larger than threshold, this connected component will be classified as an image.

After a table has been extracted, the inner and outer lines of the table are removed.

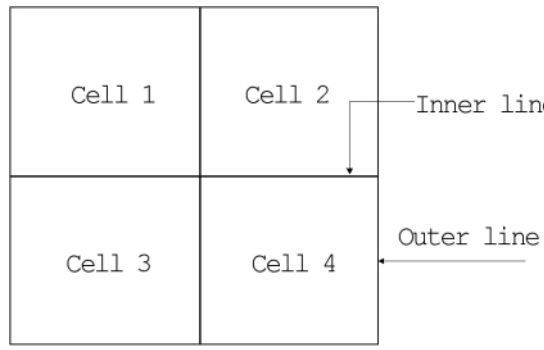


Fig. 6. The basic features of a table

2.3 Region Segmentation

After identifying the type of fax cover pages, region segmentation is proceeded as a next step. Most fax cover pages are divided into three regions. We call the separated regions as header, sender/recipient information and message, as shown in Figure 7.

The header usually contains a company logo or graphics. The sender/recipient information contains the information pertinent to fax transmission, and the message contains comments. The primary goal of region segmentation is to isolate the sender/recipient information region. In most document processing systems, the module with the longest time is the character recognition module. In addition, it has an effect on the performance of keyword extraction and recognition in the next stage, in cases of where similar words exist. Therefore, the procedure of segmentation increases the overall system performance and helps to reduce the amount of data to be processed. Until the end of this paper, the sender/recipient information region will be referred to the SRI region.

In most fax cover pages, a region separator takes the form of a long box or single bold line. Finding the region separator is not an easy task due to noise or distortion during fax transmission. First, run length smoothing is applied for restoration. The connected component(CC) is determined to be a region separator if it satisfies Eq. (6).

$$Separator(CC) = \begin{cases} True, & if \quad H < T_h \ \texttt{and} \ W > T_w \\ False, & otherwise. \end{cases} \tag{6}$$

If we discern a number of components satisfying the above conditions during analysis, the components will later be discriminated. Usually, the first and the last component will be the candidates for the region separators.

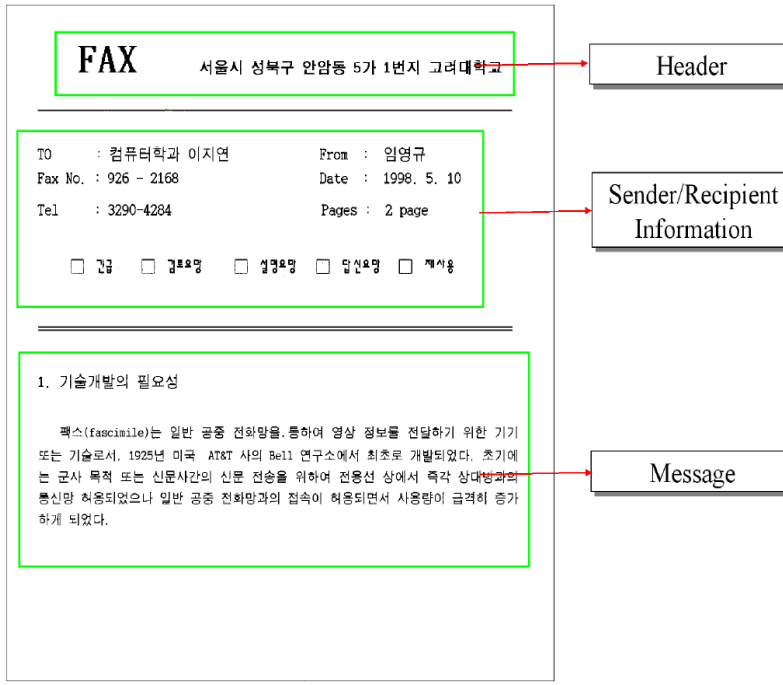


Fig. 7. Region segmentation of a fax cover page

2.4 Keyword Extraction and Recognition

Most of the useful information appears in the SRI region after region segmentation. The SRI region contains various fields related to fax transmission such as sender's name, recipient's name, fax number, etc. To easily detect the type of fields, the connected components within each field are grouped with a block corresponding to its keyword, using Eq. 7 [13].

$$Merge(C1, C2) = \begin{cases} True, & if \quad D_x < T_x \ \texttt{and} \ D_y < T_y \\ False, & otherwise. \end{cases} \tag{7}$$

where D_x is the horizontal distance between two connected components, D_y is the vertical distance between connected components.

