Segmentation-Based Retrieval of Document Images from Diverse Collections

Michael A. Moll & Henry S. Baird

Computer Science & Engineering Dept
Lehigh University
19 Memorial Dr West. Bethlehem, PA 18017 USA

E-mail: mam7@lehigh.edu, baird@cse.lehigh.edu
URL: www.cse.lehigh.edu/~mam7, www.cse.lehigh.edu/~baird

ABSTRACT

We describe a methodology for retrieving document images from large extremely diverse collections. First we perform content extraction, that is the location and measurement of regions containing handwriting, machine-printed text, photographs, blank space, etc, in documents represented as bilevel, greylevel, or color images. Recent experiments have shown that even modest per-pixel content classification accuracies can support usefully high recall and precision rates (of, *e.g.*, 80–90%) for retrieval queries within document collections seeking pages that contain a fraction of a certain type of content. When the distribution of content and error rates are uniform across the entire collection, it is possible to derive IR measures from classification measures and *vice versa*. Our largest experiments to date, consisting of 80 training images totaling over 416 million pixels, are presented to illustrate these conclusions. This data set is more representative than previous experiments, containing a more balanced distribution of content types. Contained in this data set are also images of text obtained from handheld digital cameras and the success of existing methods (with no modification) in classifying these images with are discussed. Initial experiments in discriminating line art from the four classes mentioned above are also described. We also discuss methodological issues that affect both ground-truthing and evaluation measures.

Keywords: *document content extraction, document content inventory, document content retrieval, versatility*

1. INTRODUCTION

We have developed a family of algorithms for document image content extraction, which find regions containing machine-printed text, handwriting, photographs, etc in images of documents.[1–5] These algorithms cope with a rich diversity of document, image, and content types, illustrated in Figure 1. The vast and rapidly growing scale of document image collections has been compellingly documented.[6] Information extraction[7] and retrieval[8] from document images is an increasingly important R&D field at the interface between document image analysis and information retrieval.

We classify individual *pixels*, not *regions*, in order to avoid the arbitrariness and restrictiveness of limited families of region shapes, as illustrated in Figure 2. This policy has yielded, to date, modest per-pixel classification accuracies (of, *e.g.*, 60–70%) which already support usefully high recall and precision rates (of, *e.g.*, 80–90%) for queries on collections of documents.[4,9] This flexibility has another advantage: it allows greater accuracy in *inventory* statistics, by which we mean summaries of each page estimating, for each content class, the fraction of page area dominated by that class. And, further, it thus allows useful information retrieval queries, which we will discuss in detail.

In our experimental protocol, both training and test datasets consist of pixels labeled with their ground-truth class: one of Machine Print (MP), Handwriting (HW), Photograph (PH), Blank (BL), etc. Each pixel datum is represented by scalar features extracted by image processing of a small region centered on that pixel; these features are discussed in detail in.[9] We have investigated a wide range of automatically trainable classification technologies, including brute-force k-Nearest Neighbors (kNN), fast approximate kNN using hashed k-d trees, classification and regression trees, and locality-sensitive hashing.[2,3,9]

Figure 1. Thirty-four pages selected from the one hundred and fifty in our complete test set. They are chosen to illustrate the great variety we have included: machine-print, handwriting, photographs, and of course blank regions; color, grey-level, and bilevel (black-and-white) images; English, Chinese, and Arabic languages; magazine articles, newspapers, envelopes, letters, notes; modern and historical documents; rectilinear and complex non-rectilinear layouts; and clean and degraded images.

2. EXPERIMENTAL DESIGN

In previously reported experiments,[4,9] we measured information retrieval performance on document images classified in this way. In this section we briefly summarize that work as a tutorial introduction to the new results which are reported in Section 3. The experiment involved a benchmarking set (A) containing 230 images, doubling the size of any of our previous experiments; and a development set (B) containing 4 images, which was used to test the classifiers ability to discriminate a new previously untested content type. For data set (A), 80 images were placed in the training set, totaling over 416 million pixels and the rest in the test set; then this set was used to train and test classifiers using features described in.[5] This set contains MP, HW, PH, and BL content. Its text includes English, Arabic and Chinese characters each represented by bilevel, greylevel, and color examples. The selection of test and training pages was random except that for each test image there was at least one similar, but not identical, training image. Aside from doubling the size of previous experiments, the main differences in this set from previous are the inclusion of large amounts of machine printed Arabic documents and photographs taken from a hand held digital camera of street signs, license plates and graffiti. Thus these experiments test the discriminating power of the features and weak generalization (to similar data) of the classifiers, but they do not test strong generalization to substantially different cases. Set (B) was divided into a training set and test set of two images each and include the previous four content types as well as a new fifth type, Line Art, that we have never tested before as a separate class.

