8th IPPR Conference on
Computer Vision, Graphics and Image Processing
Tao-Yuan, Taiwan, R.O.C.
Aug. 20~23, 1995

Understanding of Arrangements and Extraction of Articles in Chinese Newspaper Images

Lien-Fen Lee (李蓮芬) and Wen-Hsiang Tsai (蔡文祥)

Institute of Computer and Information Science
Department of Computer and Information Science
National Chiao Tung University, Hsinchu, Taiwan, Republic of China

ABSTRACT

A document image is composed of lots of blocks after being segmented. Among these segmented blocks, there always exist some special relationships coming from the arrangement of the document content. In order to use and retrieve the document content efficiently, understanding of document arrangement is necessary. In this study, approaches to understanding of arrangements and extraction of articles in Chinese newspaper images are proposed. In arrangement understanding, the knowledge of general layout rules, composition techniques, readers' reading habits, and information of segmented blocks are utilized. After understanding is completed, character segmentation and rearrangement are performed. Different processes are designed for segmenting characters in textlines and headlines. Also, four formats are provided for article rearrangement. In the process of rearrangement, methods for detection of paragraph ends and recognition of some special characters and symbols are proposed. Recognized special characters and symbols are rotated to fit the rearrangement format. The proposed system also provides a friendly interactive interfacing for the above woks. Experimental results show the feasibility and practicability of the proposed approaches.

1. Introduction

A document image is composed of lots of blocks after being segmented. Such segmented blocks could include head blocks, text body blocks, graphic blocks, frame blocks, line blocks, etc. To find out the relationships among these segmented blocks and to decide the reading order for the articles in a document image is a major work of document understanding. If we can find out these relationships of blocks in a document and define proper data structures to store them, we can extract articles and rearrange them in some specific format and scaling size for more convenient use.

As a test of our study on understanding of documents and rearrangement of articles, we will focus on Chinese newspapers in this study. Chinese newspapers are chosen because of their complicated layout styles and their importance as sources of information in the Chinese community.

In 1980's, Toyota, et al. [1] discussed the extraction of Japanese newspaper articles using domain specific knowledge and proposed a algorithm for identifying newspaper regions and extracting newspaper articles. Inagaki and Kato [2] built a special-purpose machine for Japanese document understanding. Higashino, et al. [3] proposed a flexible format understanding method, using a form definition language for representing the layout rules of a document. Kubota, et al. [6] proposed an experimental document understanding system using a production system concept. A document understanding method based on the tree representation of document structures was proposed by Tsujimoto and Asada [4, 5]. A symbolic learning method was proposed by Esposito, et al. [7] in order to automatically acquire the knowledge base of an expert system for document understanding. Semeraro, et al. [8] proposed a supervised inductive learning approach to document understanding.

Several applications about document understanding were studied. Futrelle, et al. [9, 10] studied understanding of diagrams in technical documents and used graphics constraint grammars for describing and analyzing diagrams. Watanabe, et al. [11] studied understanding of table-form documents and proposed a knowledge-based approach which makes use of knowledge to interpret structural features of documents. Oddo, et al. [12] proposed an image understanding approach for locating and extracting data from various business documents. Finally, a system which analyzes the layout of a Chinese newspaper article was proposed by Lau and Leung [13].

Assumptions made of the documents processed in this study include the following.

1. We focus on processing Chinese Newspapers in the proposed system.
2. The segmented blocks are assumed to be rectangular in shape and the information of the segmented blocks is known in advance.
3. We process recent Chinese newspapers which are block-oriented.

Figure 1 shows the overall system flow of the proposed document understanding and rearrangement system. The details are described in the following sections.

2.Text Block Merge

2.1 Detection of orientation of newspaper image

Before text block merge, the system has to determine the orientation of a document image at first. In a Chinese newspaper of the vertical form, the contents are arranged from right to left, and in one of the horizontal form, the contents are arranged from left to right. To detect the orientation of a newspaper image, we simply take it to be the orientation of the text blocks found in the segmented blocks.

2.2 Merge of head textlines

A headline block may be separated into several independent smaller headline blocks. The space between two neighboring broken headline blocks is less than the space between a headline block and another.

The direction of merge is the same as any of these headlines. The merge rule is described as follows.

> *Given two headline blocks, if the widths of both in the direction of merge are similar and the distance between them is small enough, then merge these two headline blocks into a headline block.*

2.3 Merge of body textlines

The system performs merge of body textlines when a text block is broken into several textlines.

The space between two neighboring textlines is less than the space between any two other neighboring blocks, the widths of textlines are small, and the textlines belonging to an identical text block usually form a cluster. If the textline is horizontal, the direction of merge must be vertical. And if the textline is vertical, the direction of merge must be horizontal.

