



PERGAMON

Pattern Recognition 36 (2003) 987–996

PATTERN
RECOGNITION

THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY

www.elsevier.com/locate/patcog

Document retrieval from compressed images

Yue Lu, Chew Lim Tan*

Department of Computer Science, School of Computing, National University of Singapore, Kent Ridge, 117543 Singapore

Received 27 November 2001; accepted 18 April 2002

Abstract

With the emergence of digital libraries, more and more documents are stored and transmitted through the Internet in the format of compressed images. It is of significant meaning to develop a system which is capable of retrieving documents from these compressed document images. Aiming at the popular compression standard-CCITT Group 4 which is widely used for compressing document images, we present an approach to retrieve the documents from CCITT Group 4 compressed document images in this paper. The black and white changing elements are extracted directly from the compressed document images to act as the feature pixels, and the connected components are detected simultaneously. Then the word boxes are bounded based on the merging of the connected components. Weighted Hausdorff distance is proposed to assign all of the word objects from both the query document and the document from database to corresponding classes by an unsupervised classifier, whereas the possible stop words are excluded. Document vectors are built by the occurrence frequency of the word object classes, and the pair-wise similarity of two document images is represented by the scalar product of the document vectors. Nine groups of articles pertaining to different domains are used to test the validity of the presented approach. Preliminary experimental results with the document images captured from students' theses show that the proposed approach has achieved a promising performance.

© 2002 Pattern Recognition Society. Published by Elsevier Science Ltd. All rights reserved.

Keywords: Document image retrieval; Compressed image; Object matching; Document similarity; Weighted Hausdorff distance

1. Introduction

With advances in Internet applications and technologies, and particularly with the development of the world wide web, a growing number of documents are published and accessed on-line. Nowadays, more and more information of various types becomes available on the Internet and Web. For example, a lot of digital libraries provide a broad collection of documents. These documents may be in either text format or image format. Undoubtedly, the text (electronic machine-readable code) format facilitates not only the storage and transmission of documents in the Internet, but also document retrieval. There has been active research on Web

content extraction using text-based techniques, on which the text retrieval community has made significant progress.

Although most of the newly generated documents are in the text format, billions of volumes distributed in the classical libraries worldwide are in paper format such as books, magazines, periodicals and students' theses. The need is evident to transfer these paper documents to their digital domain. As an approach to automatically transferring paper documents to their text format, optical character recognition (OCR) has its inherent weaknesses. In particular, manually correcting the OCR results is typically not cost effective for transferring a huge amount of paper documents to their text format. Therefore, storing documents in the image format should be an alternative way.

This poses new challenges for the research of document retrieval. To retrieve such document images, at the present time, one has to painstakingly download them and open/decompress them one by one to see their relevance. A content-based document image retrieval system is required

* Corresponding author. Tel.: +65-874-2900; fax: +65-779-4580.

E-mail addresses: luy@comp.nus.edu.sg (Y. Lu),
tancl@comp.nus.edu.sg (C.L. Tan).

to retrieve effectively and efficiently information from these document image repositories. Such a system should be able to return, in ranked order, the documents that are most likely relevant to the users' query by the degree of similarity with the query image. Information retrieval from such document images provides an even greater challenge than that from text documents because it is almost impossible to extract and utilize the semantic information directly from document images as what we do in the text retrieval.

Furthermore, in order to save storage space and speed 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.). Nowadays, great deals of document images are compressed by CCITT Group 4 recommendations on the world wide web. For example, many document images we can access in the Internet are packed in the PDF files with the compression format of CCITT Group 4.

In this paper, we present a method of retrieving document images from the CCITT Group 4 compressed images. The feature pixels composed of the changing elements are extracted directly from the compressed document images. 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 position and size. The bounded word images constitute the word objects of the documents, whereas the possible stop words are excluded. An unsupervised classifier is utilized to cluster all of the word objects of the two documents. Document vectors are built by the occurrence frequency of the word object classes, and the pair-wise similarity of two document images is represented by the scalar product of the document vectors. A weighted Hausdorff distance (WHD) is proposed to measure the similarity of the word objects. Experimental results with the document images captured from students' theses show that the proposed approach has achieved a promising performance.

The remainder of this paper is organized as follows. Section 2 briefly surveys the related research works, and the system structure proposed in our system is introduced in Section 3. Section 4 describes feature extraction from CCITT Group 4 compressed images, and word box bounding based on the feature pixels composed of the changing elements. Section 5 discusses the method of word object matching. Section 6 presents the document vector and similarity measure. Section 7 gives the experimental results. Finally, conclusions and future works are discussed in Section 8.

