

# Word Searching in CCITT Group 4 Compressed Document Images

Yue Lu, Chew Lim Tan

Department of Computer Science, School of Computing  
National University of Singapore, Kent Ridge, Singapore 117543  
{luy,tancl}@comp.nus.edu.sg

## Abstract

*In this paper, we present a compressed pattern matching method for searching user queried words in the CCITT Group 4 compressed document images, without decompressing. The feature pixels composed of black changing elements and white changing elements are extracted directly from the CCITT Group 4 compressed document images. The connected components are labeled based on a line-by-line strategy according to the relative positions between the changing elements of the current coding line and the changing elements of the reference line. Word boxes are bounded by merging the connected components. A two-stage matching strategy is constructed to measure the dissimilarity between the template image of the user's query word and the words extracted from document images. Experimental results confirmed the validity of the proposed approach.*

## 1. Introduction

The problem of compressed pattern matching was introduced by Amir and Benson[1] to perform pattern matching directly in a compressed text without any decompressing. For a given text  $T$  and pattern  $P$ , the usual approach to finding the occurrences of  $P$  in compressed  $T$  is to search  $P$  in the decompressed text  $\psi(\varepsilon(T))$ , where  $\varepsilon$  and  $\psi$  are complementary encoding and decoding functions. On the other hand, the aim of compressed pattern matching is to search  $\varepsilon(P)$  in the compressed text  $\varepsilon(T)$ . In recent years, many research efforts have been invested in the attempts of matching patterns in various compressed text[2, 3]. Their experiments found that the compressed pattern matching was generally faster than the method of decompressing followed by an ordinary pattern matching. For document image processing, the research of duplicate document detection[4] and OCR[5] on compressed images has been

reported recently. In this paper, we will concentrate on the issue of searching words in CCITT Group 4 compressed document images.

With the rapid growth of digital libraries, a huge number of documents are accessible in the Internet. These documents may be in either text format (electronic machine-readable code) or image format. Most of the newly generated documents are in the text format. But due to various reasons, many documents scanned from outdated books, magazines, periodicals and students' theses are stored in image format. We find that many documents provided by the websites of digital libraries and journal/conference publishers are in such image format.

Keyword searching is a practical and widely-used method to retrieve information from document archives. It is normally trivial to search words in the text format documents. But searching words from image format documents provides an even greater challenge than that from text format documents.

One commonly used method for word searching in a document image is to convert the document image to its machine-readable text using optical character recognition (OCR) first and then search words in the text format document. However, OCR is still not perfect for the moment, which is also the main reason why these documents are stored in the image format. Image based approach becomes an alternative way to directly search keywords in the document images without OCRing the entire document images[6]. Several methods regarding this issue have been reported in the past years[7, 8, 9].

Furthermore, in order to save storage space and speed up the transmission in the Internet, many document images are stored and transmitted in compressed formats (e.g. CCITT Group 3/4, JPEG, and JBIG2, etc). The CCITT Group 4 standard is one of the popularly used compression standards for document images. We find that, more and more document images packed in PDF files with compression by the CCITT Group 4 standards are spread worldwide through

the Internet. The considerable advantages will be realized if we can carry out word searching directly on the compressed images. In this paper, we present a method of searching user queried words from the CCITT Group 4 compressed document images, without decompressing. Experimental results with the document images provided by the Digital Library of our university show that the proposed approach has achieved a promising performance.

The remainder of this paper is organized as follows. Section 2 briefly gives the system overview. Section 3 describes feature extraction from CCITT Group 4 compressed images, and word box bounding based on the feature pixels composed of the changing elements. Section 4 discusses the method of word object matching. Section 5 gives the experimental results. Finally, conclusions are drawn in Section 6.

## 2. System Overview

Figure 1 illustrates the system diagram. The feature pixels composed of the black changing elements and white changing elements are extracted directly from the CCITT Group 4 compressed document images. At the same time, the connected components are labeled based on the line-by-line strategy according to the relative position between the changing elements of the current coding line and the changing elements of the reference line. The word boxes are bounded by merging the connected components according to their relative positions and sizes.

When a user keys in a query word, the system will generate its template image by synthesizing the bitmap images of each character in the word one after another. The template image is represented by the features of the changing elements like that in CCITT Group 4 compression standards. Then the template image is matched with each of the bounded word objects in the document to find whether the word in the document is the user's query word.

To meet the requirement for fast word matching between the template image of the user's query word and the words extracted from document images, a two-stage matching strategy is constructed to speed up the process. The first stage is a coarse-matching procedure. In the second stage, a modified Hausdorff distance is employed to measure the dissimilarity between them.