Each content type was zoned manually (using closely cropped isothetic rectangles) and the zones were ground-truthed. The training data in set (A) was decimated randomly by selecting only one out of every 15000th training sample.

We evaluated performance in two ways, per-pixel accuracy and per-page inventory accuracy:

Per-pixel accuracy: the fraction of all pixels in the document image that are correctly classified: that is, whose class label matches the class specified by the ground truth labels of the zones. Unclassified pixels are counted as incorrect. This is an objective and quantitative measure, but it is somewhat arbitrary due to the variety of ways that content can be zoned. Some content—notably handwriting—often cannot be described by rectangular zones. This in some cases will lead to a per-pixel accuracy score being worse than an image may subjectively appear to be. However, this metric does provide a simple generalization of how well the classifier is performing: for the test set for data set (A), the average per-pixel accuracy score was 73.7% (for the 416 million pixels in the entire training set).

We do not necessarily expect the per-pixel accuracy score to be extremely high due to arbitrariness and even inconsistency in zoning. While we emphasize the strength of our methods in classifying regions of arbitrary shape and layout, we acknowledge that our zoning methodology uses rectangles and that other methods of zoning do exist. Zoning is naturally a very labor intensive process and we believe the method we use is well suited to our experiments, given the time and resources available. In Figure 2, three version of the same image are shown. First is the original document image, second is the image with the rectangles drawn by our zoner overlayed and third is the output of our classification, where each pixel is assigned a color based on the class assigned to it by the classifier. Naturally, the classifier was not trained on this same image it was tested on, therefore the zoned version of this image is only used for scoring. This series of images illustrates two points. First, the inherent bias of using the per-pixel accuracy score as a measure of accuracy as subjectively an observer would (hopefully) agree that the classified output more accurately reflects the actual content and layout of the original image, such as the breaks and shapes of the paragraphs, which the zoning does not. Second, every image used for training is zoned in a similar manner, that is with little regard for fine detail (this is a practical issue) and yet classification does manage to capture more detail about layout. However, it seems to be reasonable to expect zoning to reflect the overall amount of each content type found in an image, and we hope the classifier will do the same.

Per-page inventory accuracy: for each content class, we measure the fraction of each page area that is classified as that class. That is, each page is assigned four numbers—one for each of BL, HW, MP, and PH—which sum to one. This description allows a user to query a data base of page images in a variety of natural and useful ways. For example, in an attempt to retrieve all page images with large photographs with captions, she might ask for all pages containing least 70% photograph and 10% machine print. We believe this measure is superior to per-pixel classification,

We have analyzed the performance of queries of this form: "find all images that contain at least the fraction T of pixels of content class C." This is of course an information retrieval problem[10–12] for which precision and recall are natural measures of performance: precision is the fraction of page images returned which are relevant; and recall is the fraction of relevant documents that are returned.

We issued queries, for every content class, over the full range of threshold values, and summarized the results with precision and recall curves as a function of threshold. For example, the precision and recall scores for MP are shown in Table 1.