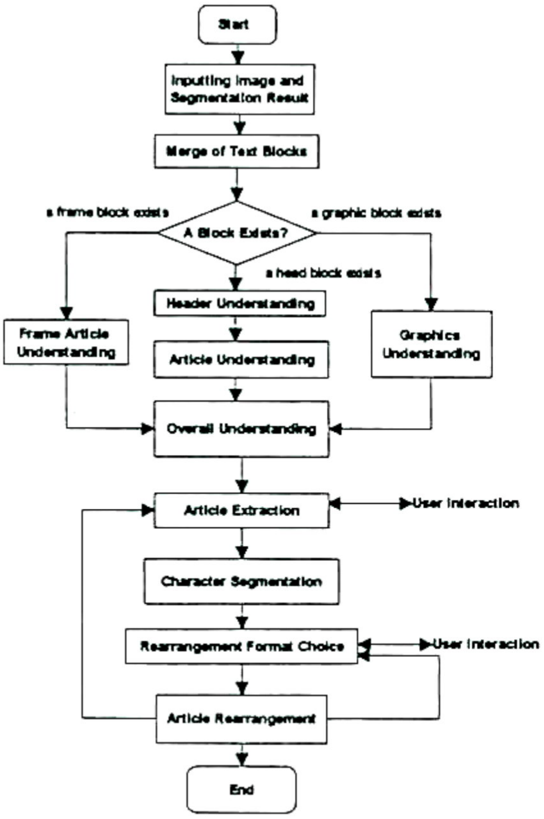


Figure 1. The flowchart of the document understanding and rearrangement system.

3 Article Arrangement Understanding

3.1 Knowledge about block features

In producing a newspaper, many rules are used in composition of the articles. We can obtain useful information by observation of newspaper contents and analysis of human begins' reading habits. Some of the information is as follows.

1. In nowadays newspapers, the main principle for arranging an article is to make it rectangular in shape.
2. Vertical Chinese newspapers are arranged from top to bottom and from right to left.
3. Horizontal Chinese newspapers are arranged from left to right and from top and bottom.
4. An article is always composed of a head block at the beginning, followed by some text blocks.
5. The comprising blocks of an article are arranged as clustered together as possible.
6. A frame block is a separator among articles. A frame block may include one or more related articles.
7. A line block usually is a separator between two articles.
8. The first text block of a vertical head block in a vertical newspaper occurs at the left side of the head block.
9. The first text block of a horizontal head block in a vertical or horizontal newspaper occurs at the bottom side of the head block.
10. Normally, there exists a caption around a graphic block.
11. The characters of the headlines in a head block are larger in size than those of the textlines in a text block.
12. The characters of the textlines in a text block usually have a uniform size, but the characters of the headlines in a head block may have two different sizes.

3.2 Article Arrangement Understanding

An article may consist of more than one head block. Thus, the correlation among head blocks has to be determined at first so that the process of further understanding can be performed more conveniently.

Given a horizontal head block, the possible location of another related to it may be at its upper or lower side. Similarly, given a vertical head block, the possible location of another related to it may be at its left or right side.

Using the results from headline arrangement understanding, the system tries to understand the association of text blocks with each virtual head block. The proposed algorithm is described as follows.

Step 1: Check the orientation of the input head block.

If the orientation is horizontal, then go to Step 2. If the orientation is vertical, then go to Step 4.

Step 2: Find text blocks downward.

Given a vertical head block, the system continues to search text blocks downward until a line block, a graphic block, a frame block, another head block, or the lower boundary of the image is encountered, which are the separators between two articles.

The way of searching downward is to calculate the differences between the left-lower corner of the source block and the left-upper corners of the other blocks, and then to decide the succeeding one whose difference is minimum.

The difference between two corners P and R of the blocks, denoted as D(P,R), is defined as

$$D(P,R) = d(p_x, r_x) + d(p_y, r_y)$$

where P denotes the assigned corner of the source block, and R denotes the assigned corner of another block, and p_x and r_x are the x-coordinate values of P and R, and p_y and r_y are the y-coordinate values of P and R.

Step 3: Check the right boundary for further searching.

If one of the ending conditions described in Step 2 occurs, it means that the lower boundary or another article is encountered and this article should be arranged rightward or be ended. So, before searching further, the system checks if the right boundary of the image is extended or not. If the boundary is not found, the system goes to Step 4 to find a text block rightward; otherwise, the system goes to Step 6 to end the article arrangement understanding of this virtual head block.

Step 4: Find a text block rightward.

Given a horizontal head block, the system searches a text block rightward. If a text block is found, the system goes to the next step. Otherwise, if the system finds a line block, a graphic block, a frame block, or another head block, which are the possible ending targets of this article, it goes to Step 6 to end the article arrangement understanding of this virtual head block.

Step 5: Check the lower boundary for further searching.

After searching rightward is completed, the system checks if the lower boundary of the image is reached or not. If the boundary is not found, then it goes to Step 2 to find text blocks downward below the text block just found. Else, if the boundary is found, the system stops searching downward and goes to Step 3 to check the right boundary again for searching rightward further.