## 3. Feature Extraction and Word Box Bounding

### 3.1. Feature Extraction from CCITT Group 4 Compressed Images

The CCITT Group 4 coding scheme for binary images uses a two-dimensional line-by-line coding method[10],

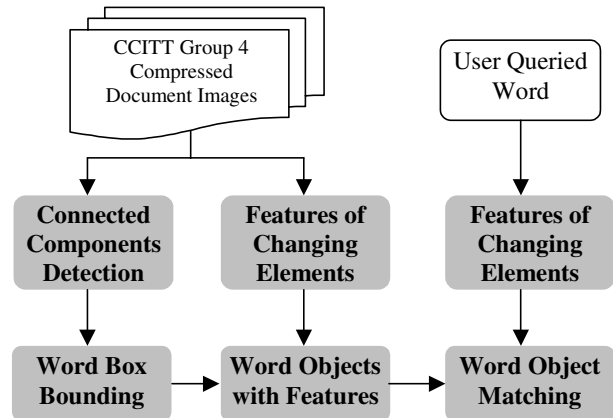


Figure 1. System diagram

in which the position of each changing element on the current coding line is coded with respect to the positions of corresponding reference elements situated on either the coding line or the reference line which is immediately above the coding line. A changing element is defined as an element whose color (i.e. black or white) is different from that of the previous element along the same line.

In the CCITT Group 4 standards, there are three coding modes: Pass Mode(P), Vertical Mode(V(0), VR(1), VR(2), VR(3), VL(1), VL(2), VL(3)), and Horizontal Mode(H). One of the three coding modes is chosen, according to the changing element and its reference elements, to code the position of each changing element along the coding line.

According to the CCITT Group 4 standards, each coded position indicates that the current pixel color is different from its previous pixel, except for the following coded positions of the pass mode. In our work, we give our attention to these changing elements in the CCITT Group 4 compressed document images, because they can be easily obtained from the compressed images directly.

In the document images, black pixels generally represent the characters' strokes. We define the black changing elements to be corresponding to the changing elements that change from white pixels to black pixels. On the other hand, the white changing elements correspond to the changing elements that change from black pixels to white pixels. The black changing elements and white changing elements can be easily discriminated while we extract the feature pixels from the compressed images.

Figure 2(a) gives a part of an original document image and Figure 2(b) and (c) demonstrate respectively the black changing elements and white changing elements extracted directly from its corresponding CCITT Group 4 compressed image, in which the changing elements following a pass mode are removed because they are not the actual changing

points according to the CCITT Group 4 standards. It can be seen that the features composed of the changing elements are roughly similar to the characters' profiles.

In the succeeding discussion, the changing elements will be utilized to segment and bound the word objects, and be used for measuring the similarity of two document images.

### 3.2. Word Bounding in CCITT Group 4 Compressed Images

While extracting the changing elements from the compressed images, the connected component detection can be performed, at the same, by analyzing the relative positions among changing elements of the current lines and the reference lines. Then the word objects can be bounded by merging the connected components.

First, all of the connected components in the document image are labeled. The connected components are detected line by line, which is similar to the procedure of compressing/decompressing the CCITT Group 4 document images. In a coding line, a black changing element  $B_C$  and its following white changing elements  $W_C$  are represented by a run  $Run(B_C, W_C)$ . Likewise, a black changing element  $B_R$  in the reference line and its following white changing element  $W_R$  are represented by a run  $Run(B_R, W_R)$ . Each run can be considered as a connected component whose leftmost and rightmost coordinates are equal to black changing element and white changing element respectively. A new connected component is generated with respect to a run  $Run(B_C, W_C)$  if it is not overlapped with any run in the reference line. On the other hand, if  $Run(B_C, W_C)$  is overlapped with  $Run(B_R, W_R)$ , it is merged to the component with respect to the run  $Run(B_C, W_C)$ . The two runs are considered to be overlapping, if they meet one of the following conditions:

$$B_R - 1 \leq B_C \leq W_R \quad (1)$$

$$B_R \leq W_C \leq W_R + 1 \quad (2)$$

Two overlapping examples are illustrated in Figure 3. Based on above processing, the connected components are detected as shown in Figure 2(d).

Second, the connected components with small areas are eliminated as noise. The connected components larger than a threshold, which may be graphics or table regions, are ignored too. Then, the punctuation, such as commas and full stop, are marked based on their relative positions with their preceding connected components.

Finally, the connected components are merged to generate the bounding boxes of word objects, according to their relative positions and sizes. Figure 2(d) demonstrates the results of word bounding boxes.

