PATTERN-BASED CONTENT LOSSLESS COMPRESSION OF CHINESE DOCUMENT IMAGES

Maggie M. K. Tsui¹, Alan Wee-Chung Liew¹, and Hong Yan^{1,2}

¹Department of Computer Engineering and Information Technology
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong

²School of Electrical and Information Engineering
University of Sydney, NSW 2006, Australia

ABSTRACT

Compression of scanned text document images is important in modern document management, communications and retrieval systems. However, most existing compression techniques have been studied extensively only for documents in English or similar alphabet-based languages. In this paper, we propose a content-lossless scheme for compression of Chinese text documents. This method utilizes the radical characteristics, which is unique to Chinese characters, to minimize the size of compressed documents. Our method consists of two main parts. The first part is the development of a radical pattern library. The second part is to utilize the radical pattern library to match character patterns in a document. The technique has been tested with many Chinese text document images with good results.

1. INTRODUCTION

Several standards have been developed in the past for lossless compression of binary images, including document images. They include JPEG, JBIG1, JBIG2, CCITT Groups 3 and 4, GIF, Photo CD, PNG, TIFF, etc. The JBIG standards have the best performance among these methods since they compress an image on the pattern level rather than the simple pixel level. In pattern matching based document compression techniques, the compression ratio is highly dependent on the results of document and character pattern analysis. Mohamed and Fahmy [1] have proposed a method that partitions the input document image into a near minimum number of non-overlapping rectangles and encodes the coordinates of the opposite vertices of each rectangle. Although this method is fast and the compression result is optimal according to a certain criterion, the partitioning algorithm may produce a lot of small rectangles around diagonal character strokes, which makes the coding less efficient. Yang, Yan and Yu [2] have recently proposed a content-lossless document image compression method based on structural analysis and pattern matching. This method can provide a high compression ratio and a low matching error than existing methods.

Although pattern matching techniques that are designed for alphabet-based languages can be directly

applied to the compression of Chinese text documents, they do not produce the optimal result since they do not exploit the special properties of Chinese character patterns. The Chinese language has a large character set and complex character structures. There are more than 47,000 distinct traditional characters and 3,000 to 5,000 of them are frequently used. It requires a much larger amount of space to store all Chinese character patterns than English characters in a computer.

In spite of the complexity of Chinese characters, they are composed of a limited number of radicals. The concept of radicals in Chinese characters has been routinely employed in applications such as Chinese dictionary indexing and Chinese computer input techniques. Using the concept of Chinese radicals would allow the construction of a set of distinct radical patterns that is much smaller than the set of all possible Chinese character patterns, thus allowing better compression rate in pattern-based document compression. Our compression method is an extension of Yang, Yan and Yu's method (the YYY method) and the idea is as follows. We first create a standard library of radicals for the font type used in the text document. The symbols segmented from a text document image are then compared with this radical set. If matched, their indices and positions are recorded. The remainders (un-matched patterns) will be passed to the second stage for further symbol matching and the construction of a second, document-specific, pattern library.

The paper is organized as follows. Section 2 discusses the preprocessing method for creating the radical pattern library. Section 3 describes the compression method utilizing the novel radical pattern library and the pattern-based compression technique. Section 4 presents our experiment results in comparison with CCITT Group 4 and the YYY method. Section 5 is the conclusion of the paper.

2. RADICAL PATTERN LIBRARY

Our scheme utilizes the concept of a standard radical pattern set. This standard radical pattern library is used in the first stage symbol matching, in which the patterns extracted from a text document image will first be compared with this radical pattern library. If a component extracted from the image is similar to a radical pattern in the radical pattern library, its relative position in the image and the

pattern index in the radical pattern library will be recorded.

Figure 1 shows a radical image that contains all possible radicals of Chinese characters (for MingLiu font). To create the standard radical pattern library for this font, we perform symbol segmentation and extraction so that all connected components in the radical patterns are indexed. Figure 2 shows the 8-connectivity matrix used for connected component labeling and the associated pixel coordinate system in our method. After analysis of the radical bitmap image in Figure 1, we obtain the radical pattern library in Figure 3. Associated with Figure 3 is a 1D database that contains full radical index information, including pattern numbers and relative positions of the radical patterns in the radical image. Figure 3 together with the 1D database constitute the radical pattern library. Similar libraries can be created for a multi-font document.

Figure 1: A bitmap image that contains all radicals of Chinese characters.

Figure 2: (a) The 8-connectivity and (b) the pixel coordinate system used in connected component labeling.

Figure 3: The radical library built from the radical bitmaps. All connected components in the library are indexed.

3. COMPRESSION ALGORITHM

The compression method proposed in this paper is an extension of the YYY method [2], adapted to accommodate the concept of Chinese radicals for Chinese text document compression. The various steps are described below.

