Printed Color Document Storage and Retrieval for Image Databases

Mehmet Celenk and Yuan Shao
School of Electrical Engineering and Computer Science, Ohio University
Athens, Ohio

Abstract

This paper describes a computationally efficient storage and retrieval method for the (R,G,B) color images of the printed documents. The proposed method is developed based on the principal component analysis of image color distribution. A new similarity measure is introduced for image retrieval based on the Tanimoto measure of recognizing similar patterns. This similarity measure is computationally effective since the vector inner product is the only operation needed for its computation.

Several feature sets are experimented in the computer simulation of the algorithm to demonstrate the efficacy of the image retrieval. It is determined experimentally that the proposed method is not affected by substantial changes in the databases. This is due to the fact that the features used for document retrieval are not predefined sets. Rather, they are extracted directly from the document images submitted for recording or searching. This makes the algorithm very robust and attractive for many applications of the image storage and retrieval systems.

Introduction

Printed document-imaging systems have become increasingly popular in the recent years[1] due to advances in computer hardware and software, imaging and video technologies, multimedia systems, and computer communications networks. The imaging systems in this context attempt to replace the established paper-based procedure and processes with intelligent office document processing. The primary purpose of these systems is to convert existing printed paper documents and office procedures using paper-based information flow into digitized documents and automated information flow. First, the printed documents are scanned and their digital images are entered into the imaging database system. The converted paper documents, which appear as images, editable forms, or free text, become objects with attributes and/or content for efficient retrieval and manipulation. If all the images being scanned pertain to a specific predefined form or have well-defined attributes that can be readily extracted from the images of the scanned documents, then form-based indexing could extract the field or attribute values from the document. Form-based indexing can be either manual, which means that an office worker views the image of a data entry and manually enters the form's field values for that image, or automatic, which means an image information system handles all the work. For the latter, an optical character recognition (OCR) system filters the images and recognizes the text for either the entire document (if it is a printed text document) or particular zones in the document (if the relevant information is contained in one or more specific areas or zones). Document-imaging systems also incorporate a work-flow processing system that allows office workers to route the digitized documents to achieve specific goals or tasks. Work-flow systems allow the user to construct office processes or procedures interactively (through a graphical editor) or procedurally (through a scripting language). The documents that are processed through a work-flow can be manipulated or updated by different workers participating in a work-flow process.

In this research, it is assumed that there are a large number of color images of the printed documents in the imaging database. We then describe a form-based indexing method for the printed document-imaging system. In the following sections, we first define the printed document storage and retrieval problem. We then present a new feature selection algorithm and an efficient retrieval method for the printed color documents. Experimental results are provided and conclusions and further research topics are presented in the end.

Problem Statement

Usually, an office document processing system contains many archives involving large numbers of printed matters such as color photographs, computer images, graphics, printed materials, etc. These are typically color images of varying sizes and stored in a computer disk or CD-ROM with R,G,B specification. Such a document archive can be considered as a database and the printed document storage and retrieval problem is defined as follows: Given a query (input) image, it is necessary to obtain a list of images from the database which are most similar in color to the query image. For solving this problem, first there is a need to find a set of features, which represents the color information of a printed matter. Second, a similarity measure has to be developed to compute the similarity between the feature values of two images.

The important criteria in any image retrieval system are the storage and computational efficiency and the retrieval outcome. The selected features as keys or indexes should not require considerable computation time and storage area. The retrieval process should bring out all the images from a

database, which are similar in the keys or feature sets to the query or the input image. The method that we describe in the following sections aims to achieve these objectives.

Description of the Method

In the following sections, first the feature selection for the (R,G,B) color images of the printed documents is developed using principal component analysis. This method requires the computation of the eigenvalues and eigenvectors of the color covariance matrix of an image. For document retrieval, a new similarity measure is proposed based on the Tanimoto measure[4] for two pattern classes. This measure involves the calculation of the inner products of the feature vectors of images being processed.