Step 6: End the article arrangement understanding.

If there exists any head block, the system goes back to Step 1 to begin the algorithm again. If no head block exists, the system goes to Step 7 to end the wc:k.

Step 7: End arrangement understanding.

3.3 Frame Article Understanding

Given a frame block in a newspaper image, understanding of the articles in the frame block is mostly the same as the understanding of the articles in a newspaper image. The arrangement in a frame block either may be the same as the arrangement in a newspaper image or will have some specially tasteful composition itself, e. g., the case in which the head block occurs in the middle of an article. In this study, we will only process the special case of frame composition with only one head block. The proposed algorithm is described as follows.

Step 1: Check the number of head blocks.

If the number of the head block is one, go to Step 2. If the number of the head block is larger than one, go to Step 7.

Step 2: Check the location of the only one head block.

If the head block is located at the center of the frame vertically, go to Step 3. If the head block is located at the center of the frame horizontally, go to Step 5. If the location of this head block is not of the above two cases, go to Step 6 to proceed general article arrangement understanding.

Step 3: Find text blocks from left to right side of the vertical head block.

Given a vertical head block, the search of text blocks must be performed from the left to the right side of the head block in order. The search method is the same as that described in the algorithm of article arrangement understanding.

Step 4: Check if the bottom boundary is reached or not.

Repetitively execute Step 3 downward until the lower boundary of the frame block is reached. Once the boundary is reached, go to Step 8 to end the understanding.

Step 5: Find text blocks from upper to lower side of the horizontal head block.

Given a horizontal head block, the search of text blocks must be performed from the upper side to the lower side of the head block in order. The search method is the same as that in the algorithm of article arrangement understanding.

Step 6: Check if the right boundary is arrived or not.

Repetitively execute Step 5 rightward until the right boundary of the frame block is reached. Once the boundary is arrived, go to Step 8 to end the understanding.

Step 7: Perform the article understanding algorithm.

This step is the same as the article arrangement understanding algorithm described in Section 3.2.

Step 8: End frame article arrangement understanding.

3.4 Graphics Understanding

Two algorithms for the associations of captions with graphics and associations of graphics with articles will be proposed in detail. Figure 2 shows an example of the association of captions with graphics.

A. Associations of captions with graphics

At first, the system extracts the text blocks which are captions of graphics. The algorithm is described as follows.

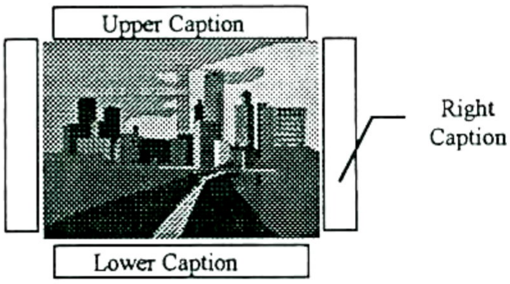


Figure 2. An example of association of captions with graphics.

Step 1: Check if there exist text blocks within graphic block or not.

If there exist some text blocks within a graphic, they may be the characters in the graphic or noise. These blocks usually are meaningless in the understanding process, so the system will ignore them.

Step 2: Search text blocks around the graphic block.

If a graphic has captions, the captions will appear around the graphic. Usually a caption has a small size, and it may be a horizontal text block if it occurs at the upper or lower side of the graphic or may be a vertical text block if it occurs at the left or right side of the graphic.

B. Associations of graphics with articles

According to the block-oriented composition principle, the graphics and its corresponding article usually is arranged in the range of a virtually integrated rectangles including this article. Under this principle, associations of graphics with articles can be found. When the result of this judgment is erroneous, the system will provide an interface for users to correct it interactively. The algorithm is described as follows.

Step 1: Find out the possible virtual integrated rectangles for all articles.

Step 2: Check if the graphic block is located in one of the virtually integrated rectangles.

Step 3: If the judgment in Step 2 makes a mistake, the user can correct it interactively. Users can correct or build the association by actualities. When the association is decided, the graphic understanding work is done.

4. Segmentation of Characters in Textlines

4.1 Overview of Proposed Character Segmentation Approach

Vertical projection [14] is widely used in the segmentation of English characters for gray images. In a Chinese newspaper, the orientations of textlines can be vertical or horizontal, so vertical projections and horizontal projections are used concurrently.

The system deals with color document images, and a color image consists of many colors, not only white and black. So the projection function has to count the pixels with background color, not black pixels. In our proposed system, the inverted projection function is used, which counts the number of pixels whose color indices are the same as the color index of the background in the direction of projection. So the projection function will have a sufficiently large value in the space between characters.

If the text is printed in a perfect condition, the space between the textlines is apparent, that is, the textlines are well separated. Then the segmentation of textlines can be accomplished directly from the projection function. But in the Chinese newspaper, a Chinese character may be separated into some components with space in between. Thus, the result of character segmentation may not be precise. Some corrective approaches, namely, merging and splitting, are employed so that most text can be segmented correctly.