2. Related work

Many approaches have been proposed for categorization and retrieval of machine-readable documents over the past decades [1–3]. They have relied on self-evident utility of

words, sentence and paragraphs for sorting, categorizing and retrieving texts. Furthermore, various means of suppressing uninformative word, removing prefixes, suffixes and endings, interpreting inflected forms, etc. have been developed. However, the traditional text retrieval system is not applicable for the document image retrieval.

One commonly used method for document image retrieval is to convert the document image to its machine-readable text using OCR first and then use the usual text retrieval techniques. However, OCR is still not perfect for the moment, which results in errors in recognition. As a consequence, manually correcting the OCR results is inevitable. It is typically not cost effective for transferring a huge amount of paper documents to their text format. Although some researchers have tried to retrieve document with toleration of recognition and segmentation errors caused in the OCR procedure [4,5], the retrieval performance is affected by both the OCR and character segmentation undoubtedly. Furthermore, the research on document layout analysis and understanding is still immature, especially for the documents with complex layout. These discourage the strategy of document image retrieval by OCRing the entire documents.

An alternative approach is to retrieve these documents based on their image contents directly without OCR [6]. A similar research is content-based image retrieval (CBIR). In these systems [7–11], either color features or texture features extracted from the images are utilized to retrieve pictures from the database. However, document images are different from the landscape pictures. Document image retrieval cannot rely on the color or texture features.

Several researchers have made the attempt to retrieve information from document images directly. For example, Chen and Bloomberg [12] described a method for automatically selecting sentences for creating a summary from a document image without recognition of the characters in each word. They built word equivalence classes by using a rank blur hit-miss transform to compare word images and use a statistical classifier to determine the likelihood of each sentence being a summary sentence. Liu and Jain [13] addressed an approach to image-based form document retrieval. They proposed a similarity measure for forms that is insensitive to translation, scaling, moderate skew and image quality fluctuations, and developed a prototype form retrieval system based on the proposed similarity measure. Niyogi and Srihari [14] described an approach to retrieve information from document images stored in a digital library by means of knowledge-based layout analysis and logical structure derivation techniques, in which significant sections of document such as title are utilized.

Moreover, in order to save the storage space and speed during the transmission in the network, document images are stored and transmitted in the compressed formats in general. The CCITT Group 4 standard is one of the popularly used compression standards for document images. Nowadays, 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 document retrieval directly on the compressed images.

Interesting research such as duplicate document detection and OCR on the compressed images has been reported recently. Hull [15] has proposed a method to detect equivalent document images by matching the pass mode codes extracted from CCITT compressed document images. He created a feature vector that counts the number of pass mode codes in each cell of a fixed grid in the image and equivalent images are located by applying the Hausdorff distance to the feature vectors. Marti et al. [16] presented a system which is capable of reading machine printed text in images compressed by CCITT Group 3 two-dimensional coding scheme. They utilized the points marked by the pass mode as the features for a hidden Markov model-based recognizer.

3. System overview

In this paper, we present a method of retrieving documents from the CCITT Group 4 compressed document images. Fig. 1 illustrates the diagram of the system for measuring the similarity of two CCITT Group 4 compressed document images.

The changing elements of the compressed images, which are the feature pixels of the subsequent processing, are extracted directly from the compressed images first. 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.

Prior to bounding the word boxes based on the connected components according to their relative position and size, some pre-processing is carried out. First, those connected components with too small area or too large area, which are noise, graphics or table regions, are excluded for further processing. Then the punctuations, such as commas and full stops, are marked and eliminated as well based on their relative positions with their preceding connected components.

The stop words are detected according to their shapes, and are excluded from further processing. The remainder of the bounded word images constitute the word objects of the documents. An unsupervised classifier is employed to cluster all of the word objects of the two documents. The occurrence frequency of each class in each document, which is normalized by dividing it by the total number of word objects in the document, builds the document's vector. Finally, the pair-wise similarity of two document images is represented by the scalar product of the document vectors.

To meet the requirement for fast processing, a hierarchical system is constructed to speed up the word object matching process. The first stage is a coarse-matching procedure. In the second stage, a modified Hausdorff distance (MHD) is employed to measure the similarity of the two word object images.