Feature Selection

The feature selection part of the algorithm is devised based on the principal component or the eigenvalue and eigenvector analysis of a square symmetric matrix. To describe the proposed feature set, we consider the color image of a printed document as a vector valued function $X(i,j)=[R(i,j)\ G(i,j)\ B(i,j)]^T$, whose domain of definition is the MxN spatial grid and whose range is the (R,G,B)-color space. Here, T denotes the matrix transposition. For each stored printed color material, we compute the mean (average) color vector, m, and the color covariance matrix, C, using the expected value operator E as follows:

$$m=E\{X\} \tag{1}$$

$$C=E\{(X - m)(X - m)^T\} \tag{2}$$

The characteristic features of each item in the database are extracted from the color covariance matrix C through the principal component analysis. Since C is a 3x3 symmetric matrix with real elements, it possesses three real eigenvalues, λ_i (i=1,2,3), and three linearly independent eigenvectors, e_i (i=1,2,3), respectively.[2] The eigenvectors of C are all three dimensional vectors, which may be given in column matrix form as $e_i=[e_{i1}\ e_{i2}\ e_{i3}]^T$. The underlying eigenvalues and eigenvectors satisfy the following well known characteristic and matrix equations of C with I being the 3x3 identity matrix:

$$|C - \lambda I| = 0 \tag{3}$$

$$(C - \lambda I)e = 0 \tag{4}$$

Let λ_1, λ_2, and λ_3 be the largest through the smallest eigenvalues and e_1, e_2, and e_3 be the corresponding eigenvectors of C. The eigenvector corresponding to the largest eigenvalue points to a direction in the (R,G,B)-tristimulus space so that the color distribution of an image is concentrated toward that direction. This means that the eigenvector of the largest eigenvalue carries the most useful information about the color content of a document, while the eigenvector due to the smallest eigenvalue has the least information. Since the mean vector determines the position of the image color distribution and the origins of the eigenvectors, then it is necessary to include the mean vector into the characteristic feature sets as one of the key attributes. Hence the feature vector F for an input document is formed using its eigenvectors (e_1, e_2, e_3), eigenvalues (λ_1, λ_2, λ_3), and mean (m). The resultant F is a fifteen-dimensional characteristic vector with a, b, and c being three constant scaling variables as defined below:

$$F = [e_1\ e_2\ e_3\ a\lambda_1\ b\lambda_2\ c\lambda_3\ m]^T \tag{5}$$

Notice that F is computed only once when a document is submitted to the database. If the input imagery is to be recorded, F is stored along with that document as the key or index (F') for search or retrieval. In the following section, we describe the document query in detail.

Printed Color Document Retrieval

For the retrieval of the printed color documents, a query or input image is submitted for search with the key or index being the feature vector F. The feature vector F of the submitted image is matched against those in the database. Let F' be the feature vector of the images stored in the database. The search algorithm brings out a set of closely matching images (i.e., those with $F' \approx F$) as the result of search output. An important criterion for testing the efficacy of the search and retrieval is that the output must include all the similar documents. This means that the output list may have other print outs than the ones similar to the search input. However, this outcome is not a performance degradation factor since the important expectation from the search is that the similar ones should not be missed in the search process. If the similar documents are not brought out, it would defeat the purpose of having an automated search. Such a criterion is important in many applications such as face recognition, digital printing, fingerprint analysis, trademark registration, etc., where the imaging system brings out the short list and the final decision or selection is made by a human expert in the loop.

The similarity measure (S) that we propose in this research for searching the office archives is based on the Tanimoto measure of similarity between two pattern classes. The Tanimoto similarity measure S between the search input with the feature vector F and a document in the database with the feature vector F' is given by

$$S(F,F') = F^T.F' / [F^T.F + F'^T.F' - F^T.F'] \tag{6}$$

Notice that the value of S is close to 1 for two similar color images while its value approaches zero for two dissimilar color documents. This can be concluded from the equation (6) by noticing that the vector inner product FT.F' is the square of the magnitude of F (i.e., |F|2) if the images are identical and it is zero if the images are totally different. Since the values of S vary between 0 and 1, it is necessary to set a threshold T on S in the automated implementation of the method. This enables the image information system to accept the match result when S>T and reject the search output when S<T, respectively.