4.3 Headline Character Segmentation

A head block may consist of more than one textline. If the head block of an article is a vertical one, then vertical projection is performed at first to obtain textlines. The space between the textlines is apparent, so the textlines in a head block can be found exactly. Next, the horizontal projection is performed to obtain the characters for each textline. We keep the number of textlines to decide the count of performing the horizontal projection for segmentation of characters.

When the segmentation of characters is performed, some special conditions will be encountered. A Chinese character may be touched by another or it may consist of separated components. The results of segmentation may be erroneous. Thus, a merging and splitting procedure is invoked to segment the textlines. From the first segmentation of characters, the most possible character height is obtained. The merging process uses this possible height to combine small neighboring components into a character. The splitting process splits any segmented region that is higher than this possible height computed from the first horizontal projection of textlines. The new result are more accurate after these two processes.

Figure 3 shows an example of segmentation of characters in a vertical head block. The merging procedure is performed in this example to obtain all isolated characters. In this figure, (a) is the histogram of horizontal projection function; (b) are the margins of the characters obtained from (a) in which the value of the histogram is sufficiently large; (c) are the margins of characters after the merging and splitting process are performed; (d) is the histogram of vertical projection function; (e) is the margins of the textlines obtained from (d) in which the value of the histogram is sufficiently large.

The segmentation of a horizontal head block is similar to the above process of segmentation of a vertical head block.

Finally, we discuss how we segment a headline with different character sizes. If the head textline includes characters with two different sizes, then the statistics of possible character heights in the vertical case or that of possible character widths in the horizontal case will yield two values. The system keeps them and uses both of them concurrently to segment the head textline again.

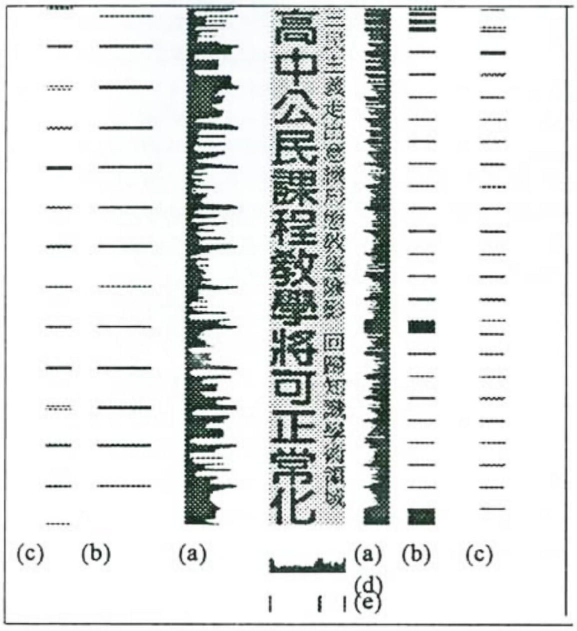


Figure 3. An example of segmentation of characters in a vertical head block. The merging procedure is performed here to obtain all isolated characters.

When the character has a small size, the region of the segmentation should be selected to be the small one of the two possible character sizes. For this, the system confirms the size of characters at first, and then chooses the possible size to segment the character again. The confirmation of the existence of characters with the small size is summarized as follows.

Step 1: If the size of the segmented character is small in the first projection, this region may include some characters with small size.

Step 2: In the vertical head case, if vertical projection is performed on the middle of this small segmented region and a sufficiently large value of the projection function is obtained, then a space really exists in the middle of the region and two characters, separated by this space, with the small size together occupy this region. Thus, the system will validate the existence of the small character size in the region.

Step 3: In the horizontal head case, horizontal projection is performed on the middle of the small segmented region to validate the existence of small character size in the region.

If the case of characters with a small size occurs, the second segmentation will be carried on using the small one. If the case of characters with large size occurs, the second segmentation will be carried on using the large one.

4.4 Text Body Character Segmentation

If all the textlines in a vertical text block are properly arranged, characters located at the same row but in different textlines will have the same character margin. Based on this characteristic, multiline projection can be used. Multiline projections will reflect the overall information and is less affected by noise or separated strokes in Chinese characters. In the vertical text case, the first step is to segment the text block to obtain textlines, and compute multiline horizontal projections for this group of vertical textlines to obtain the segmented characters. In the horizontal text case, multiline vertical projections are computed for this group of horizontal textlines, too.

If the character in a textline is printed irregularly, multiline projections cannot be used. The segmentation of characters for each textline should be performed respectively. In the proposed approach for segmentation for a text block, the technique of multiline projection and the adjustment for each textline are both adopted.

If the text block is vertical, vertical projection is performed at first to obtain textlines. Secondly, horizontal projection is performed to obtain characters for the group of these textlines together.