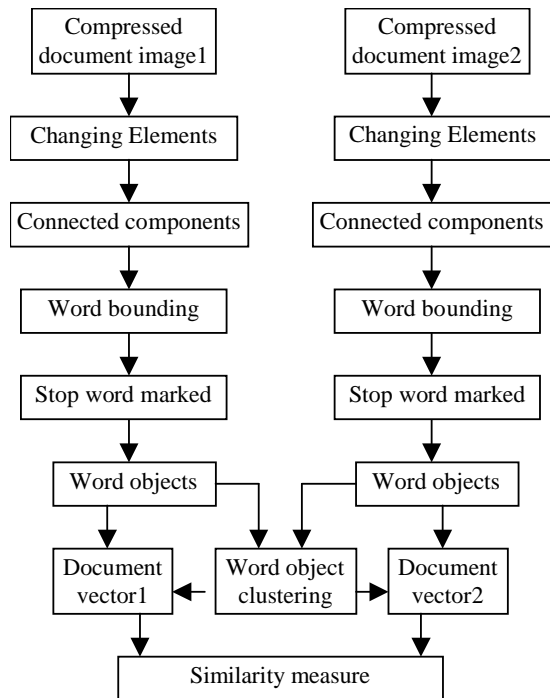


Fig. 1. Diagram of document image similarity measure.

4. Feature extraction from CCITT Group 4 compressed images and word box bounding

4.1. 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 [17], 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 standard, 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 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.

Fig. 2 gives a part of an original document image and Fig. 3(a) demonstrates the changing elements extracted

**Response digital filters, various wi
functions are the Hamming Windo
the Triangular Window and the Bl
digital filters are approximated usi**

Fig. 2. Original image.

Response digital filters, various wi
functions are the Hamming Windo
the Triangular Window and the Bl
digital filters are approximated usi

(a)

Response digital filters, various wi
functions are the Hamming Windo
the Triangular Window and the Bl
digital filters are approximated usi

(b)

Response digital filters, various wi
functions are the Hamming Windo
the Triangular Window and the Bl
digital filters are approximated usi

(c)

Fig. 3. Feature extraction and word bounding in the compressed image: (a) feature representation by changing elements, (b) bounding boxes of connected components, and (c) word bounding boxes.

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 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 document images, black pixels generally represent the characters' strokes. We define the black changing elements to correspond 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.

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

4.2. Word bounding in CCITT Group 4 compressed images

The strategy used for segmenting characters from document images can be divided into two categories, viz., top-down and bottom-up [18]. Each has its own advantages and disadvantages. The top-down approach breaks down the document images into different blocks such as headlines, text lines, graphics, tables, etc. Then the characters are extracted from the headlines and text lines. The knowledge of the document layout, therefore, is much important during the processing. This approach is simple and fast. It is very effective for processing the documents that have a specific format.

On the other hand, in the bottom-up approach all of the connected components are detected first. The connected components are merged according to their relative location and size. This approach is possible to develop algorithms which are applicable to a variety of documents, but it is time consuming.

In the previous literature, the top-down approach has been mainly employed for segmenting the machine printed characters and words, based on the use of X - Y projection. However, this approach is not capable of coping with documents with graphics and tables. To deal with the document images with complex layout, a bottom-up algorithm is utilized in this present approach to bound the word objects in the images. Moreover, the word objects are bounded based on the black and white changing elements that are directly extracted from the compressed images.

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, the pixels between one black changing element B_C and its following white changing elements W_C undoubtedly belong to one connected component.

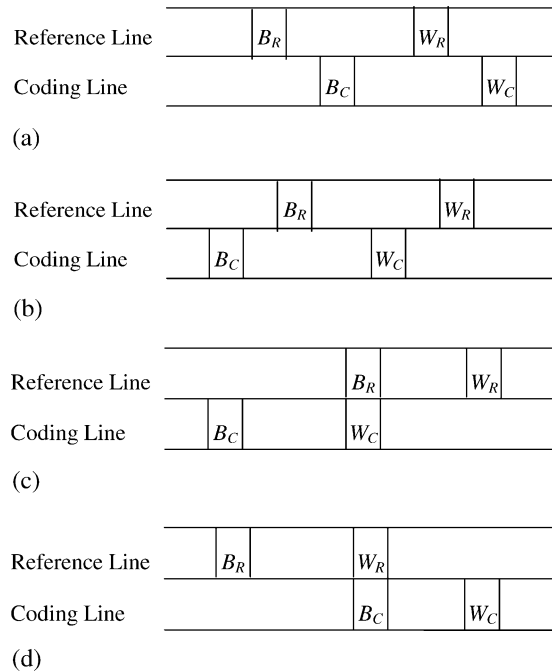


Fig. 4. Some typical examples in which the changing elements in the coding line belong to the same connected components with the changing elements in the reference line: (a) B_C is located between B_R and W_R , (b) W_C is located between B_R and W_R , (c) W_C just underlies B_R , and (d) W_C just underlies W_R .

