Document Image Matching Techniques

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

Abstract

Several techniques for document image matching developed at the Ricoh California Research Center are presented. These methods are given a document image as input and locate visually similar or identical copies of the same image in a large database.

1. Introduction

Document image matching algorithms are useful in applications where the objective is to locate visually similar or identical copies of a given document in a large database. Applications of this technology include *automatic filing* in which a user would like to store documents with a similar appearance (e.g., business letters, utility bills, etc.) in the same location. Content-based *retrieval* is another application in which a single sheet from a multi-page original is used to locate the other pages.

Figure 1 illustrates a confidential document monitoring system in which the objective is to determine whether a given document image exists in a database. Such an approach could be used to monitor facsimile traffic. The transmission or reception of specific images could be recorded or alerts issued when certain documents were processed.

The rest of this paper describes several techniques for document image matching. A general framework is presented first. This is followed by a brief presentation of two algorithms: one that uses symbolic features extracted from text and another that uses features extracted directly from CCITT group 3 or group 4 fax compressed images.

2. General Framework

A general framework for document image matching is presented in Figure 2. This follows the general technique of hypothesis generation and testing commonly used to solve computer vision problems. Features are extracted from an input document image as well as the images in a database. Those feature descriptions are compared by a similarity detection algorithm that locates a group of N documents that are visually similar to a given image.

Equivalence detection is performed by extracting another (perhaps different) feature description from both the input document image and the N visually similar documents output by the similarity detection step. The output of equivalence detection are all the duplicates of the input document that are contained in the database.

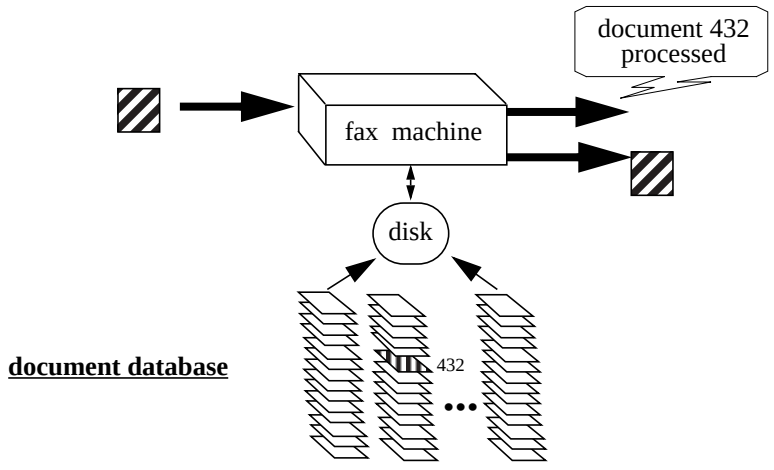


Figure 1. Fax alerting application for content-based document image matching (excerpted from [spie97]).