Figure 4 shows an example of segmentation of characters for a vertical text block arranged regularly in column and in row, in which (a) is the histogram of vertical projection function; (b) are the margins of the textlines obtained from (a) in which the values are sufficiently large; (c) is the histogram of the horizontal projection function; (d) are the margins of the characters obtained from (c) in which the values are sufficiently large. But in this example, not all the characters of the text block can be obtained directly by a vertical projection and a horizontal multiline projection. The splitting procedure is performed additionally when the range between two margins is larger than the average, and the lost margins whose projection values are under the threshold will be obtained.

But, once some textlines have more characters than other textlines owing to the movement of a punctuation mark from the head of the next textline to the tail of this textline, the global result is incorrect for these textlines. An adjustment procedure is proposed for this case. The procedure performs the projection technique to check if every margin of each character obtained from the global result lies within a space. If the corresponding space exists, the global result is correct for this textline. If the corresponding space does not exist, search of the position of the real space is performed around the character locally.

Performing this adjustment procedure, the global result is modified partially and becomes more adaptive to the actual state. When the result of character segmentation by multiline projection is not correct, the user can perform the adjustment procedure to segment the characters again.

For a horizantal block, the segmentation process is similar.

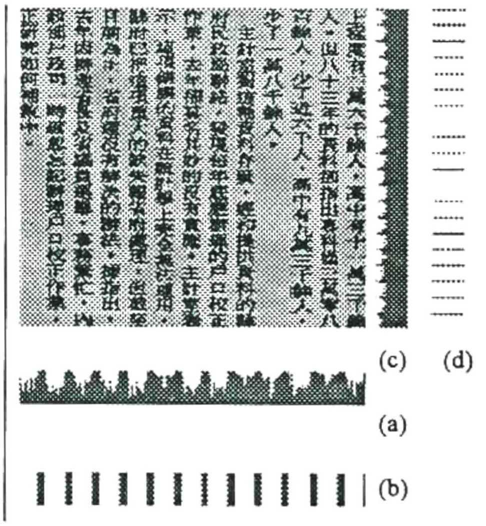


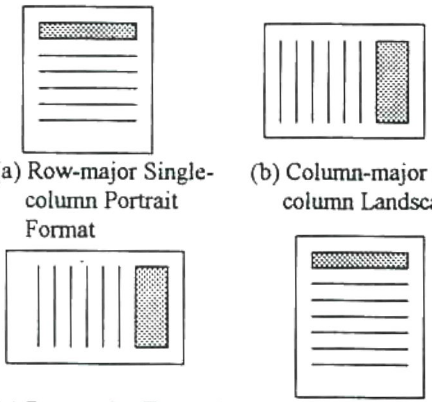
Figure 4. An example of segmentation of characters for a vertical text block arranged regularly in column and in row.

5. Extraction and Rearrangement of Articles

5.1 Introduction

After character segmentation is completed, some applications utilizing the result of the segmentation of characters can be developed. For example, the segmented characters can be rearranged into some given format.

In our system, four formats for article rearrangement are provided for users to select. Figure 5 shows the illustrations about the four formats. And the system provides three choices of scaling sizes for users to select. The three scaling sizes are 1.0, 0.8, and 0.6.



(a) Row-major Single-column Portrait Format

(b) Column-major Single-column Landscape Format

(c) Row-major Two-column Landscape Format

(d) Column-major Two-column Portrait Format

Figure 5. The four formats used for article rearrangement.

5.2 Rearrangement of Articles

A. Text rearrangement

In the row-major single-column portrait form, if R0 is the first row of the textlines and Rq is the final row, and if C0 is the first grid in a row of the textlines and Cp is the final grid in a row of the textlines, the order of grids for characters in this format is arranged as follows:

$$\begin{array}{c} (C0,R0)(C1,R0)\ldots(Cp,R0), \\ (C0,R1)(C1,R1)\ldots(Cp,R1), \\ \ldots\ldots \\ (C0,Rq)\,(C1,Rq)\ \ldots\ (Cp,Rq). \end{array} \qquad (1)$$

In the vertical head blocks of an article, if C0 is the first column of the headlines and Cn is the final column, and if R0 is the first row in a headline and Rm is the final row, the order of characters in the original vertical head block is arranged as follows:

$$\begin{array}{c} (C0,R0)\,(C0,R1)\ \ldots\ (C0,Rm)\,, \\ (C1,R0)\,(C1,R1)\ \ldots\ (C1,Rm)\,, \\ \ldots\ldots \\ (Cn,R0)\,(Cn,R1)\ \ldots\ (Cn,Rm). \end{array} \qquad (2)$$

And in the horizontal head blocks of an article, if R0 is the first row of the headlines and Rm is the final row, and if C0 is the first column in a headline and Cn is the final column, the order of characters in the original horizontal head block is arranged as follows:

$$\begin{array}{c} (C0,R0)\,(C1,R0)\ \ldots\ (Cn,R0)\,, \\ (C0,R1)\,(C1,R1)\ \ldots\ (Cn,R1)\,, \\ \ldots\ldots \\ (C0,Rm)\,(C1,Rm)\ \ldots\ (Cn,Rm). \end{array} \qquad (3)$$