Moreover, we can assign the same label to the changing elements in the coding line if their relative positions with the changing elements B_R and W_R in the reference line indicate they belong to the same connected component. Some examples are illustrated in Fig. 4. Based on above processing, the connected components of Fig. 3(a) can be detected, as shown in Fig. 3(b).

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 position with their preceding connected components.

Finally, the connected components are merged to generate the bounding boxes of word objects, according to their relative position and size. Fig. 3(c) demonstrates the results of word bounding boxes.

Furthermore, some common words called *stop words* in the linguistics are detected according to their shapes. Although width estimates of single words are unreliable for stop word identification, stop words tend to be short and are composed of a few characters. Suppose that w and h represent the width and height of a word object, respectively. If w/h of a word object is less than a certain λ , it may be a stop word. In our experiments, λ is set as 2.0. These stop words

will be excluded from further processing such as the word object clustering and the document feature computation.

5. Word object matching

To meet the requirement for fast processing, a hierarchical system is constructed to speed up the word object 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 MHD is employed to match two word object images.

5.1. Coarse matching

A character can be divided into three parts, i.e. ascender, mid-zone and descender. Similarly, we can divide a word object into different zones, namely *A*-zone, *M*-zone and *D*-zone. The top-left positions and bottom-right positions of the connected components in the word object image are utilized to decide the boundaries between two different zones, as shown in Fig. 5.

To roughly evaluate the similarity of the word object *P* and the word object *Q*, 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 *M* sub-zones. Here *M* = 4 in our experiments. The ratio of changing elements to the area of sub-zone is employed as the feature. The distance between two corresponding sub-zones of two word objects is calculated as

$$s_m^X(P, Q) = r_m^X(P) - r_m^X(Q), \quad m = 1, 2, \dots, M, \quad (1)$$

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

$$S^X(P, Q) = \sum_{m=1}^M s_m^X(P, Q). \quad (2)$$

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

$$S(P, Q) = \max(S^A(P, Q), S^M(P, Q), S^D(P, Q)). \quad (3)$$

If *S*(*P*, *Q*) is greater than a threshold δ , *P* and *Q* belong to different classes. Otherwise, *P* and *Q* are further classified by the method based on the WHD.

5.2. Character image matching based on WHD

Hausdorff distance has been widely applied in two-dimensional image matching, especially in the area of object matching [19,20].

The distance (e.g. Euclidean distance) between two points *a* and *b* is defined as $d(a, b) = \|a - b\|$, and the distance between a point *a* and a finite point set $B = \{b_1, \dots, b_{N_b}\}$ is commonly defined as

$$d(a, B) = \min_{b \in B} \|a - b\|. \quad (4)$$

Given two finite point sets $A = \{a_1, \dots, a_{N_a}\}$ and $B = \{b_1, \dots, b_{N_b}\}$, the Hausdorff distance is defined as

$$H(A, B) = \max(h(A, B), h(B, A)), \quad (5)$$

where $h(A, B)$ and $h(B, A)$ represent the directed distance between two sets *A* and *B*. The directed distance $h(A, B)$ is traditionally defined as

$$h(A, B) = \max_{a \in A} d(a, B) = \max_{a \in A} \min_{b \in B} \|a - b\|. \quad (6)$$

The function $h(A, B)$ identifies the point $a \in A$ that is farthest from any point of *B* and measures the distance from *a* to its nearest neighbor in *B*. The Hausdorff distance $H(A, B)$ measures the degree of mismatch between two point sets *A* and *B*.

Dubussion and Jain [20] presented the MHD measure by employing the summation operator over all distance, rather than the maximum operator:

$$h_{MHD}(A, B) = \frac{1}{N_a} \sum_{a \in A} d(a, B). \quad (7)$$

The MHD was proposed for the purpose of object matching in the areas of computer vision, object recognition and image analysis.

As discussed above, a word object can be divided into different parts, i.e. *A*-zone, *M*-zone or *D*-zone. We propose a 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 directed distance of WHD is computed as

$$h_{WHD}(A, B) = \frac{1}{N_a} \sum_{a \in A} w(a) \cdot d(a, B), \quad (8)$$

where $\sum_{a \in A} w(a) = N_a$.