Since the Gibbs phenomenon res might guess that the magnitude n nonrectangular window with tapt characteristic, the window must e

(a) Original image

Since the Gibbs phenomenon res might guess that the magnitude n nonrectangular window with tapt characteristic, the window must e

(b) Black changing elements

Since the Gibbs phenomenon res might guess that the magnitude n nonrectangular window with tapt characteristic, the window must e

(c) White changing elements

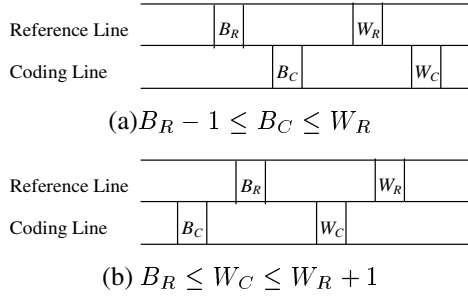
Since the Gibbs phenomenon res might guess that the magnitude n nonrectangular window with tapt characteristic, the window must e

(d) Bounding boxes of connected components

Since the Gibbs phenomenon res might guess that the magnitude n nonrectangular window with tapt characteristic, the window must e

(e) Word bounding boxes

Figure 2. Feature extraction and word bounding in the compressed image



**Figure 3. Examples of overlapping runs**

## 4. Word Image Matching

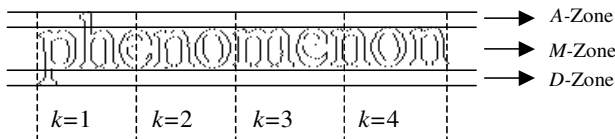
Both the word object of the compressed image and user queried word are represented by their changing elements(both black and white) to perform the word matching process. To meet the requirement for fast processing, a two-stage matching strategy is constructed to speed up the word matching process. The first stage, which is a coarse-matching procedure, is simple and fast to execute, but is not powerful enough to distinguish between similar patterns. In the second stage, a modified Hausdorff distance is employed to match two word images.

### 4.1. Coarse-matching

To roughly evaluate the dissimilarity between the word object  $P$  extracted from the document image and  $Q$  of the user's query word,  $P$  is normalized to the size of  $Q$  first.

A character can be divided into three parts, i.e. ascender, mid-zone and descender. Similarly, we can divided a word object into different zones, namely  $A$ -zone,  $M$ -zone and  $D$ -zone, as showed in Figure 4. The boundary line between different zones can be computed according the characters of the user's query word.

The distance of each zone is calculated respectively first. Each zone( $A$ -zone,  $M$ -zone and  $D$ -zone) is divided from left to right into  $K$  sub-zones of equal size. Here  $K = 4$  in our experiments. The ratio of the number of changing elements to the area of sub-zone is employed as the feature. The distance between two corresponding sub-zones of two



**Figure 4. Different zones in the word object**

word objects is calculated as:

$$d_k^X(P, Q) = |r_k^X(P) - r_k^X(Q)| \quad k = 1, 2, \dots, K \quad (3)$$

where  $X$  represents the  $A$ -zone,  $M$ -zone or  $D$ -zone.  $r_k^X$  is the ratio of the number of changing elements to the area of the  $k$ th sub-zone of  $X$ -zone. Then the distance of  $X$ -zone between  $P$  and  $Q$  is defined as:

$$D^X(P, Q) = \sum_{k=1}^K d_k^X(P, Q) \quad (4)$$

The distance between  $P$  and  $Q$  is defined as

$$D(P, Q) = \max(D^A(P, Q), D^M(P, Q), D^D(P, Q)) \quad (5)$$

If  $D(P, Q)$  is greater than a threshold  $\delta$ ,  $P$  is not similar to  $Q$ , thus no further process is needed. Otherwise,  $P$  and  $Q$  are further matched by the method based on the weighted Hausdorff distance.

### 4.2. Word Image Matching Based on Weighted Hausdorff Distance

The Hausdorff distance[11] measures the degree of mismatch between two point sets, which has been widely applied in two-dimensional image matching, especially in the area of object matching. Dubussion and Jain [12] presented the modified Hausdorff distance(MHD) measure by employing the summation operator over all distance, rather than the maximum operator in the traditional Hausdorff distance. The MHD was proposed for the purpose of object matching in the areas of computer vision, object recognition and image analysis. In our previous paper[9], we proposed the Weighted Hausdorff distance(WHD) to investigate the application of Hausdorff distance to word object matching, in which the contribution of different parts of the word object to the Hausdorff distance is not the same. The experiments have confirmed that the performance of WHD is better than that of HD and MHD for the purpose of word image matching. For further detail, readers may refer to [9].

