Document Image Database Retrieval and Browsing using Texture Analysis

John F. Cullen, Jonathan J. Hull and Peter E. Hart

Ricoh California Research Center, Ricoh Silicon Valley Inc.,
2882 Sand Hill Road #115, Menlo Park, CA 94025, USA

Abstract

A system is presented that uses texture to retrieve and browse images stored in a large document image database. A method of graphically generating a candidate search image is used that shows the visual layout and content of a target document. All images similar to this candidate are returned for the purpose of browsing or further query. The system is accessed using a World Wide Web (Web) browser. Applications include the retrieval and browsing of document images including newspapers, faxes and business letters.

1. Introduction

Given the rise in availability of low cost scanners and storage media, the creation of large image databases has become possible. Among the most common images to store in such databases are general business documents. In the last few years it has become cheaper to scan a document and save it in electronic storage than to keep the document as a sheet of paper.

One of the main challenges for document storage systems is the development of effective methods for retrieval from the database. Most systems today use text-based keywords to identify and retrieve documents. This approach depends on manual keyword entry or accurate Optical Character Recognition (OCR) and a full text indexing scheme. In general, reliance on OCR may give poor results for low image quality documents such as faxes. Also remembering text strings can be difficult, giving either zero hits or else large numbers of documents that match generic keywords.

Often a user remembers something about the appearance of a document. The layout and non-text portions of a document provide significant information about its identity. Also, a person's memory for a document can be triggered by some non-text component such as a company logo, picture, number of text columns, or the general layout.

A system is described in this paper that allows for the retrieval of document images based on such non-text features. The system includes a graphical user interface that allows the user to specify the visual characteristics of a query document. From this description, a set of features are generated that are matched against a database of document images. We use texture to describe the types of features in the document. The target document is known only to have a certain type of layout and content, which corresponds to a texture measure for that document. Texture in effect becomes the search key for a document.

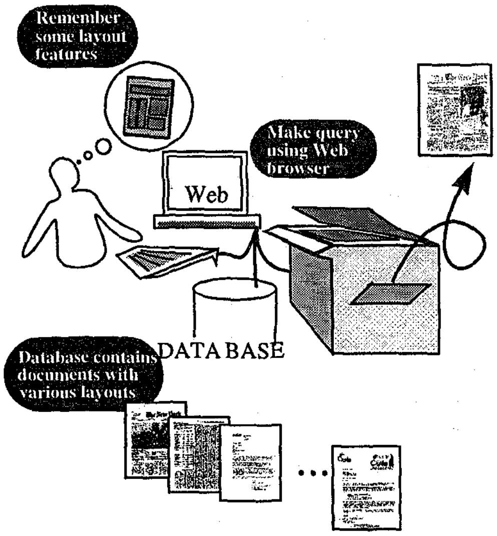
Remember some layout features
Make query using Web browser
Web
DATA BASE
Database contains documents with various layouts

Figure 1: System application scenario.

Several methods are used to specify a query that is matched against the database. These include using an example document image. Alternatively, the user interacts with a simple interface that defines a synthetic document based on the selection of a few categories. The synthetic document is then used as an example from which to extract features to search the database. Using another interface, a graphical tool is used to draw a likeness for a query

0-8186-7898-4/97 $10.00 © 1997 IEEE

document. Once the user specifies the query, a nearest neighbor classifier is used to match against the database. Several images are selected based on the similarity of their texture to that of the example document or the synthetic document. An example of this scenario is shown in Figure 1.

Several applications are envisaged, including browsing a small size image database containing faxes, newspapers, business letters or other limited sets of distinct documents. Also large databases can be searched. The texture is used to find pages that contain textures that are associated with different document layouts. For example, title pages, tables of contents or figures. Lastly, an application is possible where the texture key is combined with text-based queries to give a more accurate search specification.

2. Background

A broad field of research relating to texture analysis exists. Many of these techniques were developed for general image processing[1,2,3]. A few methods have been applied with success to document analysis. Several types of information about documents can be successfully derived using texture. These include image segmentation[11], block classification, and methods for thresholding images. As an example, the block classification scheme of Wong, Casey and Wahl[5] which is based on run-length analysis is appealing in its simplicity. This points to methods for classifying documents using the distributions of basic image components.

A number of approaches for document retrieval exist that combine aspects of image and text analysis. Bloomberg and Chen [7] use word shape features. Hull[10] uses sequences of word letter counts to assign unique identifiers to documents. These techniques can help to overcome the reliability issues associated with OCR.

The DocBrowse system, as developed by Jaisimha et al. [8], incorporates text and graphics for the analysis and retrieval of document images. The focus of the graphics retrieval component is on company logos and hand-written signatures. Doermann et al.[9] describe a framework for intelligent document image retrieval where the image retrieval component focuses on processing low-level features to high level logical structure and generating a descriptor to match against other documents. Experiments on a subset of the University of Washington database gave retrieval results in the 90% range. In general, extensive image processing was carried out to achieve these results and knowledge of the type of document being retrieved was encoded in the algorithms used.

