

SIMILARITY MEASURE FOR CCITT GROUP 4 COMPRESSED DOCUMENT IMAGES

Yue Lu, Chew Lim Tan, Liying Fan and Weihua Huang

Dept. of Computer Science, National University of Singapore, Singapore 119260

Email: {luy, tancl, fnaly, huangwei}@comp.nus.edu.sg

ABSTRACT

Similarity measure of document images acts a crucial role in the area of document image retrieval. A method of measuring the similarity of CCITT Group 4 compressed document images is proposed in this paper. The features are extracted directly from the changing elements of the compressed images. Weighted Hausdorff distance is utilized to assign all of the word objects from two document images 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. Five groups of articles relating to different domains are used to test the validity of the presented approach.

1. INTRODUCTION

The emergence of the Internet and high speed networks over the past decade have made it possible to develop digital libraries. Digital libraries provide a broad collection of documents in either text format or image format. Undoubtedly, the text(electronic code) format facilitates not only the storage and transmission of documents in the Internet, but also document retrieval 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 form such as books and students' theses, and need to be transferred to the digital domain. As an approach to automatically transferring paper documents to their text format, 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.

A content-based document image retrieval system is required 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 formulated query by the degree of similarity with the query image, in which document similarity measure is an important component. Information retrieval from document images provides an even greater challenge than that from text documents. However, some content-based document image retrieval methods have been presented by researchers in the recent years^[1-2].

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 G3/4, JPEG, and JBIG2, etc). The considerable advantages will be realized if we can carry out document image similarity measure directly on the compressed images.

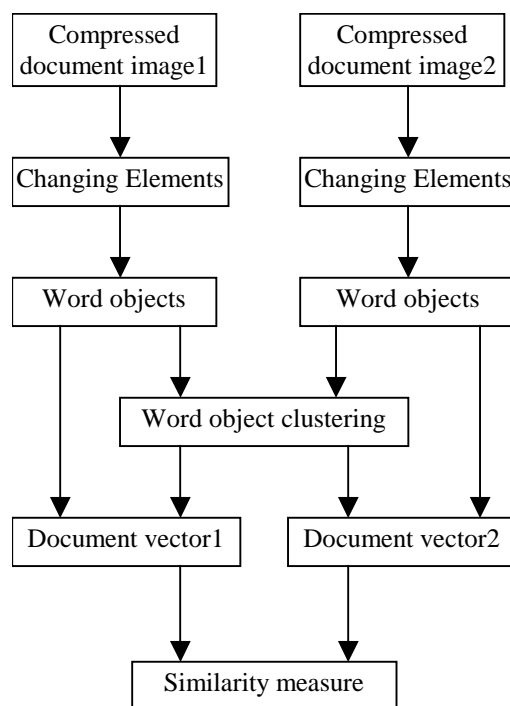


Fig. 1. Diagram of document image similarity measure

Interesting research such as duplicate document detection^[3-4] and OCR^[5] on the compressed images has been reported recently. In this paper, we will focus on the document similarity measure on the CCITT Group 4 compressed document images. The word objects are extracted directly from the changing elements of the compressed images. Weighted Hausdorff distance is utilized to assign all of the word objects from two document images to corresponding classes and the document vectors are built by the occurrence frequency of the word classes. The pair-wise similarity of two document images is represented by the scalar product of the document vectors. Figure 1 shows the diagram for measuring the similarity of two CCITT Group 4 compressed document images. Five groups of articles relating to different domains will be used to test the validity of the presented approach.

2. CHANGING ELEMENTS OF CCITT GROUP 4 COMPRESSED IMAGES

The CCITT Group 4 coding scheme for binary images uses a two-dimensional line-by-line coding method^[6], 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), V_R(1), V_R(2), V_R(3), V_L(1), V_L(2), V_L(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.

Except for the following coded positions of the pass mode, each coded position indicates that the current pixel color is different from its previous pixel. In our work, we focus on these changing elements in the CCITT Group 4 compressed document images, because they can be easily obtained from the compressed images. Figure 2 shows a part of an original document image and the changing elements extracted from its corresponding CCITT Group 4 compressed image, in which the changing element following a pass mode is removed because it is not an actual changing point according to the CCITT Group 4 standards. It can be seen from Figure 2 that the changing elements are similar to the characters' profiles, but not perfectly. We will measure the similarity between documents based on the changing elements extracted directly from the compressed images.

(a) Original image

(b) Changing elements extracted from compressed image
Fig. 2. Changing elements

3. DOCUMENT VECTOR AND SIMILARITY

3.1 Document feature Extraction

First, the word objects are bound in the “decompressed” image which is composed of the changing elements by the method of projection. Although the word objects extraction is affected by the unstable spacing between adjacent characters and punctuations sometimes, they can represent the document features generally.

Some common words called *stop words* in linguistics are eliminated 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 will be excluded from the document features. In our experiments, λ is set as 2.0.

An unsupervised classifier is employed to place each of the remaining word objects coming from both of the two document images into a set of classes, in which weighted Hausdorff distance is utilized to measure the distance between two word objects.

3.2 Weighted Hausdorff distance for word object matching

Hausdorff distance has been widely applied in two-dimensional image matching, especially in the area of object matching^[7-8].

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

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

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:

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

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

$$\begin{aligned} h(A,B) &= \max_{a \in A} d(a,B) \\ &= \max_{a \in A} \min_{b \in B} d(a,b) \\ &= \max_{a \in A} \min_{b \in B} \|a-b\| \end{aligned}$$

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 .

