Document Image Similarity and Equivalence Detection

Jonathan J. Hull and John F. Cullen
Ricoh California Research Center
2882 Sand Hill Rd., Suite 115
Menlo Park, CA 94025
{hull,cullen}@crc.ricoh.com

Abstract

A hierarchical algorithm is presented for determining the similarity and equivalence of document images. Features extracted from the CCITT fax-compressed representations of two images are compared to determine their visual similarity and whether they are equivalent. Pass codes in the compressed data are used as features. A fixed grid is imposed on the image and a feature vector is derived from the number of pass codes in each grid cell. The features vectors are compared to locate a group of documents that are visually similar to the input image. The equivalence of two documents is determined by applying the Hausdorff distance to the two-dimensional arrangement of pass codes in small patches of each image.

1. Introduction

Matching one document image to another is an important part of a document analysis system. This process determines whether a visually similar document is contained in an image database. This is useful for an automatic filing application in which scanned document images are stored in directories that contain visually similar documents. For example, business letters might be stored separately from scientific papers. Another application determines whether a specific document image is equivalent to another image in a large database.

There are several methods that could be used for matching document images. These include matching features extracted from text, such as the sequence of word lengths [2] or visual features [1]. These approaches share two steps: *feature extraction* and *matching* of the extracted features to document images in a database.

Important characteristics of the feature extraction algorithm include its *speed* and the *uniqueness* of the representation. The speed should be practical for an economical commercial implementation. The uniqueness of the representation should ensure that a given document matches itself with a high probability and matches no other document. Important characteristics of the matching algorithm include its speed and the memory required for the stored feature representation.

This paper proposes a combination of feature extraction from CCITT fax-compressed images with a matching algorithm that is efficient in its use of time and memory. The x-y positions of pass codes in the original image are used as features. These are extracted with a single pass over the compressed representation and have been used successfully to determine the skew of a document image [6].

The usefulness of this representation in finding visually similar documents is demonstrated experimentally. A fixed grid is imposed on the image and the number of pass codes in each grid cell is computed. The similarity between documents is calculated with the Euclidean distance.

The utility of the two-dimensional arrangement of pass codes in determining whether two document images are equivalent (i.e., they were scanned from the same original document) is also demonstrated experimentally. A one inch square patch is extracted from each document and the x-y locations of the centers of pass-coded runs are compared with the Hausdorff distance.

The rest of this paper contains a short explanation of the CCITT format and the proposed algorithm. An experimental investigation is reported that shows the new method is effective in finding visually similar as well as equivalent images of the same document.

2. CCITT Fax Encoding

The CCITT group 3 encoding of a binary image uses a combination of a one-dimensional coding scheme applied to every kth line (typically, k equals 4) with a two-dimensional coding method. The 2-d technique codes black and white runs in a given row with respect to corresponding runs in the previous row. In the group 4 standard, all lines are coded two-dimensionally. The first line in an image is coded with respect to an all-white reference line.

The two-dimensional coding algorithm uses three coding modes: horizontal, vertical, and pass. A pass code is used when a run on one line has no corresponding run in an adjacent line. This indicates the termination of

0-8186-7898-4/97 $10.00 © 1997 IEEE

4th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), Ulm, Germany, Aug. 18-20, 1997, pp. 308-311.

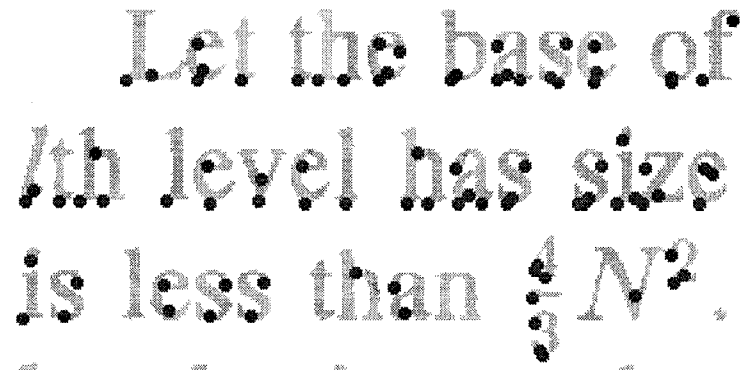
white or black components. For example, many characters will have white pass codes attached to holes and black pass codes attached to the bottoms of strokes. Approximately 80% to 90% of the components have an attached pass code. This characteristic is illustrated in the portion of the document image shown in Figure 1. The pass codes extracted from a scanned original are shown in Figure 1(a) and the pass codes extracted from a scanned photocopy of the same document are shown in Figure 1 (b).