In the second matching stage, we apply the MHD to measure the dissimilarity between the user's query word and the word objects extracted from the document images, if they appear to be similar in the first matching stage.

## 5. Experimental Results

Experiments were conducted to verify the validity of the proposed approach to searching user queried words from the CCITT Group 4 compressed document images.

The document images are selected from the scanned books and students' theses that are provided by the

Digital Library of the National University of Singapore. The document images are packed in the PDF files, and compressed by CCITT Group 4 standards.

To evaluate the performance of the system, 95 images of scanned books and 324 images of scanned students' theses are included in the test. 40 keywords are selected to search their corresponding words from the document images. The system achieves an average precision ranging from 93.72% to 98.17% and an average recall ranging from 85.16% to 95.86% depending on different thresholds of the dissimilarity measurement.

To compare the processing times, the traditional word searching method is utilized to compare with the proposed algorithm. In the former method, the CCITT Group 4 document images are decompressed first. Then, the connected component analysis, word bounding and word matching are carried out on the decompressed images. Experiments show that the proposed algorithm is approximately 2.1 times faster than a decompression followed by a traditional word matching approach. This could be explained by the fact that the proposed approach avoids the pixel-level processing for analyzing the connected components and extracting word features, whereas the processing on image pixels in the traditional approach is quite time consuming.

## 6. Conclusions

Document image has become a widespread information format of storage and transmission in the Internet, in which a huge amount of document images had been compressed by CCITT Group 4 standards. There is thus significant meaning to develop the methods of directly searching user queried words from these documents.

In this paper, we present a method of searching words from the CCITT Group 4 compressed document images, without decompressing. The feature extraction, connected component analysis and word bounding are performed directly in the CCITT Group 4 compressed document images. A metric is proposed to measure the dissimilarity between the template image of user's query word and the words extracted from document images. Experimental results with the document images captured from students' theses show that the proposed approach has achieved a promising performance, and that its speed is faster than decompressing and searching afterwards.

## Acknowledgements

This research is jointly supported by the Agency for Science, Technology and Research, and Ministry

of Education of Singapore under research grant R-252-000-071-112/303.

## References

- [1] A. Amir, G. Benson, Efficient two-dimensional compressed matching, *Processings of Data Compression Conference*, DCC'92, Snowbird, Utah, 1992, pp.279-288.
- [2] T. Kida, M. Takeda, A. Shinohara, et al. Shift-An approach to pattern matching in LZW compressed text, *Proceedings of 10th Annual Symposium on Combinatorial Pattern Matching*, LNCS 1645, pp.1-13, 1999.
- [3] S. T. Klein, D. Shapira, Searching in compressed dictionaries, *Processings of Data Compression Conference*, DCC2002, Snowbird, Utah, 2002, pp.142-151.
- [4] J. J. Hull, Document matching on CCITT group 4 compressed images, *Proceedings of SPIE, Document Recognition IV(L.M Vincent and J. J Hull edit)*, San Jose, CA, USA, 1997, vol.3027, pp.82-87.
- [5] U. V. Marti, D. Wymann and H. Bunke, OCR on compressed images using pass modes hand hidden Markov models, *Proceedings of IAPR Workshop on Document Analysis Systems*, Rio de Janeiro, Brazil, 2000, pp.77-86.
- [6] D. Doermann, The indexing and retrieval of document images: a survey, *Computer Vision and Image Understanding*, vol.70, No.3, 287-298, 1998.
- [7] F. R. Chen, L. D. Wilcox and D. S. Bloomberg, Detecting and locating partially specified keywords in scanned images using hidden Markov models, *Proc. of the International Conference on Document Analysis and Recognition*, 1993, pp. 133-138.
- [8] S. Kuo and O. F. Agazzi, Keyword spotting in poorly printed documents using pseudo 2-D hidden Markov models, *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.16, No.8, pp. 842-848, 1994.
- [9] Y. Lu, C. L. Tan, W. Huang and L. Fan, An approach to word image matching based on weighted Hausdorff distance, *Proceedings of the 6th International Conference on Document Analysis and Recognition*, 2001, Seattle, USA, pp.921-925.
- [10] W. Kou, *Digital image compression algorithms and standards*, Kluwer Academic Publishers, 1995.
- [11] D. P. Huttenlocher, G.A. Klanderman, and W. J. Rucklidge, Comparing images using the Hausdorff distance, *IEEE trans. Pattern Analysis and Machine Intelligence*, Vol.15, No.9, 850-863, 1993.
- [12] M. P. Dubuisson and A. K. Jain. A modified Hausdorff distance for object matching, *Proceedings of 12th Int. Conf. Pattern Recognition*, Jerusalem, Israel, 1994, pp.566-568.