Document Matching on CCITT Group 4 Compressed Images

Jonathan J. Hull
Ricoh California Research Center
2882 Sand Hill Road, Suite 115
Menlo Park, CA 94025
hull@crc.ricoh.com

ABSTRACT

A method is proposed for detecting whether two CCITT group 4 images were scanned from the same document. Features are extracted from rectangular patches of text and compared with a modified Hausdorff distance measure. Two images are said to be ''equivalent'' (i.e., they were scanned from the same document) if the Hausdorff measure finds that a specified number of features are located within a given distance of one another in both images. This paper explains the technique and presents experimental results that demonstrate its effectiveness. It is shown that features extracted from a one-inch square patch of image data provide better than 95% correct retrieval accuracy with no false positives on a database of 800 documents.

1. INTRODUCTION

A useful function in a document image database system is the detection of whether a given image already exists in the database. This would enable a content-based retrieval capability in which a single page could be used to retrieve other related pages. A related application, illustrated in Figure 1, would issue a message when a specific document image was transmitted or received by a fax machine. This could be applied to a database of confidential documents. The transmission or reception of such images could be recorded.

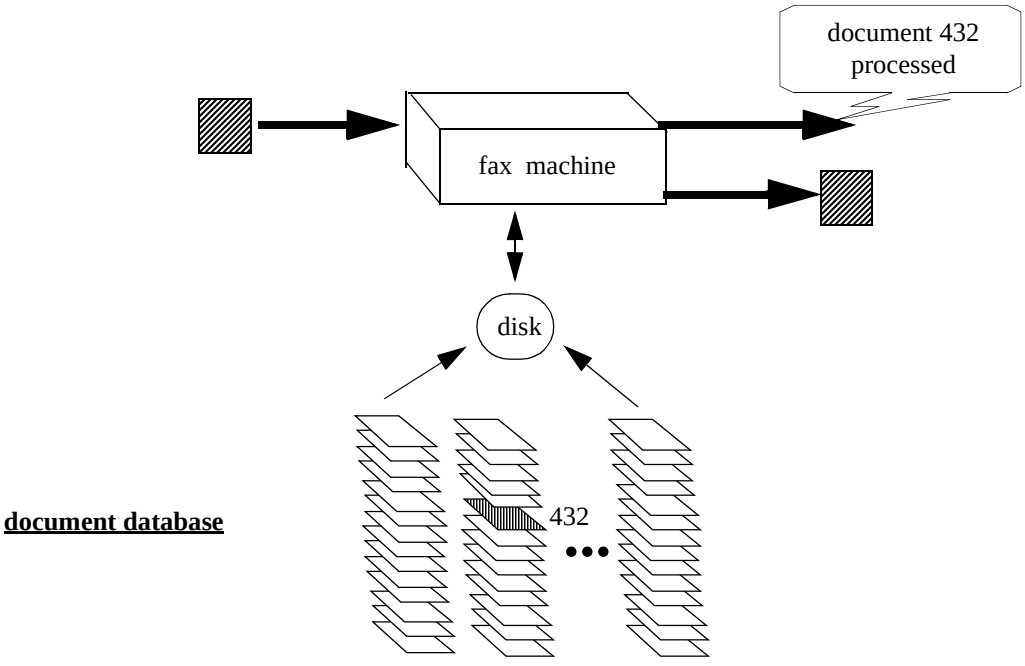


Figure 1. Fax alerting application for content-based document image matching.

appeared in SPIE Conference on Document Recognition IV, San Jose, CA, Feb. 8-14, 1997.

An earlier solution for document image matching counted the number of characters in words of text[1]. Descriptors were constructed from short sequences of word lengths and those descriptors were used as keys for a hash table. Duplicate documents were detected by applying a hash function to descriptors extracted from a query image. Database images that had a large percentage of descriptors in common with a query image were assumed to be from the same document.

Another solution for document image matching extracted features from the outline shape of words[5]. Characters were first converted to abstract shape tokens. For example, the characters *b,d,f,h,k,* and *l* were converted to the code A for *ascender.* Similar transformations were applied to descenders, x-height characters, and so on. The sequences of character shape codes derived from two document images were compared to determine the degree to which they matched.

