Document Content Inventory & Retrieval

Michael A. Moll & Henry S. Baird
Computer Science & Engineering Dept, Lehigh University
19 Memorial Drive West, Bethlehem, Pennsylvania 18017 USA
E-mail: mam7@lehigh.edu, baird@cse.lehigh.edu
URL: www.cse.lehigh.edu/~mam7, www.cse.lehigh.edu/~baird

Abstract

We give an analysis of relationships between expected retrieval performance and classification recognition accuracy in the context of document image content extraction and inventory. By content extraction we mean location and measurement of regions containing handwriting, machine-printed text, photographs, blank space, etc, in documents represented as bilevel, grey-level, 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 given minimum fraction of a certain type of content. In an effort to elucidate this interesting empirical result, we have analyzed the interdependency of classification and retrieval under a variety of assumptions about the distribution of content types in document image collections. We show that under general conditions we cannot derive precision and recall measures from per-pixel classification measures alone, but we can estimate the expected values of these measures. If however the distribution of content and error rates are uniform across the entire collection, our results suggest, it is possible to predict precision and recall measures from classification accuracy and vice versa.

Keywords: *document image analysis, document layout understanding, document content extraction, document content inventory, document content retrieval, document content frequency*

1 Introduction

We have developed a family of algorithms for document image content extraction, able to find regions containing machine-printed text, handwriting, photographs, etc [5, 4]. These algorithms must cope with a rich diversity of document, image, and content types—we have selected for our test set (B): machine-print, handwriting, photographs, and of course blank regions; color (unfortunately not visible in this Proceedings), 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images; English, Chinese, and Arabic languages; magazine articles, newspapers, envelopes, letters, notes; modern and historical documents; rectilinear and complex non-rectilinear layouts; and clean and degraded images. The vast and rapidly growing scale of document image collections has been compellingly documented[11]. Information extraction[8] and retrieval[9] from document images is an increasingly important R&D field at the interface between document image analysis (DIA) and information retrieval (IR).

Following a suggestion by Breuel [10], we classify individual *pixels*, not *regions*, and so avoid imposing arbitrary restrictions on region shape, such as the widely used Manhattan assumption [6]. This approach, in practice, respects and adapts to a wide variety of arbitrary region shapes, as illustrated in detail in [1] in this Proceedings. 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[3]. This flexibility also 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 supports a broad family of information retrieval queries, which we will describe in detail here.

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), photographs (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[3]. 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[5, 4, 3].

2 Experimental Design

In previously reported experiments[3], 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 experiments involved a development set (A) containing 28 images; and a benchmarking set (B) containing 117 images. For

data set (B), 31 images were placed in the training set, and the rest in the test set; then set (B) was used to train and test classifiers using these features and results were "cleaned" using iterated classification methods[1]; analysis of the results of these tests are reported in detail in[3] and excerpted as illustrations here. Their text includes English, Arabic and Chinese characters each represented by bilevel, grey-level, 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. 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.

Each content type was zoned manually (using closely cropped isothetic rectangles) and the zones were ground-truthed. The training data was decimated randomly by selecting only one out of every 9000th 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 B, the average per-pixel accuracy score was 79.1%. The confusion matrix is given in Table 1.

Due to the inevitability of some arbitrariness and inconsistency in zoning, we do not expect per-pixel classification to achieve perfection; however, it does seem reasonable to expect zoning to reflect the overall amount of each content type found in an image, and thus we hope that the classifier can do roughly as well.

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 [7, 2] 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.

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 $t = 0$. Also, there are very few, if any, images in this data set with greater than threshold $t > 0.7$ for any content type, further skewing our assumption that all content class distributions are equally likely. Therefore, it may be more informative to consider expected precision and recall scores for a more realistic range of thresholds, perhaps from 0.2 to 0.6.

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 60%; the fall off after that can be attributed to the rarity of such images in the test set. Most images in the test set were of mixed content type and do not contain high percentages of any single content class.