Dubousson and Jain^[8] presented the modified Hausdorff distance (MHD) measure by employing the summation operator over all distance, rather than the maximum operator:

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

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

A word object can be divided into different parts, namely the ascender, the descender, and the mid zone. We propose a 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 directed distance of WHD is computed as

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

where

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

The weight of the different parts in the word namely $w(a)$, $w(d)$ and $w(m)$ correspond to the ascender, the descender and the mid-zone respectively. In our experiments, they are set as:

$$w(a)=w(d)=2*w(m)$$

3.3 Document similarity measure

An unsupervised classifier clusters all of the word objects 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:

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

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

4. EXPERIMENTAL RESULTS

Experiments were conducted to verify the validity of the proposed approach of measuring the similarity between 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 included in the PDF files, and compressed by CCITT Group 4 standards. These articles address five different kinds of topic respectively. Six document images are arbitrarily picked from each topic. Group A(A1-A6) refers to transmission performance determination of

mobile radio communication system, group B(B1-B6) is on computation of phase and chemical equilibrium, group C(C1-C6) talks about vapor-liquid and liquid-liquid equilibrium calculations, group D(D1-D6) describes frequency optimization using genetic algorithms, and group E(E1-E6) presents adaptive equalization in teletext reception. For each group, we take the first one as the reference document(query document). Similarity of all the document images with the respective five reference document images is carried out. The results are tabulated in Table 1. From the results, we can see that the proposed approach is able to measure the similarity between the compressed document images effectively with an average of precision ranging from 81.08 % to 100% and an average recall ranging from 73.33% to 100% depending on the similarity threshold α , as showed in table 2.

5. CONCLUSIONS

We present an approach to gauging the similarity between CCITT Group 4 compressed document images. The features are extracted directly from the changing elements of the compressed images. 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.

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] D. Doermann, "The retrieval of document images: a brief survey," Proceedings of the Fourth International Conference on Document Analysis and Recognition, 1997. vol.2, pp: 945 –949
- [2] C. L. Tan, S. Y. Sung, Z. H. Yu and Y. Xu, "Text retrieval from document images based on n-gram algorithm," Text and Web Mining workshop, 6th Pacific Rim International Conference on Artificial Intelligence. 2000, Melbourne, Australia.
- [3] 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), 1997, vol.3027, pp.82-87
- [4] D. S. Lee and J. J. Hull, "Duplicate detection for symbolically compressed documents," Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999. pp: 305 –308

- [5] U. V. Marti, D. Wymann and H. Bunke, "OCR on compressed images using pass modes hand hidden Markov models," IAPR workshop on document Analysis Systems, 2000, pp.77-86
- [6] W. Kou, Digital image compression algorithms and standards, Kluwer Academic Publishers, 1995.
- [7] 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, pp850-863, 1993.
- [8] M. P. Dubuisson and A. K. Jain. "A modified Hausdorff distance for object matching", in Proc. 12th Int. Conf. Pattern Recognition, Jerusalem, Israel, Oct. 1994, pp.566-568

Table 1. Document similarity

	A1	B1	C1	D1	E1
A1	1.0000	0.1054	0.1876	0.0761	0.2235
A2	0.4948	0.0268	0.1524	0.0629	0.2078
A3	0.4976	0.0681	0.1895	0.0605	0.1559
A4	0.3401	0.0772	0.1592	0.0557	0.1568
A5	0.4979	0.0368	0.1770	0.0554	0.2287
A6	0.4821	0.0935	0.1786	0.0877	0.2265
B1	0.1245	1.0000	0.0250	0.2428	0.3299
B2	0.0851	0.3806	0.0318	0.2358	0.2400
B3	0.1969	0.2926	0.0282	0.2130	0.2948
B4	0.1960	0.3918	0.1133	0.1796	0.2207
B5	0.0759	0.3149	0.0338	0.1601	0.2255
B6	0.2180	0.2994	0.1955	0.1123	0.1035
C1	0.1522	0.0322	1.0000	0.0320	0.0626
C2	0.3193	0.0214	0.4255	0.0398	0.1198
C3	0.3083	0.0166	0.4846	0.0420	0.1071
C4	0.2405	0.0526	0.4058	0.0379	0.2000
C5	0.1958	0.0583	0.3697	0.0308	0.0866
C6	0.2526	0.0405	0.4030	0.0292	0.1027
D1	0.0834	0.2309	0.0320	1.0000	0.1847
D2	0.1605	0.2178	0.0376	0.4071	0.2915
D3	0.1317	0.2196	0.0294	0.3787	0.2180
D4	0.0234	0.1526	0.0279	0.2744	0.2096
D5	0.0485	0.1900	0.0144	0.3200	0.2259
D6	0.0735	0.1302	0.0034	0.3752	0.2311
E1	0.1920	0.2944	0.0654	0.2281	1.0000
E2	0.1822	0.2148	0.1043	0.1475	0.3901
E3	0.1378	0.1924	0.0557	0.1948	0.2905
E4	0.1792	0.2700	0.0383	0.2247	0.4352
E5	0.2830	0.2479	0.1679	0.1691	0.2906
E6	0.2037	0.1457	0.1833	0.0994	0.2879

Table 2. Recall and precision

Histogram α	0.27	0.30	0.33
Recall	100.0%	75.76%	73.33%
Precision	81.08%	98.29%	100.0%