3.1 Skew Correlation

Document skew correction is necessary for accurate character pattern matching [3-5]. We utilize

the cross-correlation method proposed by Yan [5] to estimate the skew angle of the scanned text document image. This method is based on the correlation between two vertical lines in a text document image. The correlation function is defined as follows:

$$R(s) = \sum_{x=1}^{X-d-1} R_1(x, s) = \sum_{y=s}^{Y-S-1} B(x+d, y+s)B(x, y) \quad (1)$$

where $R(s)$ is the accumulated cross-correlation function for all pairs of lines with a fixed distance d in the image and $R_1(x, s)$ is the cross-correlation function with a shift s between two parallel vertical lines at x and $x+d$ in the horizontal direction respectively. The skew angle can be estimated from the value of s , at which the function $R(s)$ is maximum, and from the value of d , as follows:

$$\alpha = \tan^{-1} \frac{s_{\max}}{d} \quad (2)$$

3.2. Symbol Analysis and Segmentation

After skew correction, we extract character symbols and their relative positions in the text document image. We create a primary library of segmented patterns with 1D data including page numbers, pattern numbers and their relative positions in the image after this process.

3.3. Symbol Matching and Clustering

The primary library created in the symbol analysis and segmentation step is first matched to the standard radical pattern library in the first stage processing, and the parameters of similar patterns will be coded in a 1D data set. The unmatched patterns will be passed to the second stage processing, in which representative patterns for unmatched patterns will be found and coded as a document-specific pattern library and their associated 1D parameters data set.

To construct the document-specific pattern library for the unmatched patterns, the redundancy arising from duplicated patterns needs to be minimized to maximize the compression ratio. To remove duplicated patterns, we perform structural clustering as in the YYY method.

Details of the process of symbol matching and the structural clustering are discuss below.

3.3.1 Symbol Matching

Since the structure of Chinese characters is much more complicated than that of English characters, the YYY method needs to be improved to minimize matching error and increase the compression ratio. In the following, we first introduce definitions of several parameters used for symbol matching and clustering and then discuss the

algorithm in details.

High tolerance (HT) error line: One type of error line in the error map, in which the number of matching errors in the two patterns is larger than $1/4$ or $1/5$ of the frame size (the total number of pixels in the pattern).

Low tolerance (LT) error line: Another type of error line, in which the number of matching errors in the two patterns is larger than 1 or 2.

Figure 4: Two examples of error maps ($P_1 - P_2$, containing pixels that are set in the first pattern but not in the second) and ($P_2 - P_1$, containing pixels that are set in the second pattern but not in the first). The error pixels are shown in black.

Figure 5: Different alignment modes for the comparison of patterns with different frame sizes, where pattern 1 in the image with frame sizes x_i and y_i is being compared with pattern 2 in the library with frame sizes x_j and y_j .

The matching criteria are:

- (1) Set size threshold S_1 to 1. Normalize the frame size of the connected component and compare it with the frame size of the pattern in the library in both horizontal (x) and vertical (y) directions. If both size differences are less than or equal to S_1 , go to step 2. Otherwise, compare the pattern with the next pattern in the library until the last patterns in the library is reached.
- (2) Set the alignment mode for pattern comparison according to the frame size differences between the two patterns compared. The alignment process is not needed when the frame sizes in both x and y directions of the two patterns compared are equal. Matching comparison is made as in step 3 for different modes (as shown in Figure 5), until a match is reached or all modes are rejected.
- (3) Two error maps, $P_1 - P_2$ and $P_2 - P_1$, are calculated, respectively. The matching decisions for

an alignment mode are as follows:

- i. The match is rejected if the number of LT error lines in any error maps is larger than $1/4$ or $1/5$ of the frame size in the perpendicular direction. Go to Step 4.
- ii. The matching is rejected if there exist more than two HT error lines in any error maps. Go to Step 4.
- iii. If the number of the HT error lines in an error map is less than or equal to 2 but not equal to 0, the edge connectivity of the HT error lines should be checked for match verification.
- (4) Repeat step 3 for the different alignment modes in Figure 5 until a match is found or all alignment modes are rejected.

Our matching criteria are different from the matching criteria in YYY method due to the properties of Chinese characters. Specifically, we introduce both HT and LT error line analysis to account for the increased complexity of Chinese characters. Chinese characters usually contain more strokes than English alphabets. A small shift of a pattern can cause one stroke in one pattern to be matched against a wrong stroke in another pattern of the same character, which increases the number of error pixels in the error maps. To take this stroke mismatch situation into account, a high tolerance can be set for HT error lines in Step 3. However, the larger the tolerance between patterns which belongs to the same class, the lower the matching accuracy. To avoid this problem, we therefore also check the connectivity of HT error lines in Step 3. Furthermore, the structures of Chinese characters are also more complex than English alphabets. Some characters cannot be distinguished by investigating the HT error lines only. As shown in Figure 6, the two characters would be matched if only using the HT error lines analysis. Hence, we add an LT error line tolerance in Step 3 to avoid pattern mismatching.