And in the vertical text blocks of an article, if C0 is the first column of the textlines and Cn is the final column, and if R0 is the first row in a textline and Rm is the final row, the order of characters in the original vertical text block is arranged as follows:

$$\begin{array}{c} (C0,R0)\,(C0,R1)\ \ldots\ (C0,Rm)\,, \\ (C1,R0)\,(C1,R1)\ \ldots\ (C1,Rm)\,, \\ \ldots\ldots \\ (Cn,R0)\,(Cn,R1)\ \ldots\ (Cn,Rm). \end{array} \qquad (4)$$

By combining the character order in the extracted article and the destination grid order in the output document, the rearrangement process for the first format can be finished well.

For the other three formats, the rearrangement processes are similar, and are omitted.

B. Decision of paragraph ends

In the process of rearrangement of text blocks, once the end of a paragraph in the original text blocks occurs, the corresponding paragraph in the output document must be ended, and another paragraph must be started from the next textline. To reserve the information and relationship among paragraphs, the decision of paragraph ends is needed. Some rules about this decision is stated as follows.

(1) If the density of the character is low, a punctuation mark or alike characters may occur in this textline. If a blank character occurs next, it indicates that the end of this paragraph is reached and the next paragraph should be started from the following textline.

(2) If the beginning of the next textline is a blank character, we are assured that the above decision is true.

The density of a character is defined as the ratio of the area of non-background pixels over the total area of the character. If rule (1) is true, rule (2) will be performed further to verify the decision. Note that, if the relationships among the paragraphs can be kept well, the quality of the rearrangement result will be better.

C. Rearrangement of graphics

In the proposed system, graphics are arranged in the tail of the text body. To fit the space in the output document, the size of each graphic block is reduced adequately to satisfy the constraint of each format, and filled into the output document at the right location.

As to captions, they will be rearranged according to the interrelationship between the graphics and the captions. But the order of the characters will be reversed if the direction of arrangement changes. In this case, the character segmentation process should be performed to obtained the individual characters at first. Then the segmented result is filled into the right location of the output document.

5.3 Handling of Special Characters and Symbols

When the characters printed in any column-major format are rearranged in any row-major format, or when the characters printed in row-major are rearranged in any column-major format, some special characters or symbols will not appear in their right orientations if the rearrangement process fills them into the output document directly. Thus some further processes should be performed. Some examples of special characters and symbols, which are handled in this study, are shown in Figure 6.

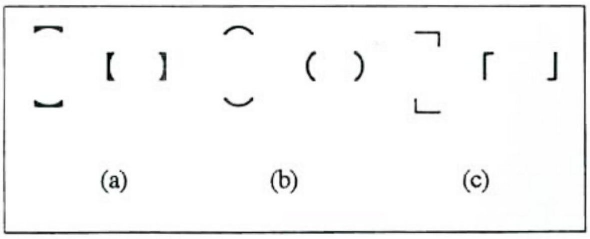


Figure 6. Some examples of special characters and symbols. In (a), (b), and (c), the left side is the right type printed in the column-major format, and the right side is the right type printed in the row-major format.

A. Recognition of special characters and symbols

In the recognition process, some features about the special characters and symbols should be extracted first. The features proposed for use in this study are introduced as follows. An illustration about these features of the special characters is shown in Figure 7.

(1) Projection function:

Using the vertical projection function, the real width of the character printed in the column-major format can be obtained, and the width of the special characters and symbols should be large enough.

Using the horizontal projection function, the number of the space rows and the real thickness of the character will be obtained. In the case of the special characters and symbols, the number of the space rows should be larger than half the height of a segmented character, and the thickness of them will not be smaller than a threshold value.

(2) Crossing count:

The crossing count of each of the special characters and symbols is always one.

According to our experimental experience, the proposed features can be used to recognize the special characters and symbols quite well.

B. Reorienting the shapes of special characters and symbols

After the recognition of a special character or symbol, the process of reorienting the character shape is needed. In this study, a rotation of the original character is performed. If the character is printed in a row-major text block, a clockwise rotation is applied.

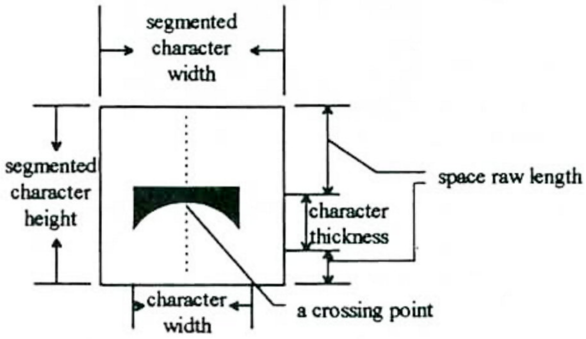


Figure 7. An example of the structure of a special character and symbol.