The x-y locations of pass-coded runs can be extracted directly from the compressed data. The speed-up compared with extracting similar information from an uncompressed image is proportional to the compression factor. In practice an even greater speed-up should be achieved since the fax compression algorithm performs additional comparisons that must be done when an uncompressed image is processed.

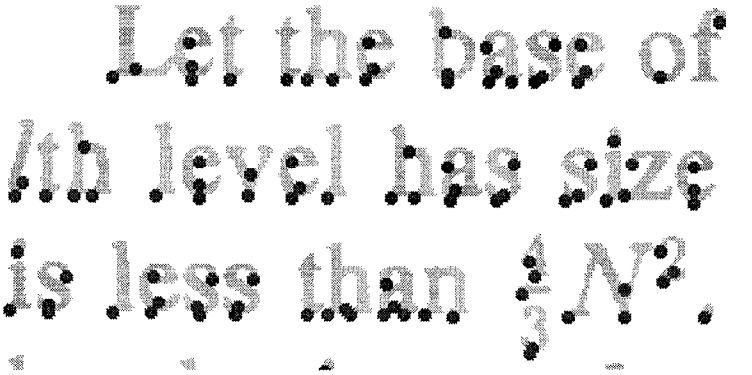
3. Proposed Algorithm

The proposed algorithm for document image matching is shown in Figure 2. First, the x-y locations of pass codes are extracted from an input fax-encoded image. Then the number of pass codes in each cell of a fixed grid are determined. A group of visually similar documents is located by calculating the Euclidean distance between these feature vectors. A fixed number of images with the smallest distances provide a group of documents that are likely to contain an equivalent image, if one exists in the database.

Equivalent documents are detected by locating a patch (e.g., the upper left 1-inch by 1-inch region of the first paragraph) in each image that could be the same in both documents. The x and y locations of the centers of pass-coded runs in each patch are compared using the Hausdorff distance. Two patches with a Hausdorff distance below a



(a)



(b)

Figure 1. Pass codes in a CCITT group 4 fax coded image.Scanned original (a) and scanned photocopy (b) of the same

4th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), Ulm, Germany, Aug. 18-20, 1997, pp. 308-311.

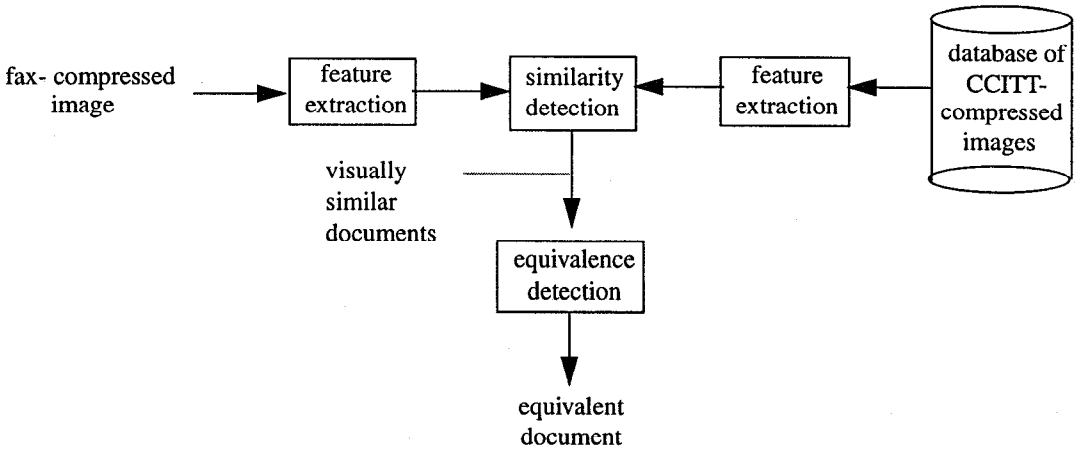


Figure 2. Document image similarity and equivalence detection algorithm.

threshold are assumed to have been derived from the same image. This technique has been used successfully without a similarity detection algorithm [3].