There are two formats in unstructured fax cover pages as shown in Fig. 4(a), called one column and semi-two column. The format is determined using the distance between connected components in process of grouping. Where some blocks are located closely together, not only the distance but also the relationship between the blocks has to be investigated. Some blocks may contain two or more

text lines. We use left-alignment characteristic of keyword for the merging of blocks associated with the same keyword. The distance between the blocks associated with different keywords(D_d) is smaller than that of the blocks associated with the same keyword(D_s) as shown in Figure 8.

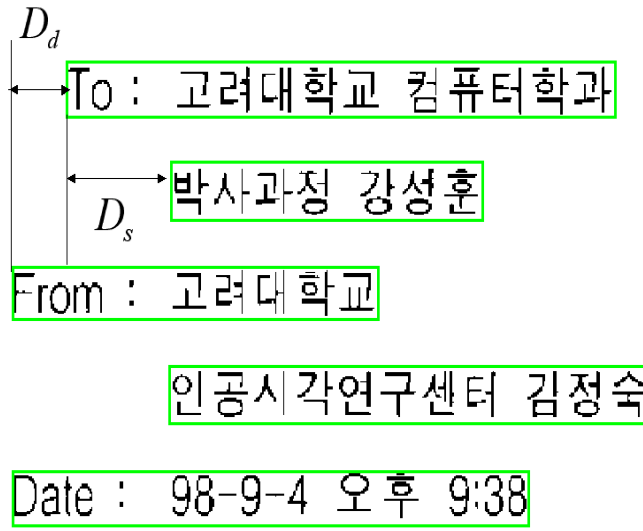


Fig. 8. The comparison of distances between text blocks

As the result of grouping process, each block contains a keyword and contents related to that keyword as shown in Figure 9. A keyword can be easily extracted, provided that the keyword is always located on the left-most side of each block.

Keyword recognition and identification are performed by combination of the small predefined keyword-lexicon with the character recognizer. Korean characters usually consist of three units, called chosung, jungsung and jongsung. Keyword matching entails finding word similarities by comparing the predefined keyword with the recognition results. The similarity is calculated by using the distance between units, which has been defined in advance. The HMM-based approaches have been reported for faxed word recognition [15]. But it does not suitable to korean language. So, the recognition of keywords is done by the Hangul character recognizer, based on multi-layer backpropagation neural network, which shows good performance on low resolution faxed characters [16].

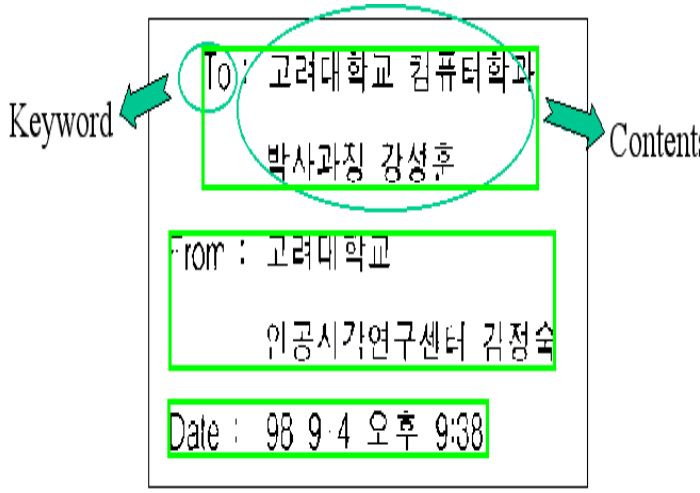


Fig. 9. Example of blocks

The algorithm of the keyword matching is given below.

Let the result of recognition be W and the similarity array be S[i]
For all the predefined keywords in lexicon (W[i]) {

Segment each keyword into three units

Calculate euclidean distance(W,W[i]) between three units

save the distance in S[i]

}

Find minimum in S[i]

The characters within each block are connected and form a linked-list. If a keyword can not be recognized on the left-most side in a block, all the remaining characters in the block are skipped without recognition. This enables fast searching and recognition of keywords. Figure 10 shows the whole process of extracting the recipient's field on fax cover page.