Experimental Results

The method that we have described in this paper for the storage and retrieval of the printed color documents has been implemented in a Sun workstation using MatLab. For the experimentation, we have created eleven different databases. Each database has included the (R,G,B) color images of various printed materials, such as computer graphics, photographs, images, diagrams, etc. Each color document has been scanned into a spatial grid of varying sizes and quantized into 8 bits per pixel per color component. The resulting digitized color images have been subjected to the principal component analysis for feature extraction before storing them into the databases. First, the mean (average) color vector m' is computed using the unbiased mean estimator for N points of an image data, which is given by

$$m' = \{\sum_{i=1}^{N} X_i [R\ G\ B]^T\}/N - [128\ 128\ 128]^T \quad (7)$$

The color covariance matrix C' of the image is then calculated using the biased covariance estimator for N points of the input data. It is given by

$$C' = \{\sum_{i=1}^{N} X_i [R\ G\ B]^T\ X_i [R\ G\ B]\}/N - m'^T m' \quad (8)$$

Three eigenvalues (λ_1, λ_2, and λ_3)' and the corresponding eigenvectors (e_1, e_2, and e_3)' of the 3x3 real and symmetric matrix C' are computed using the MatLab functions for the equations (3) and (4). The document image is then stored into the database along with its estimated mean vector (m)', eigenvalues (λ_1, λ_2, and λ_3)', and eigenvectors (e_1, e_2, and e_3)' as the key or index parameters (F') for that image.

For the retrieval of a document, an input or query image is submitted to the database system. The query image is analyzed as described above and its feature set (F) is extracted in accordance with the equation (5) for matching against the images in the databases. The matching is performed using the similarity measure given by the equation (6). To determine the effectiveness of the selected feature set as the index of retrieval, a case study is conducted by changing the feature vectors of the input and the stored images. Seven different feature vectors F_i (i=1,2, …,7) are defined using the color mean vector, eigenvalues, and eigenvectors of the search image as follows:

$$F_1 = [e_1]^T \quad (9)$$

$$F_2 = [e_1\ \ e_2]^T \quad (10)$$

$$F_3 = [e_1\ \ e_2\ \ e_3]^T \quad (11)$$

$$F_4 = [e_1\ m]^T \quad (12)$$

$$F_5 = [e_1\ \ e_2\ \ m]^T \quad (13)$$

$$F_6 = [e_1\ e_2\ e_3\ m]^T \quad (14)$$

$$F_7 = [e_1\ e_2\ e_3\ 10^{-3}\lambda_1\ 10^{-3}\lambda_2\ 10^{-3}\lambda_3\ m]^T \quad (15)$$

Similar feature vectors F'_i (i=1,2,…,7) are formed for the database images as given below:

$$F'_1 = [e_1]'^T \quad (16)$$

$$F'_2 = [e_1\ \ e_2]'^T \quad (17)$$

$$F'_3 = [e_1\ \ e_2\ \ e_3]'^T \quad (18)$$

$$F'_4 = [e_1\ \ m]'^T \quad (19)$$

$$F'_5 = [e_1\ \ e_2\ \ m]'^T \quad (20)$$

$$F'_6 = [e_1\ \ e_2\ \ e_3\ \ m]'^T \quad (21)$$

$$F'_7 = [e_1\ e_2\ e_3\ 10^{-3}\lambda_1\ 10^{-3}\lambda_2\ 10^{-3}\lambda_3\ m]'^T \quad (22)$$

Each search image is matched against the database images seven times using the similarity measure $S(F_i, F_i')$, i=1,2, …,7, given by the equation (6), where F_i is the i*th* feature vector of the input image described by the equations (9) through (15) and F_i' is the corresponding feature vector of a database image as defined by equations (16) through (22). Outcome of the similarity measure $S(F_i, F_i')$ between the query image and the database image in question is compared with the threshold T=0.75, which is determined experimentally. If $S(F_i, F_i') \geq 0.75$, the database image is accepted as similar to the input image; otherwise, it is rejected from the output list. The summary of the results are presented in Table I for three different databases (DB1, DB2 and DB3) using the seven different feature sets as described above.