A modified Hausdorff distance has been proposed for locating models of objects in two-dimensional scenes [4]. This modified distance measures the percentage p_m of model points that are within d_m pixels of some image point for some given translation of the model. It also measures the percentage p_i of image points that are within d_i pixels of some model point, for the same translation. An efficient implementation of the modified Hausdorff distance has also been developed [5].

Figure 1 helps explain why the Hausdorff distance applied to pass code locations is effective in locating equivalent document images. Even though a number of pass codes occur in one image and not the other, if they do occur in both images, they are very likely to be close to one another, for an x-y translation that registers the documents. This is because of the natural constraint provided by the content of the text passage (i.e., the sequence of characters it contains), as well as the font, point size, pitch, and spacing in a typeset passage.

4. Experimental Results

A series of experiments were performed that investigated the performance of the similarity and equivalence detection steps. The 979 images on the first University of Washington CDROM were used as test data. Each image on the CDROM is labeled with a four character identifier. 125 images are labeled with an initial "E" and 125 images are labeled with an initial "S". Each image labeled E was scanned from the same original as an image labeled with an S. There are also 21 images labeled with an initial "IG" that were scanned from the same original as an image labeled with an initial "I00". The difference between the members of such a pair is that they were usually scanned from different generations of photocopies. The "E" documents were all scanned from first generation copies, 115 of the "S" documents were scanned directly from the original document (0th generation), the other ten were scanned from first generation copies. All the "IG" documents were scanned from first generation copies and all the "I00" documents were scanned from second generation copies.

4.1 Similarity detection

The objective of testing the similarity detection algorithm was to determine whether it finds the corresponding member of a pair among its ten best choices when it is given a document labelled E, S, IG, or I00 as input. Also, the vector size needed to obtain the best performance was investigated.

The performance of the proposed technique was calculated using grid sizes 2x2, 3x3, 4x4 and 5x5. This provided feature vectors that have 4, 9, 16, and 25 integers, respectively. The feature vector from each of the 979 images on the CDROM was compared to the feature vectors for all the other 978 images to obtain the results summarized in Table 1. The absolute number of correct choices in each of the first ten positions is shown as well as the overall percent correct in the top ten.

4th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), Ulm, Germany, Aug. 18-20, 1997, pp. 308-311.

The results show that the technique was successful in finding the correct document among the first ten choices in up to 98% of all cases. This varied between 94% and 98% for the different vector lengths. Of the 7 images that could not be matched when a 5x5 grid was used, 4 of them (E02H, E02I, S02H, and S02I) were caused by documents on the CDROM that were scanned at different magnifications. These errors are understandable since this technique was not designed to handle scale changes. Two of the errors (IG0G and I00G) apparently occured because one photocopy was made with the cover closed and the other without the cover closed.

4.3 Equivalence detection

Another set of experiments explored the ability of the proposed algorithm to correctly match the image of a given document to a different image of the same document in a large database. A one-inch by one-inch patch was extracted from the upper-left corner of the first body text zone of 800 document images on the CDROM. The upper-left corners were located from the truth files. The 800 images were those that contain at least one zone classified as "text-body" that is at least 1-inch by 1-inch. The other 179 images either contained no body text or the largest body text zone was smaller than one inch square.

This was designed to be a stress test for the equivalence testing algorithm. That is, the previous results showed that the equivalence detection algorithm may only have to be executed on as few as 10 image hypotheses chosen from an initial set of 979 documents. Successful equivalence detection on 800 images implies that the combined approach of similarity detection followed by equivalence detection could also work well on a larger initial set of documents.