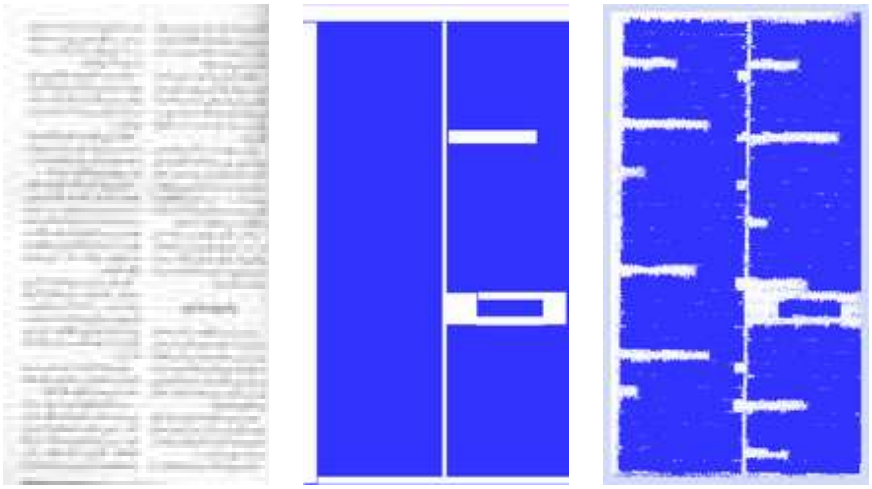

Figure 2. An example of a document image in our test set, the ground truth version of the image and the classified result. On the left is the original greyscale image of machine printed Arabic. The middle image is the ground truth of that image with the rectangles drawn by our zoner overlaying the original image. The image on the right is the output of our classifier. The dark pixels are MP and the white pixels are BL. This test image is never used in the training set for the classifier and therefore the ground truth is only used for scoring purposes. However, all of our training data is also zoned using a similar methodology of using rectangles to capture large-scale detail of the content and layout. Our methods, despite this, are very succesful in capturing the true layout of a document image.

Threshold	Recall	Precision
0.0	1.0000	1.0000
0.1	0.9744	0.7525
0.2	0.9865	0.9125
0.3	0.9698	0.8986
0.4	0.9322	0.8730
0.5	0.9074	0.8901
0.6	0.7436	0.7250
0.7	0.7600	0.6129
0.8	1.0000	0.6000
0.9	0.6667	0.2856

	Recall	Precision
MP	0.894	0.755
PH	0.814	0.835

Table 1. Left: Recall and precision scores for the query "Find all pages with at least the fraction T of machine-print (MP) pixels," over a range of thresholds T from 0.0 to 0.9 (it is rare to find a document image in our test set that is one hundred percent of one content type) , on the test set of data set (A). Values left blank reflect queries which do not return any images. Right: Expected precision and recall scores for each class assuming equal distribution of content across all thresholds.

Threshold	Recall	Precision
0.00 - 0.24	1.0000	1.0000
0.25 - 0.49	0.8236	0.7222
0.50 - 0.74	0.8645	0.8667
0.75 - 1.00	1.0000	0.6800

Threshold	Recall	Precision
0.00 - 0.24	1.0000	1.0000
0.25 - 0.49	0.5833	0.6000
0.50 - 0.74	0.6000	0.7857
0.75 - 1.00	0.7000	0.8750

Table 2. Left: Recall and precision scores for the query "Find all pages with between fraction T1 and T2 of machine-print (MP) pixels," over a range of four "bins" of thresholds, on the test set of data set (A). Right: Similiar table for photograph content in data set (A).

This new method of analysis of infortmation retrieval queries appears to be more accurate than the previously described method. The values for the second lowest bin (the second smallest threshold values) have decreased noticably from the earlier analysis in Table 1. This can be explained by the bins now being independent of each other and not including documents with content greater than the current threshold. However, we still generally see precision and recall scores that are higher than the per-pixel accuracy rate and thus we have a richer, more descriptive way of discusing the inventory of a data set, comparing the clasification of separate images and evaluating the performance of a classifier.

Thus the query "Find every image containing at least 30% machine print" can be answered with 97% recall and 90% precision for data set (A).

If we assume that all threshold values (from 0.0 through 1.0) are equally likely, we can compute expected recall and precision scores for each class (assuming equal distribution of content across all thresholds) as seen in Table 1. This is generalization that must be reconsidered in future work. As we mention, all images must trivially have expected precision and recall scores of 1.0 for the threshold $t = 0$. Unlike previous test set where there were very few, if any, images in the data sets with greater than threshold $t > 0.7$ for any content type, skewing our assumption that all content class distributions are equally likely, this set (A) was designed to inclue a more even distribution of content across all thresholds. The results shown here are on par or higher than previously published results[4] showing that our assumptions hold for more balanced distributions of documents as well.

It is interesting that even at this early stage of development of these document inventory methods, MP and PH enjoy usefully high expected recall and precision, far higher than the per-pixel classification accuracy scores would suggest. This good performance persists up to a threshold of about 90%; the fall off after that can be attributed to the rarity of such images in the test set. While some images at these thresholds were included in the set, this analysis unintentionally favors lower thresholds having higher scores since the queries we are issuing ask for at least some threshold, thus including all images of a greater threshold.