3 Information Retrieval Performance Analysis

We believe the development of evaluation metrics and ways of interpreting classification results—beyond simply looking at raw pixel accuracies—is of vital importance to the development of these classification techniques. In the previous section we discussed three methods of evaluating performance. Per-pixel accuracy scores give a simple to calculate, quick measure of the performance of a classifier but are of marginal use to an end-user or process of these classifiers. A per-pixel accuracy score is constrained by the methodology used in zoning images and is deteriorated by classifier errors that may or may not be unique to a set of images. Therefore, it is not useful as a means of comparing images, comparing classifiers and certainly not for the retrieval of images. They are useful from a diagnostic point of view, as a confusion matrix can be extracted from them, allowing a developer to identify potential flaws in the classifier, but this is of little use to a user or downstream process.

We also discussed the possibility of a measure of subjective segmentation quality of an image. Thinking about implementation of this measure is more difficult and could naturally be domain or use specific. This measure would not be very useful for diagnostic purposes and while helpful to an end user in providing a rough estimate of classifier performance, is not very useful if identifying content in an image is a goal.

This leads us to applying to the classifier output Information Retrieval queries of the form: "find all images that contain at least the fraction T of pixels of content class C." These queries build on the information provided by per-pixel accuracy scores and we

	BL	HW	MP	PH	Type1
BL	0.178	0.022	0.022	0.005	0.050
HW	0.015	0.050	0.007	0.001	0.024
MP	0.022	0.035	0.383	0.034	0.091
PH	0.013	0.007	0.034	0.170	0.054
Type2	0.051	0.065	0.064	0.039	0.219

Thresh.	Recall	Prec.
0.0	1.000	1.000
0.1	1.000	0.945
0.2	0.953	0.938
0.3	0.803	0.900
0.4	0.795	0.866
0.5	0.812	0.764
0.6	0.714	0.789
0.7	0.875	0.636
0.8	0.750	1.000

	Recall	Prec.
MP	0.856	0.871
PH	0.890	0.735

Table 1. Left: Confusion matrix for per-pixel classification of the entire 75 page test set (B), over 208 million test pixels (47M BL; 24M HW; 93M MP; and 44M PH). The rows label ground truth content types; the columns label the content types assigned by the classifier. The bottom right entry gives the overall error rate: 21.9%. This matrix is discussed in detail in [1]. Center: 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 1.0, on the test set of data set (B). Values left blank (0.9 and 1.0) 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.

believe are of much more practical use. These queries would allow a user to quickly and efficiently search a large set of documents for objects containing specific amounts of content. Typical IR performance measures such as precision and recall, also naturally apply to these queries. We will argue that this measure is less prone to being detrimentally affected by zoning methodology or minor classification errors. We also believe that looking at the content inventory of an image and associated information retrieval queries provide a much richer and useful measure of classifier performance.

3.1 Analysis of Per-Pixel Classification Accuracy and Per-Class Information Retrieval Performance Measures

We begin by asking a series of questions about confusion matrices and precision and recall curves: Is there any relationship between them? Can one predict the other? Is one set more descriptive than the other? We will begin to answer these questions by analyzing in detail several special cases.

In our current experiments we perform classification and IR queries on a 4 content class problem, but for this analysis we will focus on a hypothetical 2 content class problem. We assume the test set contains these two content classes in the following distribution: half of the documents are 100% class MP and half are 100% class PH. For the rest of this analysis we also assume that a confusion matrix contains error rates that are distributed uniformly across the entire collection. We will consider the performance of three different classifiers for this set of documents categorized by the type of errors they commit: a perfect classifier (classifies each content correctly, every time), a one class error classifier (classifies one class correctly all the time and occasionally misclassifies the second class at a known rate) and a two class error classifier (misclassifies each class at a known rate)

3.1.1 Perfect Classifier

Assuming an ideal classifier that never makes a mistake in classifying either of the two classes, the analysis is trivial. The error rates and expected precision and recall values are found in the left table of Table 2. Obviously, if no errors are made in classification, the expected values for both precision and recall are 1. In this case, we obtain the same amount of information from either measure and can infer one from the other.

3.1.2 One-Class Error

We assume that the classifier now classifies MP correctly at a rate α and never misclassifies PH. The error rates and expected precision and recall values are found in the middle table of Table 2. The expected recall value for PH is not affected. The expected recall value for MP is reduced to α as some of the MP documents are now misclassified as PH. These values are derived from the following equations,

$$E[recall] = TP/(TP + FN) = \alpha/(\alpha - (1 - \alpha)) = \alpha$$

where TP (True Positive) is documents that contain MP classified as MP and FN (False Negative) is documents that contain MP misclassified as PH. The denominator must add to 1 as every pixel must be classified as only one of two classes.