6. Experimental Results

Several Chinese newspaper images with various arrangement styles were tested in our experiment. The input images were obtained by an HP ScanJet IIC color scanner at 250 dpi, and the results of segmentation were obtained in another study in advance. The proposed system was implemented on a Sun SPARC 10 workstation based on the C language. The result and speed of arrangement understanding is satisfactory. The outputs of rearrangement of articles are good as expected.

The rearrangement results of some sample are shown in Figure 8.

7. Conclusions

A system of semi-automatic arrangement understanding and real-time extraction and rearrangement of articles as been successfully implemented. Several achievements in different phases are summarized as follows.

In the text block merge phase, some merge rules were used to merge broken smaller headline blocks into complete headline blocks or to merge separated textlines into complete text blocks.

In the article arrangement understanding phase, some arrangement rules and reading habits of human begins were used to extract the relationships among segmented blocks and multi-articled arrangement, including the arrangement of articles with vertical titles or horizontal titles, frame articles, and graphics with captions.

In the character segmentation phase, some knowledge about the size of characters and vertical and horizontal projection technique were used to segment text blocks into individual characters.

In the article rearrangement phase, four formats and three scaling sizes were used to rearrange the extracted article in a new format. In the rearrangement process, some additional issues were solved, including the decision of paragraph ends and the handling of some special characters and symbols.

To sum up, the proposed system is useful for understanding, extracting, and achieving Chinese newspaper articles in image forms.

References

[1] J. Toyoda, Y. Noguchi and Y. Nishimura, "Study of Extracting Japanese Newspaper Article," *Proc. 6th ICPR*, Munich, pp. 1113-1115, 1982.

[2] K. Inagaki and T. Kato, "MACSYM: A Herarchical Parallel Image processing System for Event-Driven Pattern Understanding of Documents," *Pattern Recognition*, Vol. 17, No. 1, pp. 85-108, 1984.

[3] J. Higashino, H. Fujisawa, Y. Nakano and M. Ejiri, "A Knowledge-based Segmentation method for Document Understanding," *Proc. 8th ICPR*, Paris, pp. 745-748, 1986.

[4] S. Tsujimoto and H. Asada, "Understaning Multi-articled Documents," *Proc. 10th Int. Conf. Pattern recognition (Atlantic City, NJ)*, pp. 551-556, 1990.

[5] S. Tsujimoto and H. Asada, "Major Components of a Complete Text Reading System," *Proceeding of the IEEE*, Vol. 80, No. 7, pp. 1133-1149, 1992.

[6] K. Kubota, O. Iwaki and H. Arakawa, "Document Understanding System," *Proc. 7th Int. Conf. Pattern Recognition*, pp. 612-614, 1984.

[7] F. Esposito, D. Malerba, and G. Semeraro, "Contextual Supervised Learning for Document Understanding," *Proc. 7th Int. Conf. on Image Analysis and Processing*, pp. 291-297, 1994.

[8] G. Semeraro, F. Esposito, and D. Malerba, "Learning Contextual Rules for Document Understanding," *Proc. 10th Conf. on Artificial Intelligence for Applications*, pp.108-115, 1994.

[9] R. P. Futrelle, I. A. Kakadiaris, "Diagram Understanding Using Graphics Constraint Grammars," *Engineering Systems with Intelligence. Concepts, Tools, and Applications*, pp. 73-81, 1991.

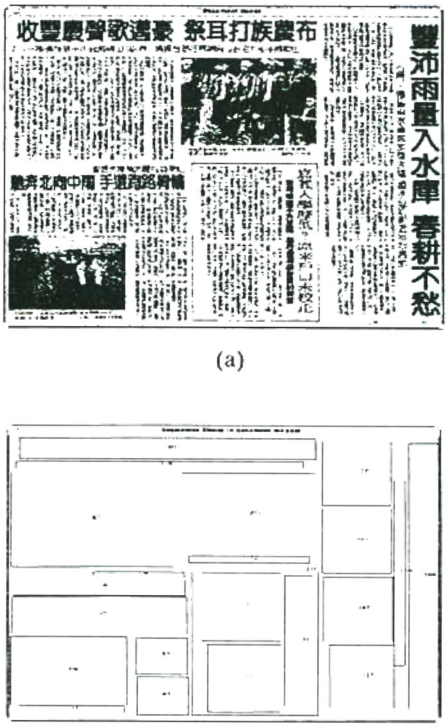
[10] R. P. Futrelle, I. A. Kakadiaris, J. Alexander, C. M. Carriero, N. Nikolakis, and J. M. Futrelle, "Understanding Diagrams in Technical Documents," *Computer*, Vol. 25, No. 7, pp. 75-78, 1992.

[11] T. Watanabe, Q. Luo, and N. Sugie, "Knowledge for Understanding Table-form Documents," *IEICE Transactions on Information and System*, Vol. E77-D, Iss. 7, pp. 761-769, 1994.