Fig. 5. Different zones in the word object.

Table 1
Corpus used in the experiments

Group	Number	Topic
A	16	Programmable digital filtering based on the TMS320C30 digital signal processor
B	25	Method for multiphase equilibrium calculations
C	13	Error probability of continuous-phase FSK with discriminator detection
D	17	Frequency optimization using genetic algorithm
E	33	Adaptive equalization in teletext reception
F	20	Computation of phase and chemical equilibrium by direct search optimization
G	16	Three-dimensional model of drug delivery to brain tumors
H	17	Using marching cubes for real time surface rendering
I	15	Micro-refrigerators for cold electronics

The weight of the different zone in the word namely $w(a)$, $w(m)$ and $w(d)$ correspond to the A -zone, M -zone or D -zone, respectively. In our experiments, they are set as

$$w(a) = w(d) = 2 \times w(m). \quad (9)$$

6. Document vector and similarity

An unsupervised classifier is employed to cluster all of the word objects (except the stop words) of the two documents to K object classes. The occurrence frequency of each class in each document builds the document's vector. The occurrence frequency is normalized by dividing it by the total number of word objects in the document.

The similarity score between two document vectors is defined as their scalar product divided by their lengths. A scalar product is calculated through summing up the products of the corresponding elements. This is equivalent to the cosine of the angle between two document vectors seen from the origin. So the similarity between document X and Y will be

$$S(\vec{X}, \vec{Y}) = \frac{\sum_{k=1}^K x_k \cdot y_k}{\sqrt{\sum_{k=1}^K x_k^2 \sum_{k=1}^K y_k^2}}, \quad (10)$$

where, \vec{X} and \vec{Y} are the document vectors of image X and Y respectively. K is the dimension number of document vector, and $\vec{X} = (x_1, x_2, \dots, x_K)^T$, $\vec{Y} = (y_1, y_2, \dots, y_K)^T$.

7. Experimental results

Experiments were conducted to verify the validity of the proposed approach to retrieving documents from the CCITT Group 4 compressed document images.

The document images are selected from the scanned students' theses that are provided by the Central Library of the National University of Singapore. The document images are packed in the PDF files, and compressed by CCITT Group 4 standards. These articles address nine different kinds of topic, respectively. They are composed of the students' theses pertaining to electrical engineering,

chemical engineering, mechanical engineering, medical engineering, etc. Their topics and numbers used in the experiments are listed in Table 1.

For each group, we take the first one as the query document (reference document). Similarity of all the document images with the respective nine reference document images is carried out. To demonstrate the validity of the proposed approach, the similarity of nine query documents with the first three documents taken from each group is tabulated in Table 2, in which the similarities with document B1 and D1 are illustrated in Figs. 6 and 7, respectively. From the results, we can see that the similarity represented by the proposed approach is effective for document image retrieval.

The retrieval performance of each group are tabulated in Tables 3–5 with respect to different similarity threshold α . The system achieves an average of precision ranging from 78.12% to 96.52% and an average recall ranging from 71.45% to 97.56% depending on different similarity thresholds. The three tables show the effect of α on precision and recall. A lower α allows more relevant documents to be retrieved (hence higher recall) but at the expense of false alarm (hence lower precision). The reverse is true for a higher α . Thus, if the goal is to retrieve as many relevant documents as possible allowing irrelevant ones to be rejected by manual reading, then a lower α should be set. On the other hand, for low tolerance to false alarm, then α should be raised.

8. Conclusions and future works

With the emergence of digital libraries, document image has become a widespread information format of storage and transmission in the world wide web, in which a huge amount of document images had been compressed by CCITT Group 4 standards. There is thus an urgent need for effective document image retrieval system.