Likewise, the opposite is true for the expected precision values. The expected precision for MP is not affected as every document returned as MP is correctly classified as such. However, the expected precision value for PH is now reduced as some documents returned for a query seeking PH will also now return MP documents. These values are derived from the following equations,

$$E[precision] = TP/(TP+FP) = 1/(1+(1-\alpha)) = 1/(2-\alpha)$$

where FP (False Positive) is documents that contain MP misclassified as PH.

3.1.3 Two-Class Error

Finally, we generalize to assume that class MP is misclassified at a rate α and PH is misclassified at a rate β (obviously, in our two class problem we are discussing any pixel misclassified is classified as the other class, there is no third, or error class). The error rates and expected precision and recall values are found in the right table of Table 2. The same discussion and derivations that preceded the One-Class Error classifier applies here as the classifier now misclassifies both classes at a constant rate across all documents in the set. One important thing to remember in this part of the analysis is that we assume that errors are distributed uniformly across all documents in the set. This assumption is necessary at this point to allow us to estimate the expected precision and recall values from the confusion matrices and to infer the estimated precision and recall curves.

3.1.4 Perfect Classifier

The next step in this analysis is to consider a slightly more complex case by considering the same three classifiers applied to a different data set. Previously, we assumed that each document in the set was entirely one content type and there was an equal distribution of each type of document. Now we will assume that every document contains content of both classes, at a fixed rate across the entire set. We assume the classifier never makes a mistake in classifying either content class resulting in a similar analysis as with the first set discussed in the right table of Table 3. We use f_{MP} to represent the frequency of MP content in each image and f_{PH} to be the frequency of PH content. These rates are constant across the entire data set. We see that the confusion matrix remains the same as does the expected precision values. The only difference in these models due to the change in content distribution is a new limit on the maximum expected recall values. Since each document now contains both types of content, recall can no longer be 1 since there is no document that contains entirely one content class only. Therefore, the recall we found for the previous set is multiplied by the frequency of its content class to find the new expected recall value.

3.1.5 One-Class Error

We assume the classifier correctly classifies PH every time and correctly classifies MP at rate α in the center table of Table 3. Same observation as above, the confusion matrix remains the same, however the expected recall value must be adjusted. The precision values remain the same and are not affected by the different content inventory of the images. The derivations of these values are the same as the analysis in the first One-Class Error discussion.

3.1.6 Two-Class Error

We assume the classifier correctly classifies MP at a rate α and PH at a rate β in the right table of Table 3. As with the first document set, for this split content set we again estimate the precision and recall curves for each classifier from the confusion matrix for that classifier and vice versa.

3.2 Conclusions

Beginning with this generalized analysis and also thinking of the data from our experiments, we can start to answer the questions we asked at the start of this section. First, obviously there is a relationship between the measures of per-pixel classification accuracy (seen here in the form of confusion matrices) and per-class information retrieval performance measures. This is not surprising as the IR measures use the per-pixel accuracy scores in part to answer their queries. For some limited cases in the above analysis we see that the it is possible to calculate expected values for precision and recall and the shape their curves will take under some simplifying assumptions. The main assumptions being that that distribution of content and error rates were uniform across the entire data set. With this assumption it appears that there is no difference in what either measure (per-pixel accuracies versus IR measures) tells us about the data and that one can be derived from the other.

In practice however, we cannot make these assumptions. Content will not be distributed uniformly across all documents in a set and as a result error rates certainly will not be uniform. If we are given a confusion matrix for per-pixel accuracies, we cannot calculate what the measures of IR queries performance will be, we can only estimate the expected values of them if we apply the assumptions that we did. As we saw in our experimental data, the actual precision and recall curves present a much more descriptive analysis of the data and allow the documents to be searched and organized in a way not possible with simply per-pixel accuracies alone.

We have concluded that under general conditions we cannot extract precision and recall measures from per-pixel accuracies alone (only estimates of the expected values of these measures). Now we must answer the inverse question: given a complete set of precision and recall curves for every content class, can we derive the confusion matrix for that data set? If not always, are there assumptions that we can apply as before to do so? Can we at least claim that the average per-pixel accuracy score corresponds to the average precision and recall scores? Also, if we cannot directly derive one from the other, can we find for a particular content class for which the per-pixel accuracy is always greater than (or, perhaps, always less than) its precision and recall scores?