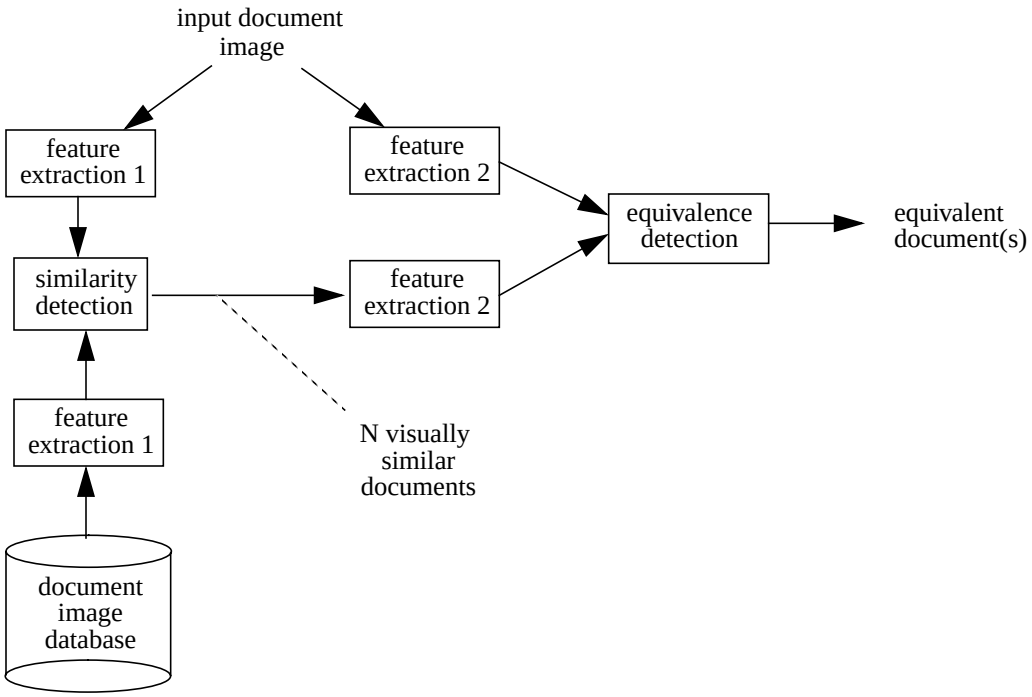


Figure 2. General framework for document image matching.

2.1 Similarity Detection

A technique for similarity detection is reported in [3] that calculates a feature vector by imposing a fixed grid on a document image and counting the number of connected components of a certain size that occur in each grid cell. Only connected components that were approximately the size of characters were counted.

Experimental results showed that as few as 9 features (a 3x3 grid) could be used to locate duplicate documents in a database of 979 images with a 98% accuracy. The test set for this application was extracted from University of Washington CDROM1 [6] and the object of the test was to locate duplicates of the 125 images tagged as 'E' that are indicated by a corresponding 'S' tag. The algorithm was applied to each of the 979 images in turn. The ten images with the minimum Euclidean distance to the input document were output. The result quoted above means that 98% of the time the correct match was contained among the ten documents with the minimum Euclidean distance to the input.

2.2 Equivalence Detection Using Word Lengths

A method for detecting equivalent document images in a large database is reported in [1]. This technique estimated the number of characters in each word of text and formed features by concatenating the lengths of M adjacent words. For example, with M=3, the phrase "the rain in Spain" can be described by the descriptors 3-4-2 and 4-2-5. This feature description is tolerant to noise in that incorrectly estimating the length of any word changes at most M descriptors. Also, this feature description maps easily from images of documents to their symbolic representations as ASCII files, independent of how they are formatted.

Each such "descriptor" was used as a key for a hash function and the identity of the passage of text from which each descriptor was extracted was stored in a hash table. At run-time, the descriptors extracted from a text passage were hashed and votes were accumulated in the hash table. Documents in the database that obtained a suitable number of votes were assumed to match the input image.

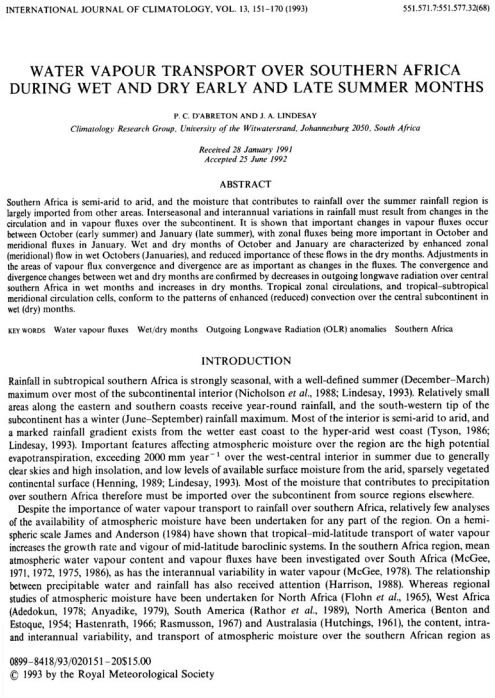
A result of this work is the observation that the sequence of word lengths extracted from a passage of text can provide a unique identifier for the passage. Experimental results showed that as few as 50 descriptors of length M=6 can be used to locate a matching document in a database of 997 images. Increasing the value of M reduces the number of descriptors and increases the number of documents that can be described by this method.

An adaptation of the word length hashing method was used in a paper-based technique for document image retrieval [5]. Small iconic representations for document images were printed in such a way that the visual appearance of the document was retained. However, an address of the original high resolution scanned image of the document in a large database could still be derived directly from the icon. This was done by printing text in such a way that the numbers of characters in each word could be determined from a scanned image of the icon.