An alternative method of performing this analysis is to instead consider creating "bins" of these thesholds. That is to answer queries of this form: $0.25 < t < 0.49$ The results of these queries are shown in Table 2.

3. DISCUSSION OF EXPERIMENTAL RESULTS

In this section we will analyze individual classificaton results from data set (A). They are shown in Figures (3, 4, and 5). In all examples shown, each test image is on the left with the results of classification next to it on the right as a *classification image* where the content classes are shown in color: machine print (MP) in dark blue, handwriting (HW) in red, photograph (PH) in aqua and blank (BL) in white. As mentioned previously, this set was twice as large as any of our previous experiments in both training (80 images) and test sets (150 images, over 400 million pixels classified). We are aware of the literature on extracting text from photographs and found it interesting how well our current classifier and features performed at locating text within a photograph without any modification from previous experiments.

While we used the same features as discussed in previous experiments and used the same methodology in collecting and dividing the training and test sets (ensuring at least one image from the same source occuring in the training set as the test set), an old problem reoccurred in a new manner. In all previous experiments, the classifier had the most trouble discriminating line art from machine print and blank space, but generally still got handwriting correct about fifty percent of the time. In this experiment however, the classifier completely failed to classify handwriting, classifying only a neglible amount of pixels as such. Instead of getting handwriting wrong or mistaking other classes for it, it now no longer classifies any pixels as such. We attribute this to one new source that is featured much more predominantly in this data set: machine printed Arabic documents. While we have not yet attempted to discriminate between handwritten and machine printed Arabic, we have been including machine printed Arabic in a number of our recent experiments. This is the first set that included nearly as much Arabic as English documents however. Visually, Arabic script looks more

Figure 3. Test page image of a picture from a digital handheld camera taken while in a moving vehicle of a cluster of street signs. In this image 69.8% of the pixels are classified correctly according to the ground truth. The three signs in the image are all at different orientations, are on different backgrounds and are receiving differing amounts of light and shadows. However, the classifier still locates all three signs and little else in the scene as machine print (with the exception of the long horizontal power lines in the background). Also note the large amount of BL (white) pixels identified in this photograph. This brings up an important methodological issue with zoning as a decision must be made as to whether these pixels which are apart of a photograph should be zoned and classifed as PH or BL.

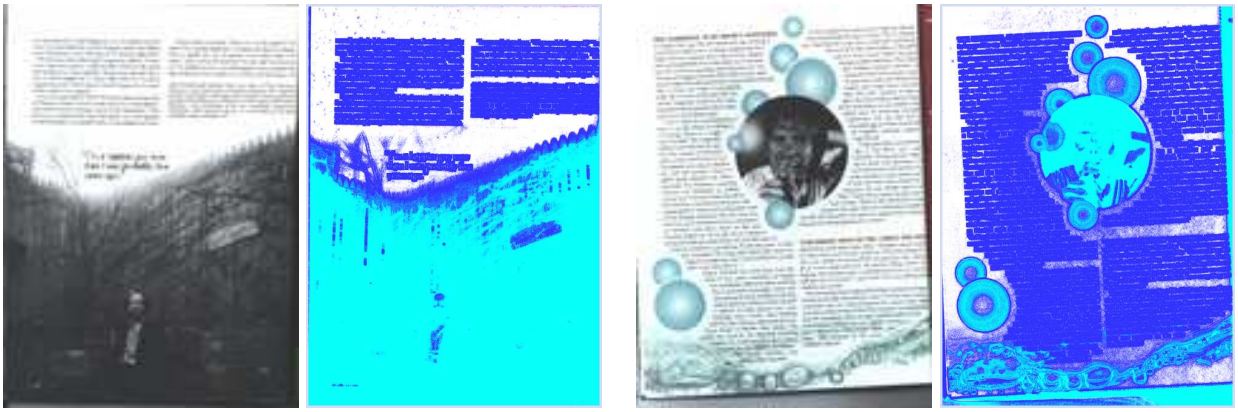

Figure 4. Test page images of two magazine articles containing MP, PH and BL. The image on the left has a per-pixel accuracy score of 77.6% and the right has a score of 61.3%. The image on the left is a greyscale image that has a photograph that bleeds into the text. The classifier nicely captures the actual layout of the paragraphs of text as well as the quote above the photo which is in a different typeface. The image on the right displays two interesting phenomena. First, it shows that our classifier operates independent of skew. Second, it highlights the classifier's ability to handle complex, non-rectilinear layouts.

similar to the handwritten document images in our set than English machine printed images do and we believe that they ended up dominating the handwriting images. This is an issue for future research, as we must consider new features for discriminating between these classes and perhaps consider making machine printed Arabic a unique class.