Undoubtedly, it is quite time consuming to retrieve the documents based on their image format directly, compared to the traditional document retrieval based on their text format. A feasible scheme is computing the similarity measure

Table 2
Document similarity

	A1	B1	C1	D1	E1	F1	G1	H1	I1
A1	1.0000	0.0467	0.1653	0.1755	0.3190	0.4008	0.2263	0.3529	0.2456
A2	0.4189	0.1185	0.3021	0.2442	0.2494	0.2504	0.2635	0.3021	0.2517
A3	0.4074	0.1252	0.2318	0.1303	0.3207	0.3359	0.1250	0.3639	0.3441
B1	0.1281	1.0000	0.1935	0.0656	0.1630	0.0629	0.1676	0.1990	0.3497
B2	0.1146	0.4036	0.3196	0.0993	0.2243	0.1072	0.2443	0.3902	0.3382
B3	0.1333	0.4566	0.1835	0.0552	0.2280	0.0746	0.2776	0.2452	0.3188
C1	0.1896	0.2480	1.0000	0.1265	0.2670	0.1162	0.3050	0.4156	0.2406
C2	0.3448	0.2566	0.4939	0.2235	0.3345	0.1612	0.1655	0.4481	0.3496
C3	0.2776	0.1969	0.5107	0.1533	0.2840	0.1388	0.1607	0.3947	0.3872
D1	0.0517	0.0266	0.0712	1.0000	0.2217	0.2445	0.0646	0.1004	0.0581
D2	0.1167	0.0452	0.1409	0.4055	0.2390	0.2776	0.1767	0.1445	0.1086
D3	0.1514	0.0489	0.1134	0.4526	0.2095	0.2245	0.1842	0.1748	0.1137
E1	0.2203	0.1155	0.2388	0.2618	1.0000	0.3174	0.2576	0.3523	0.3098
E2	0.2594	0.1324	0.2635	0.2110	0.4148	0.2584	0.3410	0.3458	0.3261
E3	0.1578	0.0090	0.0719	0.2954	0.3990	0.3687	0.1617	0.1382	0.1476
F1	0.1640	0.0258	0.0690	0.2926	0.3641	1.0000	0.1808	0.2590	0.1598
F2	0.1992	0.0436	0.1626	0.3142	0.3370	0.4672	0.1841	0.2983	0.1611
F3	0.2164	0.0538	0.1212	0.3241	0.3419	0.4930	0.2035	0.2068	0.1717
G1	0.2873	0.1520	0.2391	0.2052	0.2756	0.3257	1.0000	0.3885	0.2443
G2	0.2076	0.1530	0.2828	0.1362	0.2751	0.2428	0.4373	0.2627	0.2747
G3	0.2948	0.0711	0.2724	0.1748	0.2751	0.3355	0.3526	0.3683	0.1953
H1	0.3536	0.1672	0.2877	0.2260	0.3298	0.1954	0.3102	1.0000	0.3334
H2	0.3215	0.2501	0.3304	0.1039	0.3062	0.1803	0.2444	0.4776	0.3884
H3	0.2901	0.1269	0.3797	0.1988	0.3094	0.1484	0.1630	0.3928	0.3633
I1	0.2474	0.2175	0.2796	0.0845	0.2082	0.2568	0.2500	0.3604	1.0000
I2	0.2632	0.2576	0.2693	0.0857	0.3140	0.1844	0.1683	0.3592	0.4559
I3	0.2609	0.1942	0.1826	0.1039	0.3391	0.1940	0.2082	0.3177	0.3776

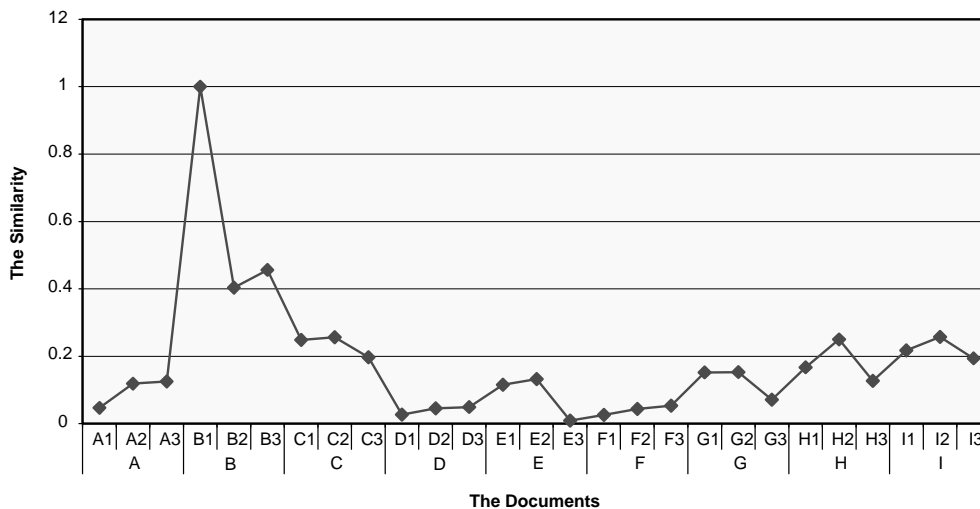


Fig. 6. The similarity with document B1.

between any two document images in advance off-line. Then the retrieval can be carried out according to the corresponding index information.