For any query, we define the efficiency of retrieval, E, similar to the one given in reference [3] for a short list of size N. It is given as the ratio of the number of similar images, n, retrieved in the short list to the total number of similar images, N, in the database as follows:

$$E = n / N \quad (23)$$

Table I. Retrieval efficiency (E) of printed documents.

Features	DB1 (E)	DB2 (E)	DB3 (E)	Avr. E
F_1 , F'_1	0.5	0.5	1.0	0.66
F_2 , F'_2	0.5	0.5	1.0	0.66
F_3 , F'_3	0.5	0.83	1.0	0.77
F_4 , F'_4	1.0	0.66	1.0	0.89
F_5 , F'_5	1.0	0.66	1.0	0.89
F_6 , F'_6	1.0	0.66	1.0	0.89
F_7 , F'_7	1.0	0.66	1.0	0.89

Table I summarizes the retrieval efficiency, E, for seven keys of three different databases and the overall average value per feature used. These results are obtained for a short list of documents selected from the databases as the representative samples. The averaging efficiency listed in the same table indicates that the feature sets F_4, F_5, F_6, and F_7 are more effective for retrieval than F_1, F_2, and F_3 for the databases used in the experiment. As far as the computational complexity, the feature vector F_4 is the least computationally complex set with high average search efficiency.

Conclusions and Further Research Topics

In this paper, we have described a computationally efficient storage and retrieval method for the color images of the printed documents. The proposed method is based on the principal component analysis of the color covariance matrix of an image data. A new similarity measure is introduced based on the pattern recognition principal. It is computationally effective since the vector inner product is the only operation that is needed for the similarity computation.

Several feature sets are experimented in the computer simulation of the algorithm to demonstrate the efficacy of the image retrieval. It is determined experimentally that the proposed method is not affected by substantial changes in the databases. This is due to the fact that the features used for image storage and retrieval are not predefined prior to the operation. They are extracted directly from the document images submitted either for storage or retrieval. This makes the algorithm very robust and effective in many applications of the image storage and retrieval systems.

Although the selected color features have produced high retrieval efficiency for the databases used in the implementation, other image characteristics such as histogram features,[3] shape measurements, and texture properties should be included into the search process. This requires further research on the selection and use of additional picture characteristics.

References

1. S. Khoshafian and A. B. Baker, Multimedia and Imaging Databases, Morgan Kaufmann Publishers, 1996.
2. C. R. Wylie, Advanced Engineering Mathematics, 4th edition, McGraw Hill, 1975.
3. M. S. Kankanhalli et al., "Cluster-based color matching for image retrieval," *Pattern Recognition*, Vol. **29**, No. 4, pp. 701 - 708, 1996.
4. J. T. Tou and R. C. Gonzalez, Pattern Recognition Principles, Addison-Wesley, 1974

Biography

Mehmet Celenk received his B.S. and M.S. in Electrical Engineering from the Istanbul Technical University, Istanbul, Turkey, in 1974 and 1976, respectively, and his Ph.D. in Electrical Engineering and Computer Science from the Stevens Institute of Technology, Hoboken, NJ, in 1983. After serving in the Turkish Army as a second lieutenant for two years, Dr. Celenk joined the Ohio University in 1985, where he is now an Associate Professor of Electrical Engineering and Computer Science. His current interests are in computer vision, image processing, distributed computing, and multimedia communi-cations systems. He is a member of IEEE, ACM, SPIE, IS&T, and Eta Kappa Nu.