2.5 Identification of the Recipient's Name

The recipient's name on fax cover pages appears in the block including the recipient's keyword. A field indicator like the colon is sometimes followed by the recipient's keyword. Sometimes, a fax cover page without a colon may also

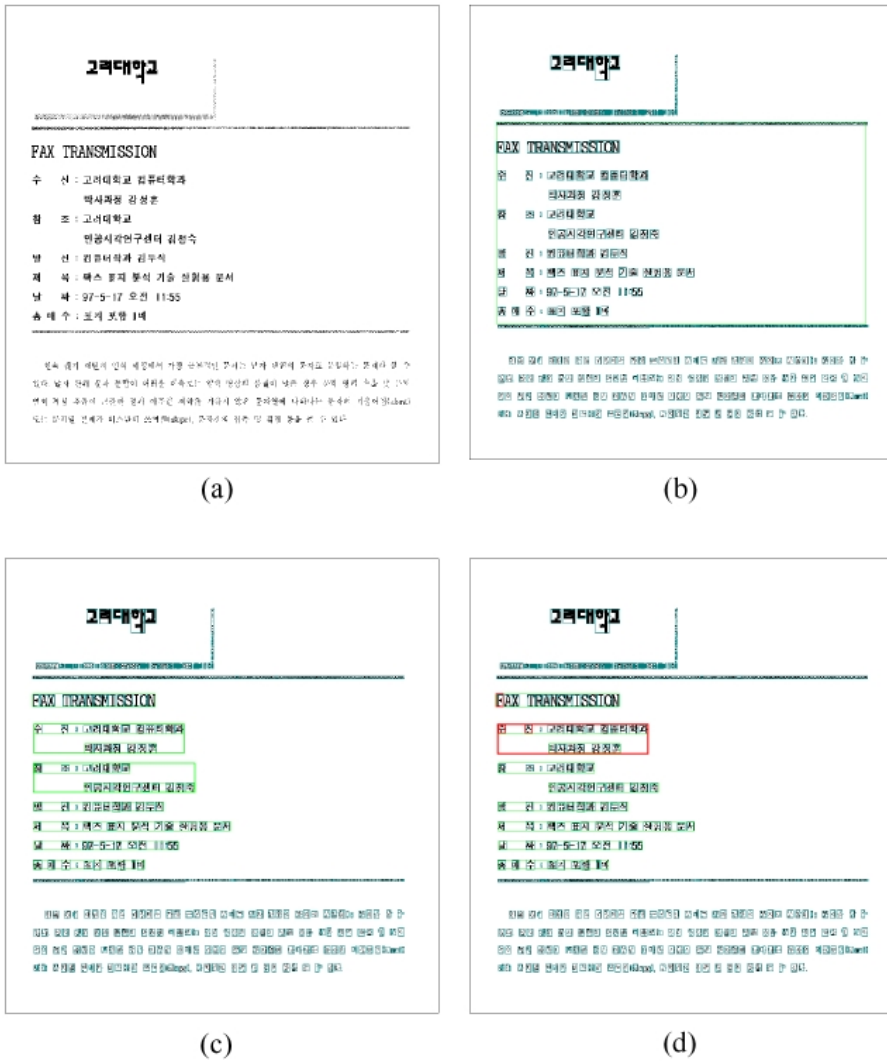


Fig. 10. Process of extracting the recipient's field. a) Input image. b) Region segmentation. c) Grouping. d) Keyword recognition.

exist; therefore it is necessary to check whether a field indicator exists or not. The existence of field indicator is determined using the average size of characters within the block and then the recipient's name is recognized.

There are various pieces of information in the recipient's field like duties, department, etc. This system requires the extraction of only the name in the recipient's field. This is carried out by using word similarity within a small lexicon

containing the names of personnel employed by the company or the institution, based on the properties of Korean names. A Korean name usually consists of 2-4 Hangul characters, so the length of a word is first used for comparison to check whether it is a name candidate or not. An extracted word is verified by using a method similar to the keyword matching algorithm.

As the recipient's name cannot always be recognized correctly by the difficulties of low resolution faxed character recognition, postprocessing is required [17]. It is carried out by confusion matrix.

3 Experimental Results and Analysis

The proposed method was implemented with MS Visual C++ 1.52 on a Pentium 166MHz machine. We collected 164 fax cover sheets which were scanned by facsimile in 2 modes (fine and normal). The test data has resolutions of 200 by 200 dpi in fine mode and 200 by 100 dpi in normal mode, respectively. The examples of test data were shown in Figure 4 in section 2. Table 1 and 2 show the experimental results of the proposed method on two different types of fax cover pages. In Table 2, F_r, F_k and F_n mean the number of failure in region segmentation, failure in keyword recognition and failure in recipient's name recognition respectively.

The experimental results show that the proposed method appeared to work better on fax cover pages scanned in fine mode than on those scanned in normal mode. The result of error analysis allows us to infer that most of errors occurred in the process of keyword recognition or recipient's name recognition. This result is mainly due to the difficulty of low resolution faxed character recognition.