4 Discussion and Future Work

Our developing intuition, reinforced by experiments and confirmed by analysis, is that precision and recall curves provide a richer and potentially more useful methodology for analyzing content classification, compared with per-pixel accuracy results (in the form of, *e.g.*, confusion matrices). We have shown here that information retrieval performance metrics can be derived from per-pixel accuracies with the assistance of detailed distribution models, with the result that we can predict the expected values of precision and recall over ranges of natural queries.

In our experiments over the past year, we have consistently observed that the overall average precision and recall scores are substantially higher than per-pixel accuracy scores for the same data for machine print and photographs (we don't have enough data yet to draw conclusions about handwriting and blank). This may be,

	MP	PH
MP	1	0
PH	0	1
E[recall]	1	1
E[precision]	1	1

	MP	PH
MP	α	1-α
PH	0	1
E[recall]	α	1
E[precision]	1	1/(2-α)

	MP	PH
MP	α	1-α
PH	1-β	β
E[recall]	α	β
E[precision]	1/(2-β)	1/(2-α)

Table 2. Summary of algebra discussed in Sections 3.1.1-3.1.3. The left table assumes a perfect classifier, the middle table a one-class error classifier, and the right table a two-class error classifier.

	MP	PH
MP	1	0
PH	0	1
E[recall]	f_{MP}	f_{PH}
E[precision]	1	1

	MP	PH
MP	α	1-α
PH	0	1
E[recall]	α*f_{MP}	f_{PH}
E[precision]	1	1/(2-α)

	MP	PH
MP	α	1-α
PH	1-β	β
E[recall]	α*f_{MP}	β*f_{PH}
E[precision]	1/(2-β)	1/(2-α)

Table 3. Summary of algebra discussed in Sections 3.1.4-3.1.6. Same description as above.

in part, an artifact of our choice of data sets to emphasize documents containing substantial mixtures of content types on each page. More testing on expected data sets with a greater, and perhaps more representative variety of per–page mixtures may be necessary to confirm our observations or to reveal deeper correlations among these measures. We hypothesize that the use of information retrieval queries and their metrics will not be constrained by factors that limit the utility of per-pixel accuracy scores. For example, per-pixel accuracy scores are often highly sensitive to the choice of zoning and ground-truthing protocols, are occasionally greatly affected by small fluctuations in classifier performance. Future experiments will be conducted to test the hypothesis that information-retrieval metrics are often more stable than per-pixel accuracies. One referee kindly suggested an alternative analysis of information retrieval queries that would "bin" precision and recall scores (within narrow ranges of thresholds). For example, rather than search for images containing at least 30 percent of a given class, instead search for all images containing between 20 and 30 percent of that class. Larger test sets—perhaps an order of magnitude larger—will be needed to investigate this.

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] 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.

[2] H. S. Baird and F. Chen. Document image retrieval. *Information Retrieval journal (Special Issue)*, 2(2/3), May 2000.

[3] 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.

[4] 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`.

[5] 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.

[6] R. Cattoni, T. Coianiz, S. Messelodi, and C. Modena. Geometric layout analysis techniques for document image understanding: a review, January 1998. Technical Report, ITC-IRST (Centro per la Ricerca Scientifica e Tecnologica), Trento, Italy.

[7] D. Doermann. The indexing and retrieval of document images: A survey. *Computer Vision and Image Understanding*, 70(3), June 1998. Special Issue on "Document Image Understanding and Retrieval," J. Kanai and H. S. Baird (Eds.).

[8] Y. Ishitani. Model-based information extraction method tolerant of ocr errors for document images. *icdar*, 00:0908, 2001.

[9] M. Mitra and B. B. Chaudhuri. Information retrieval from documents: A survey. *Information Retrieval*, 2(2–3):141–163, 2000.

[10] F. Shafait, D. Keysers, and T. M. Breuel. Pixel-accurate representation and evaluation of page segmentation in document images. In *Proc., IAPR 18th Int'l Conf. on Pattern Recognition (ICPR20006)*, Hong Kong, China, August 2006.

[11] I. H. Witten, A. Moffat, and T. C. Bell. *Managing Gigagytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann, 1999.