3. System

The proposed system is composed of three components. Firstly, a graphical user interface is required for the user to set up queries and examine results. Secondly, features are extracted from the images in the database. Lastly a matching metric is used that measures the similarity between two document images. A query is performed when a synthetic document or candidate image is supplied to the database and used to return sets of similar documents. A document image in the query result can also be used to initiate another search.

4. User Interface

In many cases, a simple user interface suffices to allow the user to specify a query. In its most basic case, an example image can be supplied to the system and used to retrieve other documents with a similar appearance. This would allow a user to scan a document and retrieve other similar documents in the database. Then using a relevance feedback method, images returned by an earlier query can be used to make more queries.

The next level of sophistication is provided by a user interface on a web page with buttons that specify the characteristics of the query image (see Figure 2). The first column of buttons titled *TEXTURE* allows the user to

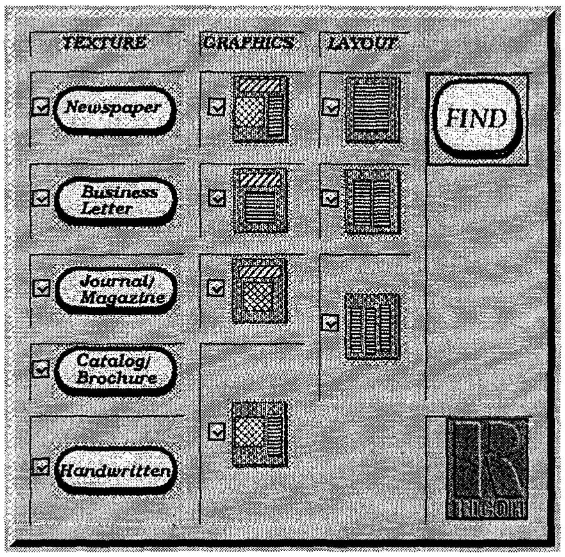

Figure 2: The Web graphical user interface that is used to specify document texture and layout.

specify the document type. This refers to different classes of documents that were considered in this work. Connected components and "interest point" densities over the documents are used to determine texture values. A second column of buttons called *GRAPHICS* allows the user to select the composition of title text and picture of the search document. The third column of buttons called *LAYOUT* allows the user select the numbers of columns of text in the document. By making a selection from each of the three columns the user generates a specification for the type of document that is being sought. The system then generates a 'synthetic document' and the query is processed as in the case where an example document is presented to the system.

In a more sophisticated version of this system, the user generates a query using a graphical tool (see Figure 3).

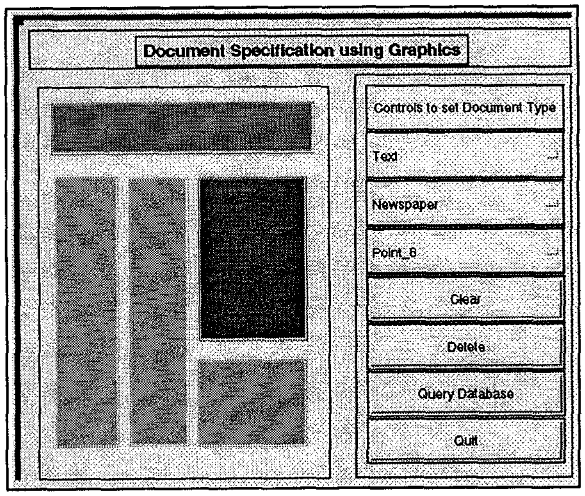
Document Specification using Graphics
Controls to set Document Type
Text
Newspaper
Point_8
Clear
Delete
Query Database
Quit

Figure 3: The Web user interface with a Java applet for drawing a query.

Using a mouse the user draws outlines of a query document. The user selects a component type such as header, author, title, body, footnote, or image, selects an appropriate size and then draws the object as a series of rectangles. After several objects are drawn on the page it is submitted to the document generator. This generates an image from which features are extracted. A query is made to the document image database as before.

5. Features

Many of the differences between document types can be directly mapped to the size of text and to the distribution of text size throughout a document. The extraction of "interest points" and their distribution gives a general estimate of font size. The extraction of connected components and their distribution gives a description of the layout of different size objects in the document.

A feature vector of 80 elements is constructed by analyzing texture features and layout in the document. The texture features are extracted by finding "interest points," calculated using the Moravec Operator [6], and histograms of connected components. The first 20 elements are obtained from interest point densities. An example of how "interest points" vary is shown in Figure 4. The next 20 elements are obtained from a histogram of connected component sizes for the whole document. Twenty more elements are extracted from a vertical projection histogram that emphasizes the number of columns in the document. Lastly, 20 elements are obtained by calculating the density of connected components in each cell of a 5x4 grid over the document. Each feature element corresponds to a histogram bin and are normalized for each document.