Table 1. Experimental results of the proposed method

Item (mode)	Type I		Type II	
	Fine	Normal	Fine	Normal
# of Total	74	74	8	8
# of Success	69	62	6	5
# of Failure	5	12	2	3

Table 2. Error Analysis

Item (mode)	Type I		Type II	
	Fine	Normal	Fine	Normal
F_r	1	3	0	0
F_k	1	3	1	2
F_n	3	6	1	1

4 Conclusion

In this paper, we proposed an effective structure analysis method for low resolution fax cover pages, based on region segmentation and keyword recognition. The major contribution of the proposed method is the reduction of the dependency on character recognition by using region segmentation and keyword verification based on distance measure. It can enable fast and correct recognition of the recipient's name.

Experimental results confirmed that the proposed method worked well on various types of fax cover pages. We also showed that the fax cover pages were correctly analyzed when the keyword recognition was poor.

In the future, we will extend our research to handwritten fax cover page recognition.

Acknowledgments

This research was supported by Creative Research Initiatives of the Korea Ministry of Science and Technology.

References

1. S.-W. Lee, Character Recognition: Theory and Practice, Hongneung Publisher, Seoul, 1993. (in Korean)
2. J. Li and S. N. Srihari: Location of Name and Address on Fax Cover Pages. Proc. of 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 756-759.
3. Y. Y. Tang, S.-W. Lee and C. Y. Suen: Automatic Document Processing: A Survey. Pattern Recognition, Vol. 29, No. 12, 1996, pp. 1931-1952.
4. T. Akiyama: Addressee Recognition for Automated FAX Mail Distribution. Proc. of SPIE Conference on Document Recognition(III), Vol. 2660, San Jose, California, January 1996, pp. 677-680.
5. G. Ricker and A. Winkler: Recognition of Faxed Documents. Proc. of SPIE Conference on Document Recognition, Vol. 2181, San Jose, California, February 1994, pp. 371-377.
6. J. Ha, R. M. Haralick and I. T. Philips: Document Page Decomposition by the Bounding-Box Projection Technique. Proc. of 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 1119-1122.
7. M. Y. Yoon, S.-W. Lee and J. S. Kim: Faxed Image Restoration Using Kalman Filtering. Proc. 3rd Int. Conf. on Document Analysis and Recognition, Montreal, Canada, August 1995, pp. 677-680.
8. J. C Handley and E. R. Dougherty: Optimal Nonlinear Fax Restoration. Proc. of SPIE Conference on Document Recognition, Vol. 2181, San Jose, California, February 1994, pp. 232-235.
9. J. Liang and R. M. Haralick: Document Image Restoration Using Binary Morphological Filters. Proc. of SPIE Conference on Document Recognition(III), Vol. 2660, San Jose, California, January 1996, pp. 274-285.

10. B. Yu and A. Jain: A Robust and Fast Skew Detection Algorithm for Generic Documents. Pattern Recognition, Vol. 29, 1996, pp. 1599-1629.
11. D. S. Kim and S.-W. Lee: An Efficient Skew Correction and Character Segmentation Method for Constructing Digital Libraries from Mixed Documents. Proc. of The 23rd KISS Fall Conference, Vol. 23, Taegu, Korea, April 1996, pp. 293-206. (in Korean)
12. K. Fan and L. Wang: Classification of Document Block Using Density Features and Connectivity Histogram. Pattern Recognition Letters, Vol. 16, 1995, pp. 955-962.
13. S. W. Lam, L. Javanbakht and S. N. Srihari: Anatomy of a From Reader. Proc. of the 2th Int. Conf. on Document Analysis and Recognition, Tsukuba Science City, Japan, October 1993, pp. 506-509.
14. Y. Katsuyama and S. Naoi: Fast Title Extraction Method for Business Documents. Proc. of SPIE Conference on Document Recognition(IV), Vol 3027, San Jose, California.February 1997, pp. 192-201.
15. A. J. Elms, S. Procter and J. Illingworth: The Advantage of Using HMM-based Approach for Faxed Word Recognition. International Journal on Document Analysis and Recognition, Vol. 1, No.1, 1998, pp. 18-36.
16. D. S. Kim and S.-W. Lee: Performance Comparison of Two Methods for Low Resolution Printed Hangul Recognition. Proc. of The 23rd KISS Fall Conference, Vol. 23, Seoul, Korea, October 1996, pp. 587-590. (in Korean)
17. S.-W. Lee and E. S. Kim: Efficient Postprocessing Algorithms for Error Correction in Handwritten Hangul Address and Human Name Recognition. Pattern Recognition, Vol. 27, No. 12, 1994, pp. 1-10.