The 800 images contain 266 images that have a duplicate elsewhere in the database that had been scanned from a different generation photocopy. This includes 114 images labeled E, 114 images labeled S, 19 images labeled I, and 19 images labeled IG. For any image, there is at most one other equivalent version in the database.

The two-dimensional arrangement of pass codes in each of the 800 test patches was matched to the two-dimensional arrangement of pass codes in all the 800 patches in the database. The parameters used were $p_m = 70$, $d_m = 4.0$, $p_i = 50$, and $d_i = 4.0$.

The results of this experiment are shown in Table 2. This shows that every patch was uniquely matched to itself (100% correct, 0% error). Also, 95% of the patches that occurred twice (original and photocopy) in the database were also correctly located. That is, there was a 5% miss rate. This indicates that it is possible to derive an identifier for a complete page of English text from a relatively small amount of image data (one square inch) extracted from it. Also, this representation is tolerant to the noise that typically occurs when a document is photocopied.

The run time of the matching algorithm was measured to be about 10 minutes for each patch on the given database of 800 documents. This was calculated on a 70 Mhz Sparcstation 20.

5. Conclusions and Future Directions

A hierarchical algorithm for document image matching was proposed that uses the two-dimensional arrangement of pass codes in a fax-compressed document image. A group of similar images is first located with a feature vector that counts the number of pass codes in each cell of a fixed grid. Equivalent images are then located by applying the modified Hausdorff distance to the documents returned by the first step.

data set	vect. size	choice										cum. %
		1	2	3	4	5	6	7	8	9	10	
979	2x2	215	24	12	7	4	2	6	1	3	1	94%
979	3x3	256	13	5	3	2	0	1	2	0	1	97%
979	4x4	267	7	4	1	4	0	0	2	0	0	98%
979	5x5	271	6	4	1	1	0	1	1	0	0	98%

Table 1. Performance in finding duplicate documents

4th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), Ulm, Germany, Aug. 18-20, 1997, pp. 308-311.

Areas for improving the similarity detection algorithm include incorporation of structural information about the segmentation of a document into zones such as text, graphics, photographs, etc. Such information may improve performance. However, the run time overhead of such a method should be considered. The experiments with the equivalence detection method assume corresponding image patches can be located in two equivalent images. An automatic technique for calculating this from compressed data should be developed.

Future work on this approach should include application to larger databases of document images. Also, a method for locating text zones in a group-4 encoded image should be developed.

References

1. D. Doermann, C. Shin, A. Rosenfeld, H. Kauniskangas, J. Sauvola and M. Pietikainen, "The development of a general framework for intelligent document image retrieval," Proceedings of the IAPR Workshop on Document Analysis Systems, Malvern, PA, October 14-16, 1996, 605-632.

2. J. J. Hull, "Document image matching and retrieval with multiple distortion-invariant descriptors," International Association for Pattern Recognition Workshop on Document Analysis Systems (DAS94), Kaiserlautern, Germany, October 18-20, 1994, 383-399.

3. J.J. Hull, "Document matching on CCITT group 4 compressed images," SPIE Conference on Document Recognition IV, San Jose, CA, February 12-13, 1997, 82-87.

4. D. Huttenlocher, G. Klanderman and W. Rucklidge, "Comparing images using the Hausdorff distance," IEEE Transactions on Pattern Analysis and Machine Intelligence 15, 9 (September, 1993), 850-863.

5. W. Rucklidge, "Efficient computation of the minimum Hausdorff distance for visual recognition," TR94-1454, Cornell University, Department of Computer Science, September 8, 1994.

6. 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.

test condition	N	% corr.	% error (false positives)	% missed
original: original	800	100%	0%	0%
original: copy	266	95%	0%	5% (12 images)

Table 2. Equivalence detection performance.

4th IAPR Int. Conf. on Document Analysis and Recognition (ICDAR), Ulm, Germany, Aug. 18-20, 1997, pp. 308-311.