This paper proposes an alternative technique for detecting duplicate document images. Features are extracted from a rectangular patch of image data in a document. The two-dimensional arrangement the features are compared to detect whether two documents are equivalent (that is, the images were scanned from the same document). A version of this approach is reported in which the features are extracted directly from CCITT group 3 or group 4 fax-compressed representations. This eliminates some of the image processing needed for the previous methods discussed above.

The rest of this paper presents the proposed algorithm in detail. An experimental application of the technique to a database of document images is described and the results are analyzed. Areas for future investigation are also outlined.

2. PROPOSED ALGORITHM

The proposed algorithm for document image matching is outlined in Figure 2. The input image (also referred to as the query) and the images in the database are assumed to be compressed in CCITT group 3 or group 4 format. The x,y locations of the centers of pass coded runs in each image are extracted. A subset of pass code locations in each image are chosen that are contained in rectangular patches of text. The two-dimensional arrangements of x,y locations are compared using a modified Hausdorff distance measure that compensates for x-y translation[2]. It is assumed that skew would be normalized by prepro-

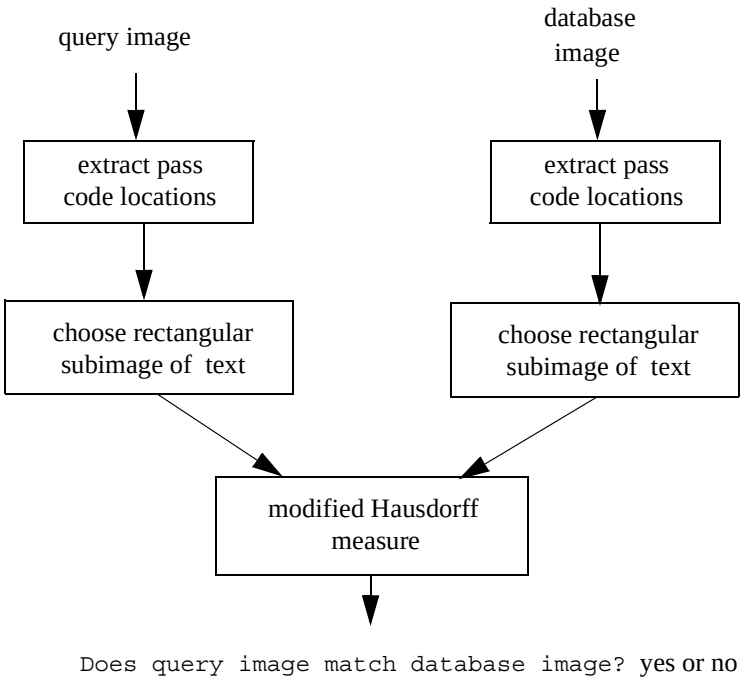


Figure 2. Document image matching algorithm.

cessing. It is further assumed that it is not necessary to compensate for scale change. A binary decision is output that indicates whether the query image is equivalent to a given image from the database. This procedure is used to compare a query image sequentially to each image in the database.

Pass codes are used in the CCITT compression technique to encode black or white runs on a given row that are not connected to a black or white run (of the same color) on an adjacent row. In many cases this attaches a pass-coded run to the bottom of black components or white holes. The locations of pass codes in document images have been used to estimate the angle at which a document was skewed[4]. Results showed that this approach was accurate to within a fraction of a degree.

Figure 3 shows a one inch square patch of image data and the pass codes extracted from it. It can be seen that the two-dimensional arrangement of pass codes represents the rigid arrangement of characters in a passage of text. Aspects of the identities of the characters, their font size, line spacing, and word spacing are reflected in the layout of pass codes.