4. DISCUSSION OF LINE ART CLASSIFICATION EXPERIMENTS

We also conduct experiments using the much smaller development set (B) to attempt to discriminate a fifth class, line art. This is early work for us and we have limited results to report at this time, but hope to have a more thorough discussion with larger test sets by the camera ready deadline. As we collected images for this set, we had trouble deciding how to zone some image as line art. Broadly, we noticed that most thing we thought of as line art could be divided into two

Figure 5. Test page images of two MP documents. The image on the left has a per-pixel accuracy score of 56.0% and the image on the right has a score of 71.8%. The image on the left illustrates the point made previously about the inadequacies of relying on the per-pixel accuracy score to evaluate performance. Subjectively, we would rate this classification as excellent. There is little noise and the classifier has nearly perfectly classified each individual line of text. The low per-pixel score comes as a result of the methodology used to zone this image, as the zoning did not capture the exact layout of the text (instead using large rectangles that ignored the fine details of the layout). Also note that the classifier is not affected by orientation or skew of the image. A shortcoming of the classifier is highlighted here in the large text headlines which show the interior of the letters as PH. The image on the right is a relatively clean and accurate classification of a document image that again begins to classify individual lines of text.

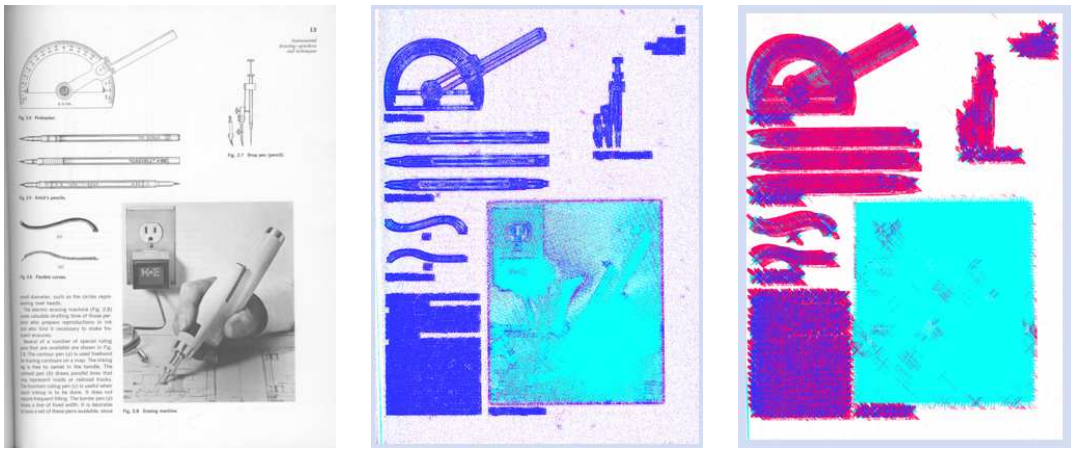

Figure 6. Results of our experiments in classifying Line Art (LA). On the left is the original image, scanned from a book, containing MP, PH, LA (shown as bright red), and BL content. In the middle is the classified image from the features used for the other experiments described in this paper. They focus on the local area around each sample, looking in no larger than a 20 × 20 window of pixels. This resulted in a complete failure to recognize Line Art. On the right is the classified image that resulted from using the same features, but doubling the size of the window. Line Art is now identified, but now there are also more confusions of machine print for line art.

categories: those that resembled machine print, such as engineering drawings, etc and those that resembled handwriting, such as drawing made by hand, etc. Further, thigns like engineering drawings have very large amounts of white (blank) space in them, raising another methodological problem with zoning.

Our first experiment simply divided the images into a training set of two images and a test set of two images, all from the same source, and all strongly resembling each other. Using our current feature set from the previous experiments, the classifier absolutely failed to identify line art. Instead it was classified almost uniformly as machine print (we chose ot focus this experiment on line art of the machine print variety). see Figure 6.