Commonly used retrieval method is query by example. How to measure the similarity between the query document

image and any other document image from the database acts a crucial role in the retrieval systems. In this paper, we have presented an approach to gauging the similarity between CCITT Group 4 compressed document images. The feature pixels are extracted from the changing ele-

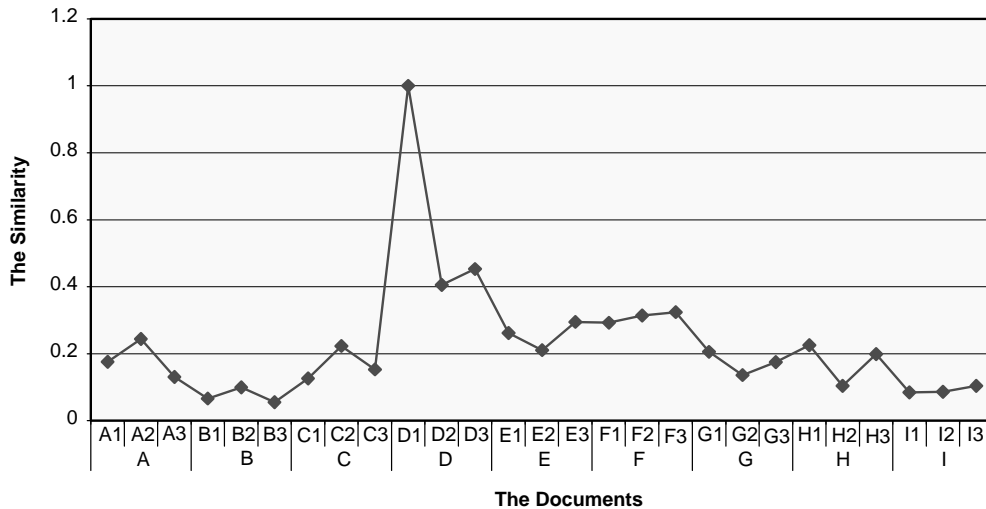


Fig. 7. The similarity with document D1.

Table 3
Retrieval performance for $\alpha = 0.32$

Group	A	B	C	D	E	F	G	H	I
Recall (%)	100.0	92.00	92.30	100.0	96.97	100.0	100.0	88.24	100.0
Precision (%)	75.36	75.15 C	72.48	68.34	70.83	78.62	65.23	77.52	69.98

Table 4
Retrieval performance for $\alpha = 0.36$

Group	A	B	C	D	E	F	G	H	I
Recall (%)	87.50	84.00	76.92	82.35	81.81	85.00	87.50	76.47	73.33
Precision (%)	87.50	80.77	76.92	83.33	81.25	93.75	78.57	90.91	83.33

Table 5
Retrieval performance for $\alpha = 0.40$

Group	A	B	C	D	E	F	G	H	I
Recall (%)	75.00	64.00	69.23	70.59	72.73	75.00	68.75	58.52	73.33
Precision (%)	92.30	100.0	100.0	85.71	92.30	100.0	91.67	100.0	100.0

ments of the compressed images. Word objects are bounded based on the changing elements directly. Weighted Hausdorff distance is proposed to match word objects. The experimental results have show that the method can measure the similarity between CCITT Group 4 documents effectively.

Document image retrieval is a challenging problem. In this research, only the occurrence frequency of word objects is employed as the vector feature of document, whereas the semantic information is not used. In further work, we want to

extract such information from document images to improve the retrieval performance.

Acknowledgements

This project is supported by the National Science and Technology Board and Ministry of Education of Singapore under research grant R255-000-071-112/303. The authors also would like to thank Mr. Ng Kok Koon and Mrs. Khoo

Yee Hoon of the Central Library of the National University of Singapore for providing us the document images.