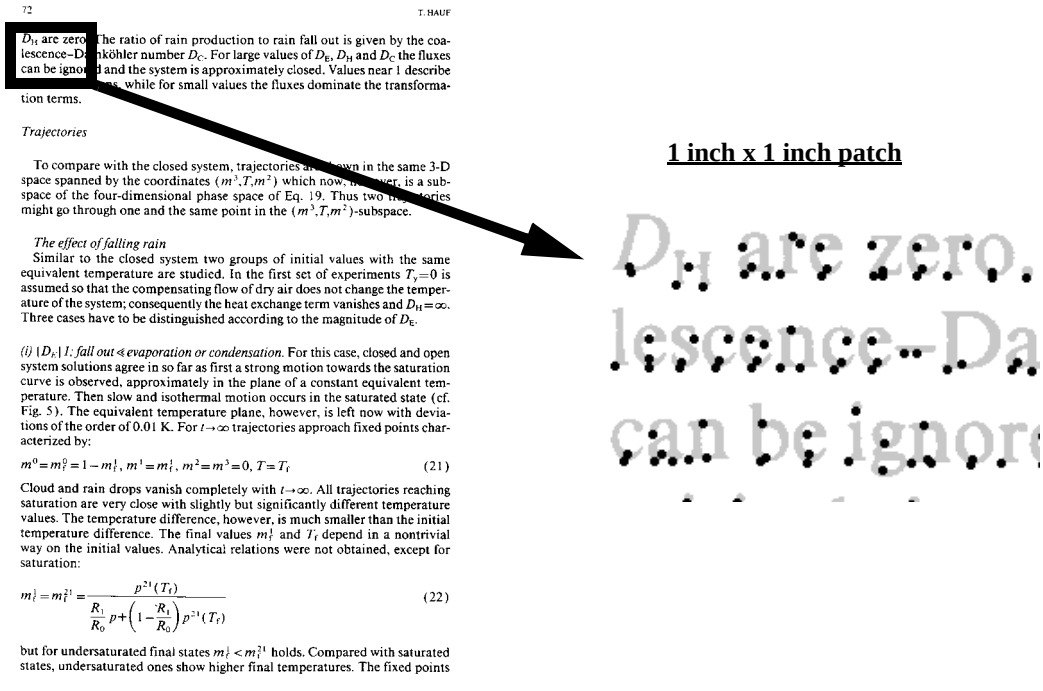


Figure 3. Pass code locations extracted from a document image.

3. EXPERIMENTAL RESULTS

An experiment was conducted that tested the ability of the proposed algorithm to detect duplicate document images. A one-inch square patch of image data was extracted from the upper left corner of one zone classified as ''body text'' in each of 800 images on UW CDROM 1[3]. The other 179 images on the CDROM did not contain a zone classified as body text that was at least one-inch square. Each image had been scanned at 300 dpi. The coordinates and zone classifications were determined from truth files provided on the CDROM.

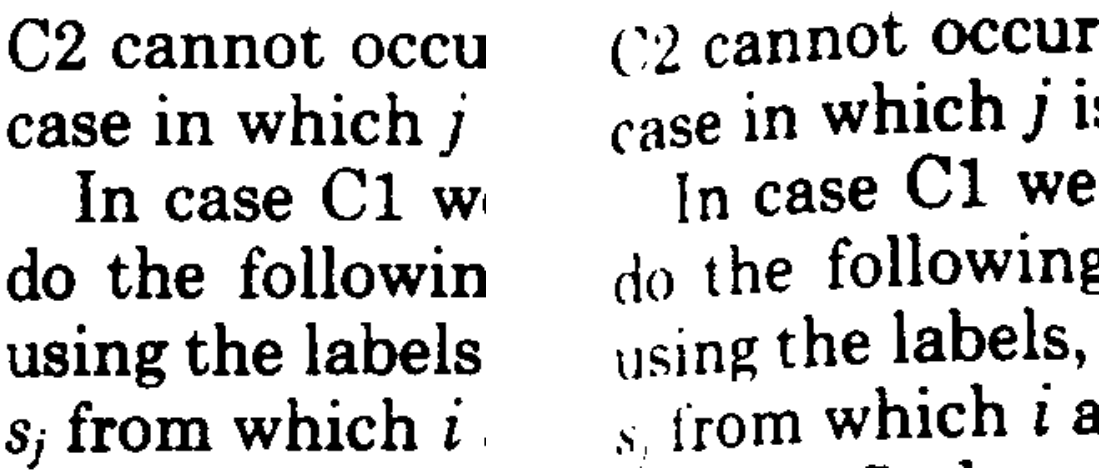
On average each patch contained 190 pass-coded runs. Thus, only a small percentage of the 90,000 possible locations in the patch are filled.

Out of the 800 images, there were 133 pairs (266 total images) that contained an image scanned from an original document and an image scanned from a photocopy of the original. The generation of the photocopies ranged from first to third.