Figure 6: An example of two patterns differentiated by the LT error line criterion but not by the HT error line criterion.

3.3.2 Structural Clustering

In symbol matching, we usually provide a very low matching tolerance in order to ensure the accuracy. This produces many components with similar shapes in the pattern library. To minimize this redundancy, we perform structure clustering. After that, representative patterns will be extracted and used in compression. An example is shown in Figure 7. We have used similar procedure as in the YYY method for pattern clustering.

After pattern matching and the extraction of

pattern information in the image, the document-specific pattern library and the pattern information are coded using adaptive arithmetic coding with different orders and the Q-Coder, respectively. The method is similar to the one discussed in [2].

我 我我我我我我我我我
中 中中中中中中中中中中

Figure 7: Pattern clustering results, where the left pattern in each row is used to represent all similar patterns on the right.

1. EXPERIMENTAL RESULTS

We have tested our method on many scanned text documents. Figure 8 shows an example that we have tested. Table 1 shows the compression result. We can see that our method can achieve better compression rate than alternative techniques.

5. CONCLUSION

In this paper, we have developed an efficient method for content-lossless compression of Chinese text document images. Our method is based on character pattern matching and a high compression rate is achieved since a group of pixels can be represented by a single symbol. Although Chinese characters have much more patterns and complicated structures than English alphabets, they are composed of a limited number of radicals. In our methods, radicals are used to build the radical pattern library for matching and compression. The extracted patterns in a text document image are first matched to the radical pattern library in the first stage matching. The remaining un-matched patterns are then subjected to a second stage pattern-based compression scheme similar to the YYY method with modifications to accommodate the special properties of Chinese characters. By exploiting the special properties of Chinese characters in a two stage scheme, we are able to improve the compression efficiency of Chinese text document considerably.

Table 1: Compression result. “Original” is the bit-packed bitmap, “Group 4” is CCITT Group 4 standard, “YYY” is Yang, Yan and Yu’s method and “Our” is our method.

		Data Size (Bytes)				Compression Ratio		
Image		Original	Group4	YYY	Our	Group4	YYY	Our
Doc1	0	2240458	1352276	95794	77839	1.65	2.57	2.83
Doc1	1	2240458	1063168	9747	8330	2.11	22.62	26.47
Doc2	0	2240458	1285916	91806	74607	1.74	2.40	2.95
Doc2	1	2240458	1241024	57472	47451	1.81	3.84	4.65
Doc3	0	2240458	1306484	85821	69714	1.71	2.57	3.16
Doc3	1	2240458	1294336	72915	59233	1.73	3.02	3.72
Doc3	2	2240458	1280844	71799	57141	1.75	3.07	3.86
Doc3	3	2240458	1161744	43932	36606	1.93	5.02	6.11

驢心母煤坐日起行

《吉林教育》1974年第(20)期(20-21)

[illegible]

在過去只要為時、勢、地、宜都可以被解釋為「除了地理和商會外」，所以當時的地理師、命理師、堪輿師甚至算命的人都可以憑著玄秘的經書、按圖索驥、移星拿斗和推排以此為生計。事實上，這一切經驗與智慧都係前賢們訂下的關防以謹防後人的誤用，故其秘訣、禁忌、與此也互斥的觀點，皆被社會大眾認為虛構和迷信而為等等，無非數此輩的自甘墮落。

[illegible][illegible]

第貳、在學理與實踐的討論中，香港傳媒以「網際」的提法有別於「網際」，而「網際」又與「網際」不同。在「網際」中，「網」與「際」是兩個不同的概念，「網」是指「網際」的「網」，而「際」是指「網際」的「際」。在「網際」中，「網」與「際」是兩個不同的概念，「網」是指「網際」的「網」，而「際」是指「網際」的「際」。

Figure 8: Example of a scanned page of Chinese text document image.

Acknowledgement: The work described in this paper was fully supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project CityU 1128/01E).

6. REFERENCES

- [1] S. A. Mohamed and M. M. Fahmy, "Binary Image Compression Using Efficient Partitioning into Rectangular Regions", IEEE Trans. on Communications, Vol. 43, No. 5, May 1995, pp.1888-1893.
- [2] Y. Yang, H. Yan, D. Yu. Content-lossless document image compression based on structural analysis and pattern matching. Pattern Recognition 33, 2000, pp.1277-1293.
- [3] L. O'Gorman. The Document Spectrum for Page Layout Analysis. IEEE Transaction on Pattern Analysis and Machine Intelligence, 1993, 15(11): pp. 1162-1173.
- [4] D. S. Le, G. R. Thoma and H. Wechsler. Automated Page Orientation and Skew Angle Detection for Binary Document Images. Pattern Recognition, 1997, 27(10): pp1325-1344.
- [5] H. Yan. Skew Correction of Document Image Using Interline Cross-Correlation. Computer Vision Graphics Image Process: Graphical Models and Image Process, 1993, 55(6), pp. 538-534.