An example of an original document and the icon derived from it are shown in Figure 3. It can be seen that the general appearance of the document is retained in the icon. This figure also shows the actual size of an icon. Experimental results showed that 49 of these icons could be printed on a single sheet of paper.

2.3 Equivalence Detection Using Pass Codes in Fax Images

Another technique for equivalence detection (outlined in Figure 4) addresses the application scenario presented in Figure 1 [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 [4]. It is assumed that skew would be normalized by preprocessing using a technique similar to that proposed in [7]. 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.



INTERNATIONAL JOURNAL OF CLIMATOLOGY, VOL. 13, 151–170 (1993) 551.571.7:551.577.32(68)

WATER VAPOUR TRANSPORT OVER SOUTHERN AFRICA
DURING WET AND DRY EARLY AND LATE SUMMER MONTHS

P. C. D'ABRETON AND J. A. LINDESAY

Climatology Research Group, University of the Witwatersrand, Johannesburg 2050, South Africa

Received 28 January 1991
Accepted 25 June 1992

ABSTRACT

Southern Africa is semi-arid to arid, and the moisture that contributes to rainfall over the summer rainfall region is largely imported from other areas. Interseasonal and interannual variations in rainfall must result from changes in the circulation and in vapour fluxes over the subcontinent. It is shown that important changes in vapour fluxes occur between October (early summer) and January (late summer), with zonal fluxes being more important in October and meridional fluxes in January. Wet and dry months of October and January are characterized by enhanced zonal (meridional) flow in wet Octobers (Januaries), and reduced importance of these flows in the dry months. Adjustments in the areas of vapour flux convergence and divergence are as important as changes in the fluxes. The convergence and divergence changes between wet and dry months are confirmed by decreases in outgoing longwave radiation over central southern Africa in wet months and increases in dry months. Tropical zonal circulations, and tropical–subtropical meridional circulation cells, conform to the patterns of enhanced (reduced) convection over the central subcontinent in wet (dry) months.

KEY WORDS Water vapour fluxes Wet/dry months Outgoing Longwave Radiation (OLR) anomalies Southern Africa

INTRODUCTION

Rainfall in subtropical southern Africa is strongly seasonal, with a well-defined summer (December–March) maximum over most of the subcontinental interior (Nicholson *et al.*, 1988; Lindesay, 1993). Relatively small areas along the eastern and southern coasts receive year-round rainfall, and the south-western tip of the subcontinent has a winter (June–September) rainfall maximum. Most of the interior is semi-arid to arid, and a marked rainfall gradient exists from the wetter east coast to the hyper-arid west coast (Tyson, 1986; Lindesay, 1993). Important features affecting atmospheric moisture over the region are the high potential evapotranspiration, exceeding 2000 mm $year^{-1}$ over the west-central interior in summer due to generally clear skies and high insolation, and low levels of available surface moisture from the arid, sparsely vegetated continental surface (Henning, 1989; Lindesay, 1993). Most of the moisture that contributes to precipitation over southern Africa therefore must be imported over the subcontinent from source regions elsewhere.

Despite the importance of water vapour transport to rainfall over southern Africa, relatively few analyses of the availability of atmospheric moisture have been undertaken for any part of the region. On a hemispheric scale James and Anderson (1984) have shown that tropical–mid-latitude transport of water vapour increases the growth rate and vigour of mid-latitude baroclinic systems. In the southern Africa region, mean atmospheric water vapour content and vapour fluxes have been investigated over South Africa (McGee, 1971, 1972, 1975, 1986), as has the interannual variability in water vapour (McGee, 1978). The relationship between precipitable water and rainfall has also received attention (Harrison, 1988). Whereas regional studies of atmospheric moisture have been undertaken for North Africa (Flohn *et al.*, 1965), West Africa (Adedokun, 1978; Anyadike, 1979), South America (Rathor *et al.*, 1989), North America (Benton and Estoque, 1954; Hastenrath, 1966; Rasmusson, 1967) and Australasia (Hutchings, 1961), the content, intra- and interannual variability, and transport of atmospheric moisture over the southern African region as

0899–8418/93/020151–20$15.00
© 1993 by the Royal Meteorological Society

(a)

(b)

Figure 3. Original image (a) and the icon derived from it (b).