The features that we currently use are focused on the local area around each pixel, none looking in more than a 20 by 20 window around each test sample (a pixel). We make the assumption that the small size of this window is preventing the features from seeing what might make this line art (which strongly resembles machine print) different than pure machine print. Therefore, we reran the classifier with the same features, however we simply doubled the size of the window all of the features use. The classifier now succesfully identified most line art as line art but also started to make more confusions of machine print as line art. In future work, new features will be considered and we will also consider different forms of line art.

5. DISCUSSION AND FUTURE WORK

Intuition and the analysis so far suggests that precision and recall curves provide a richer and more descriptive means of analyzing classification results. We have already concluded that information retrieval performance metrics cannot be derived from per-pixel accuracy scores alone, without substantial assumptions, which still result in only expected value results. We have also observed in the most recently completed experiment that the overall average precision and recall scores are substantially higher than the per-pixel accuracy scores for the same data for machine print and photographs (we do not have enough data from this test set and have known issues with classifying handwriting and blank to make any conclusions about those content types). Per-pixel accuracy scores are highly dependent on the zoning methodology used and greatly affected by small disturbances and fluctuations in the classifier. Future experiments will be conducted to confirm this hypothesis and we believe they will show that these new measures of classifying document image content are more robust than simple per-pixel accuracy scores and confusion matrices.

Inital experiments in classifying line art suggest that discriminating it as a new content class will not be as simple a collecting new images. Also, the addition of large amounts of Arabic machine print appears to have greatly diminished the abilitiy of our classifier to identify handwriting. Both of these occurences indicate the need for careful thought about both feature selection as well as the separation of content classes. It is possible that some of the classes we are attempting to extact may be better thought of as subsets of some classes, rather than independent classes of their own.

Acknowledgements

We are grateful for insights and encouragement offered by Jean Nonnemaker, Pingping Xiu, and Sui-Yu Wang. We acknowledge the continually helpful advice and cooperation of Professor Dan Lopresti, co-director of the Lehigh Pattern Recognition Research laboratory. The support of DARPA Program Manager Joseph Olive under the terms of a seedling grant is also warmly appreciated.

REFERENCES

1. H. S. Baird, M. A. Moll, J. Nonnemaker, M. R. Casey, and D. L. Delorenzo, "Versatile document image content extraction," in *Proc., SPIE/IS&T Document Recognition & Retrieval XIII Conf.*, (San Jose, CA), January 2006.
2. M. R. Casey and H. S. Baird, "Towards versatile document analysis systems," in *Proceedings., 7th IAPR Document Analysis Workshop (DAS'06)*, (Nelson, New Zealand), February 2006.
3. M. R. Casey, *Fast Approximate Nearest Neighbors*, Computer Science & Engineering Dept, Lehigh University, Bethlehem, Pennsylvania, May 2006. M.S. Thesis; PDF available at `www.cse.lehigh.edu/~baird/students.html`.
4. M. A. Moll and H. S. Baird, "Document content inventory & retrieval," in *Proc., IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR2007)*, (Curitiba, Brazil), September 2007.
5. C. An, H. S. Baird, and P. Xiu, "Iterated document content classification," in *Proc., IAPR 9th Int'l Conf. on Document Analysis and Recognition (ICDAR2007)*, (Curitiba, Brazil), September 2007.
6. I. H. Witten, A. Moffat, and T. C. Bell, *Managing Gigagytes: Compressing and Indexing Documents and Images*, Morgan Kaufmann, 1999.
7. Y. Ishitani, "Model-based information extraction method tolerant of ocr errors for document images," *icdar* **00**, p. 0908, 2001.

8. M. Mitra and B. B. Chaudhuri, "Information retrieval from documents: A survey," *Information Retrieval* **2**(2–3), pp. 141–163, 2000.
9. H. S. Baird, M. A. Moll, and C. An, "Document image content inventories," in *Proc., SPIE/IS&T Document Recognition & Retrieval XIV Conf.*, (San Jose, CA), January 2007.
10. J. F. Cullen, J. J. Hull, and P. E. Hart, "Document image database retrieval and browsing using texture analysis," in *Proc., Int'l Conf. on Document Analysis and Recognition (ICDAR97)*, pp. 718–721, August 1997.
11. D. Doermann, "The indexing and retrieval of document images: A survey," *Computer Vision and Image Understanding* **70**, June 1998. Special Issue on "Document Image Understanding and Retrieval," J. Kanai and H. S. Baird (Eds.).
12. H. S. Baird and F. Chen, "Document image retrieval," *Information Retrieval journal (Special Issue)* **2**, May 2000.