Parameters for the modified Hausdorff comparison were chosen after brief experimentation with about 10 images. In order for two images to be ''equivalent,'' at least 70% of the pass codes in a query image had to be within 4 pixels of a pass code in a database image. Also, at least 50% of the pass codes in that database image had to be within 4 pixels of a pass code in the query image. About 8 minutes of CPU time on a Sun Sparcstation 20 were needed to compare each patch to the complete set of 800 patches.

The results showed that the proposed algorithm was 100% accurate in matching the 800 document images to themselves. There were no false positives. 95% of the 133 pairs of duplicate images were also correctly located with no false positives. The other 5% (12 images) failed for a variety of reasons.

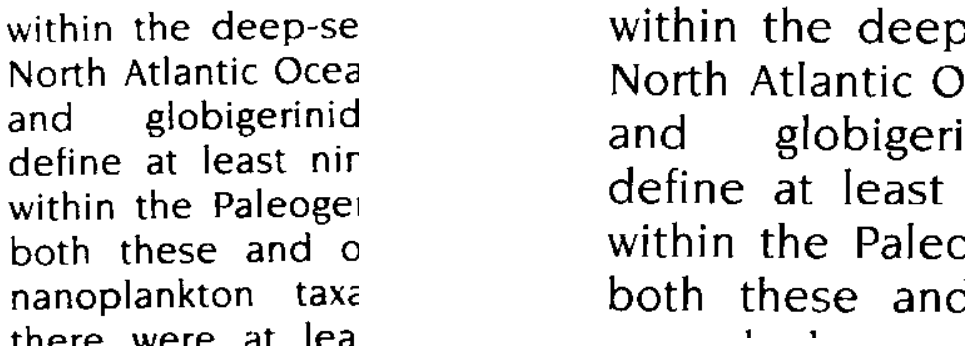
An analysis of the errors showed that four of them were caused by non-linear distortions such as those shown in Figure 4 which can occur when copying pages near the binding of a book. Four of the errors were caused by scale differences between the original and the photocopy (see Figure 5 for an example). The other four errors were caused for miscellaneous reasons that might be corrected by further adjustment of the parameters.



IG0D: 1st generation photocopy

I00D: 2nd generation photocopy

Figure 4. Error caused by non-linear distortion



S02H: original scanned (93%)

E02H: 1st generation photocopy 100%

Figure 5. Error caused by scale difference between original and photocopy.

4. CONCLUSIONS AND FUTURE DIRECTIONS

A technique for determining whether two images were scanned from the same document was proposed. The x-y coordinates of the centers of pass-coded runs were extracted directly from CCITT group 3 and group 4 document images and compared with a modified Hausdorff distance measure.

Experimental results showed that the pass codes extracted from a one-inch square patch of text provided a unique signature for a document. Furthermore, this signature was robust to the noise caused by photocopying a document before scanning it. Also, in the experiments discussed here, on average 190 x-y pairs (only 380 bytes) were needed to encode each square patch.

Limitations of the present experimwnt include the use of fixed coordinates from truth files. These were determined manually when the CDROM containing the image database was created. An automatic technique for locating text zones in document images should be used instead.

Further improvements in run time could be obtained by using a hierarchical approach. Global features of a document image could be used to reduce the number of potential matches that are compared with the Hausdorff distance measure. Such an approach could also be useful in comparing documents that contain little or no text,

REFERENCES

1. J.J. Hull, ‘‘Document image matching and retrieval with multiple distortion-invariant descriptors,’’ in *Document Analysis Systems*, World Scientific, 1995, 379-396.

2. D. Huttenlocher, G. Klanderman, and W. Rucklidge, ‘‘Comparing images using the Hausdorff distance,’’ *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 15, no. 9, September, 1993, 850-863.

3. I. T. Phillips, S. Chen, R. M. Haralick, ‘‘CD-ROM document database standard,’’ Proceedings of the Second International Conference on Document Analysis and Recognition, October 20-22, 1993, Tsukuba Science City, Japan, 478-483.

4. A.L. Spitz, ‘‘Skew determination in CCITT group 4 compressed document images,’’ Proceedings of the Symposium on Document Analysis and Information Retrieval, Las Vegas, Nevada, March 16-18, 1992, 11-25.

5. A.L. Spitz and A.P. Dias, ‘‘Method for matching text images and documents using character shape codes,’’ U.S. Patent 5,438,628, August 1, 1995.