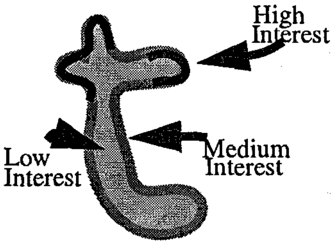

Figure 4: An example of the distribution of 'interest points' around the letter 't'. Interest values change with the amount of curvature in the character. Different densities occur for different point sizes.

6. Matching

A distance measure is used that returns the similarity of two documents. The distance between documents is calculated using the Euclidean distance. The distances are sorted and the nearest documents are selected as hits. A threshold can be used at this stage to limit the number of documents returned by the search procedure.

7. Results

Experiments were carried out on a set of 963 images, which represented a total of 7 different classes of document. These classes were "journal," "business letter," "brochure," "handwritten," "newspaper," "catalog" and "magazine." The principle experiment to validate this system consisted of collecting the feature vectors from each image in the database. Then using a leave one out nearest neighbor classifier the closest three images in the database were found. The class of the top hit was assigned to the unknown document. Figure 5 shows some example results returned by the procedure. Using this method, 939 documents were found to have one of the three nearest documents in the same class as the query. The results indicate that the method had a 97% hit rate for the correct class. The documents that were miss-classified were found to be difficult to classify in general. As an example, the difference between a business letter and a brochure can sometimes be small. An important aspect of this approach is that more than one image can be located that is visually similar in feature space to the query document and the retrieved documents can then be used to locate other relevant documents.

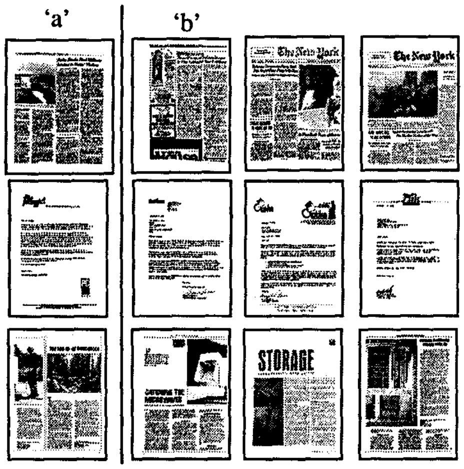

Figure 5: Some examples of querying the database with an example image. 'a' refers to the query, 'b' refers to the results

8. Conclusion

This paper proposed a technique for document image retrieval and browsing that uses image features. Documents are described by texture features such as font size, page layout information and diagrams, that a user can recall and use to specify a query. A Web browser interface lets the user specify searches and examine returned document images.

9. References

[1] Haralick, R. Shanmugam, M. Dinstin, I., "Textural features for Image Classification," IEEE Transactions on Systems Man and Cybernetics. SMC-3, No. 6, 1973.

[2] Chen C., "A study of texture classification using spectral features", International Conference on Pattern Recognition, Munich, 1982.

[3] Laws, K., "Texture Energy Measures", Proceedings on Image Understanding Workshop, Nov. 1979.

[4] Bloomberg, D. "Multi-resolution Morphological Analysis of Documents", SPIE Vol. 1818, pp. 648-662, 1992.

[5] Wong, K., Casey, R., Wahl, F., "Document Analysis System", Proc. 6th International Conference on Pattern Recognition, Munich, October 1982.

[6] Yan Lu. "Interest Operators and Fast Implementation" International Archives of Photogrammetry and Remote Sensing, Vol 27-II, Japan, pp. 491-500, 1988.

[7] Bloomberg, D. Chen, F. "Extraction of text-related features for condensing image documents", SPIE Vol. 2660, 1996

[8] Jaisimha, M.Y., Bruce, A., Nguyen, T., "DocBrowse: A system for Textual and Graphical Querying on Degraded Document Image Data", International Association for Pattern Recognition Workshop, Document Analysis Systems, Oct. 1996.

[9] Doermann, D. Shin, C. Rosenfeld, A., Kauniskangas, H. Sauvola, J., Pietikainen, M. "Development of a General Framework for Intelligent Document Retrieval", International Association for Pattern Recognition Workshop, Document Analysis Systems, Oct. 1996.

[10] Hull, J.J. Document image matching and retrieval with multiple distortion-invariant descriptors," pp383-400, IAPR workshop, Document Analysis Systems, 1994.

[11] Jain, A. Farrokhnia, F., "Unsupervised Texture Segmentation using Gabor Filters", Pattern Recognition, Vol 24, No. 12, pp. 1167-1185, 1991.