References

- [1] G. Salton, Developments in automatic text retrieval, *Science* 253 (1991) 974–980.
- [2] G. Salton, J. Allan, C. Buckley, A. Singhal, Automatic analysis, theme generation, and summarization of machine-readable text, *Science* 264 (1994) 1421–1426.
- [3] Y. Yang, X. Liu, A re-examination of text categorization methods, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, California, USA, 1999, pp. 42–49.
- [4] T. Kameshiro, T. Hirano, Y. Okada, F. Yoda, A document image retrieval method tolerating recognition and segmentation errors of OCR using shape-feature and multiple candidates, Proceedings of the Fifth International Conference on Document Analysis and Recognition, Bangalore, India, 1999, pp. 681–684.
- [5] S. Senda, M. Minoh, K. Ikeda, Document image retrieval system using character candidates generated by character recognition process, Proceedings of the Second International Conference on Document Analysis and Recognition, Tsukuba, Japan, 1993, pp. 541–546.
- [6] D. Doermann, The indexing and retrieval of document images: a survey, *Comput. Vision Image Understanding* 70 (3) (1998) 287–298.
- [7] V.N. Gudivada, V.V. Raghavan, Content-based image retrieval systems, *Computer* (1995) 18–22.
- [8] Y. Wu, Y.T. Zhuang, Y.H. Pan, Image retrieval system for web: Webscope-CBIR, Proceedings of the 11th International Workshop on Database and Expert Systems Applications, London, UK, 2000, pp. 620–624.
- [9] A.H. Kam, T.T. Ng, N.G. Kingsbury, W.J. Fitzgerald, Context based image retrieval through object extraction and querying, Proceedings of the IEEE Workshop on Content-Based Access of Image and Video Libraries, South Carolina, USA, 2000, pp. 91–95.
- [10] C.Y. Fung, K.F. Loe, A new approach for image classification and retrieval, Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, California, USA, 1999, pp. 301–302.
- [11] A. Vailaya, A. Jain, H.J. Zhang, On image classification: city images vs. landscapes, *Pattern Recognition* 31 (12) (1998) 1921–1935.
- [12] F.R. Chen, D.S. Bloomberg, Extraction of thematically relevant text from images, Proceedings of the Symposium on Document Analysis and Information Retrieval, Las Vegas, NV, USA, 1996, pp. 163–178.
- [13] J. Liu, A.K. Jain, Image-based form document retrieval, *Pattern Recognition* 33 (3) (2000) 503–513.
- [14] D. Niyogi, S. Srihari, Use of document structure analysis to retrieve information from documents in digital libraries, in: L.M. Vincent, J.J. Hull (Eds.), Proceedings of the SPIE, Document Recognition IV, Vol. 3027, San Jose, CA, USA, 1997, pp. 207–218.
- [15] J.J. Hull, Document matching on CCITT Group 4 compressed images, in: L.M. Vincent, J.J. Hull (Eds.), Proceedings of the SPIE, Document Recognition IV, Vol. 3027, San Jose, CA, USA, 1997, pp. 82–87.
- [16] U.V. Marti, D. Wymann, H. Bunke, OCR on compressed images using pass modes hand hidden Markov models, Proceedings of the IAPR Workshop on Document Analysis Systems, Rio de Janeiro, Brazil, 2000, pp. 77–86.
- [17] W. Kou, Digital image compression algorithms and standards, Kluwer Academic Publishers, Dordrecht, 1995.
- [18] Y.Y. Tang, S.W. Lee, C.Y. Suen, Automatic document processing: a survey, *Pattern Recognition* 29 (12) (1996) 1931–1952.
- [19] D.P. Huttenlocher, G.A. Klanderman, W.J. Rucklidge, Comparing images using the Hausdorff distance, *IEEE Trans. Pattern Anal. Mach. Intell.* 15 (9) (1993) 850–863.
- [20] M.P. Dubuisson, A.K. Jain, A modified Hausdorff distance for object matching, Proceedings of the 12th International Conference on Pattern Recognition, Jerusalem, Israel, 1994, pp. 566–568.

About the Author—YUE LU is a research fellow at the Department of Computer Science, National University of Singapore. He received his B.E. and M.E. degree in Communication and Electronic Engineering from Zhejiang University, Zhejiang, China in 1990, 1993, respectively, and his Ph.D. degree in Pattern Recognition and Intelligence System from Shanghai Jiao Tong University, Shanghai, China in 2000. His research interests include document image processing, document retrieval, character recognition and intelligence system.

About the Author—CHEW LIM TAN is an associate professor in Computer Science at the School of Computing, National University of Singapore. His research interests are computer vision, document image analysis, intelligent text processing and neural networks. He obtained a B.Sc. (Hons.) degree in Physics in 1971 from the University of Singapore, an M.Sc. degree in Radiation Studies in 1973 from the University of Surrey in UK, and a Ph.D. degree in Computer Science in 1986 from the University of Virginia, USA.