[12] L. A. Oddo, J. F. McNamara, and R. W. Smith, "An Image Understanding Approach to Reading Business Transaction Documents," *Proc. of the SPIE - The Int. Society for Optical Engineering*, Vol. 2103, pp. 142-154, 1994.

[13] K. K. Lau and C. H. Leung, "Layout Analysis and Segmentation of Chinese Newspaper Articles," *Computer Processing of Chinese and Oriental Languages*, Vol. 8, No. 1, pp. 97-114, 1994.

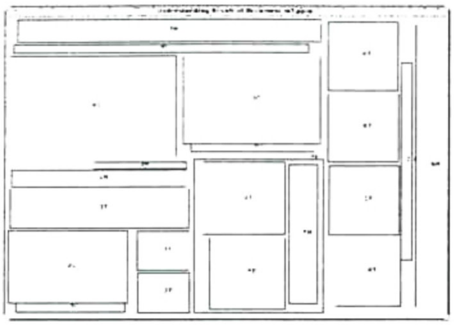
[14] Y. Lu, "Machine Printed Character Segmentation - An Overview," *Pattern Recognition*, Vol. 28, No. 1, pp. 67-80, 1995.



(a)

(b)

Figure 8. Example of experimental results. (a) A tested Chinese newspaper image with size 2950X2456. (b) The input of the segmented result of (a). (c) The result of article arrangement understanding of (a). (d) The result of character segmentation in the extracted article. (e) The result of rearrangement in the row-major single-column portrait format. (f) The result of rearrangement in the row-major two-column landscape format. (g) The result of rearrangement in the column-major single-column landscape format. (h) The result of rearrangement in the column-major two-column portrait format.



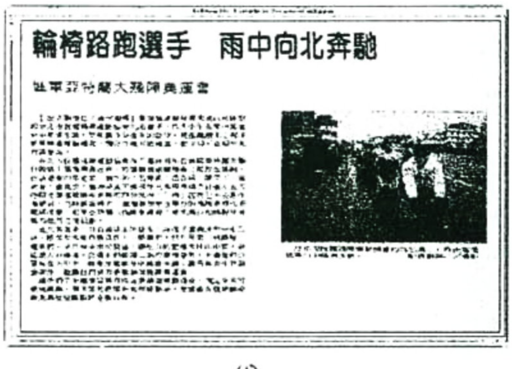
(c)

布農族打耳祭 豪邁歌聲慶豐收
豐沛雨量入水庫 春耕不愁

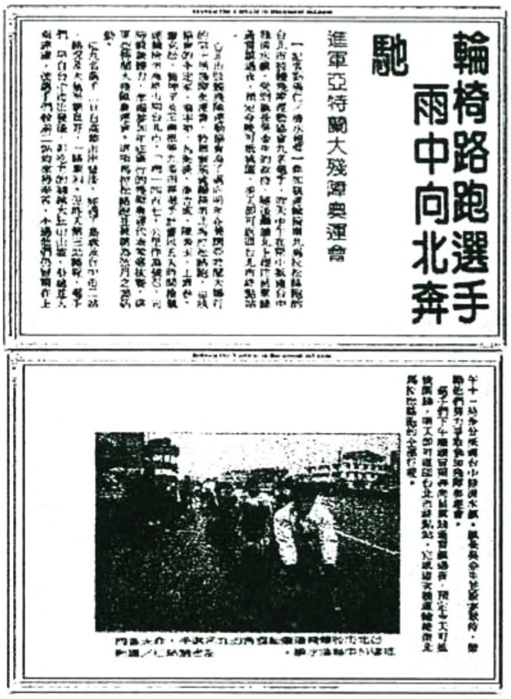
(d)

輪椅路跑選手 雨中向北奔馳
進軍亞特蘭大殘障奧運會

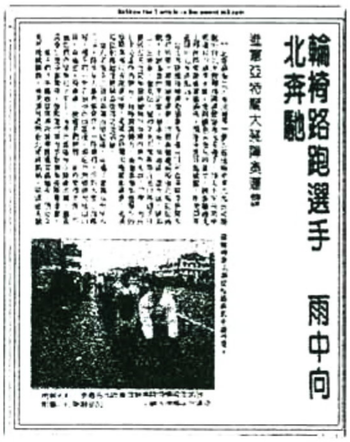
(e)



(f)



(g)



(h)

Figure 8. Example of experimental results. (a) A tested Chinese newspaper image with size 2950X2456. (b) The input of the segmented result of (a). (c) The result of article arrangement understanding of (a). (d) The result of character segmentation in the extracted article. (e) The result of rearrangement in the row-major single-column portrait format. (f) The result of rearrangement in the row-major two-column landscape format. (g) The result of rearrangement in the column-major single-column landscape format. (h) The result of rearrangement in the column-major two-column portrait format. (countinued)