Classification of Document Page Images

Christian K. Shin and David S. Doermann
Language and Media Processing Laboratory
Institute for Advanced Computer Studies
University of Maryland
College Park, MD 20742-3275
cshin@cfar.umd.edu, 301-405-1745(tel), 301-314-9115(fax)
doermann@cfar.umd.edu, 301-405-1767(tel), 301-314-9115(fax)

1 Introduction

Searching in a large heterogeneous collection of scanned document images often produces uncertain results in part because of the size of the collection and the lack of an ability to focus queries appropriately. Searching for documents by their type is a natural way to enhance the effectiveness of document retrieval in the workplace [2], and a such system is proposed in [4]. The goal of our work is to build *classifiers* that can determine the type or genre of a document image. We use primarily layout features since the layout of a document contains a significant amount of information that can be used to identify a document's type. Layout analysis is necessary since our input image has no structural definition that is immediately perceivable by a computer. Classification is thus based on "visual similarity" of the structure without reference to models of particular kinds of pages. There has been some classification work reported but most require either domain specific models [3, 5, 6, 8] or are based on text obtained by optical character recognition (OCR) [3, 6, 8].

2 Proposed Method

We propose a method for using layout structures of documents (i.e., visual appearance) to facilitate the search and retrieval of a document stored in a multi-genre database by building a supervised classifier. Ideally, we need tools to automatically generate layout features that are relevant for the specific classification task at hand. Class labels for training samples can be obtained manually or by clustering examples. Once the image features and their types are obtained from a set of training images, classifiers can be built.

In our experiment, we used 64 image features derived from the University of Washington Image Database I (UW-I) groundtruth [1] including the percentages of text and non-text (graphics, image, table, and ruling) zones, the presence of bold font style, font size, and density of content area measured by dividing the total content area by the page area. To obtain

The support of this effort by the Department of Defense under contract MDA 9049-6C-1250 is gratefully acknowledged.

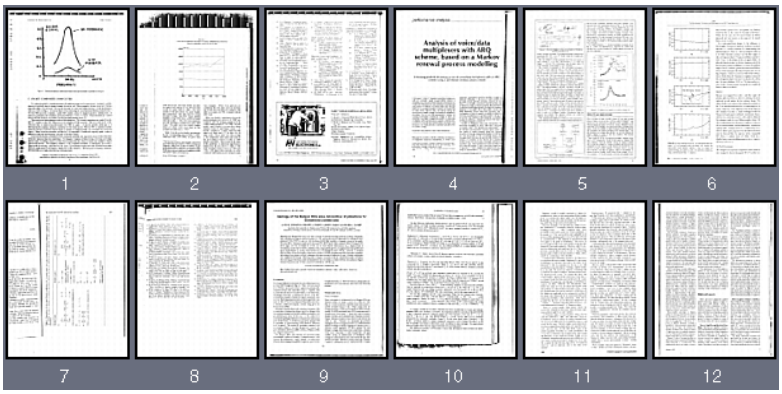
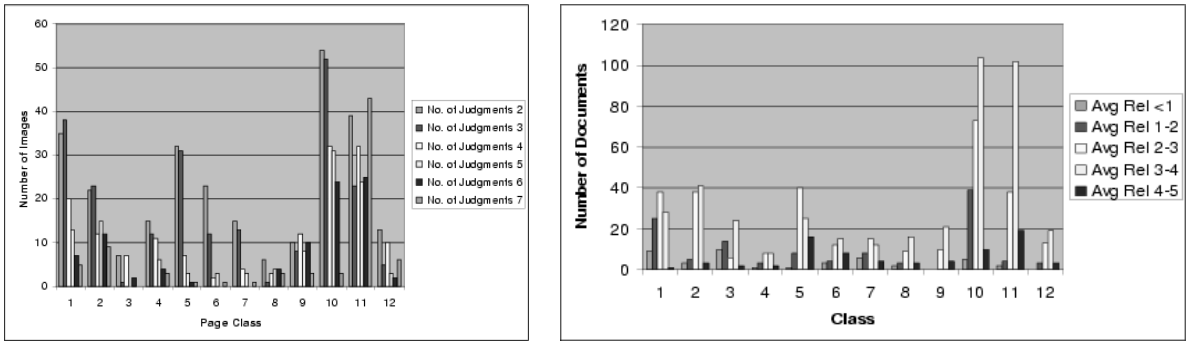


Figure 1: 12 selected representative classes of images for the similarity experiment.

class labels for training samples, we conducted a user relevance test where subjects ranked UW-I document images with respect to a set of query images. We used the relevance rating obtained from the experiment to assign class labels with varying degrees of confidence (Section 3). We implemented our classification scheme using the OC1, decision tree classification software [7] (Section 4).



(a) (b)

Figure 2: (a) Distribution of number of judgments, and (b) distribution of relevance ranking.

3 Similarity Experiment

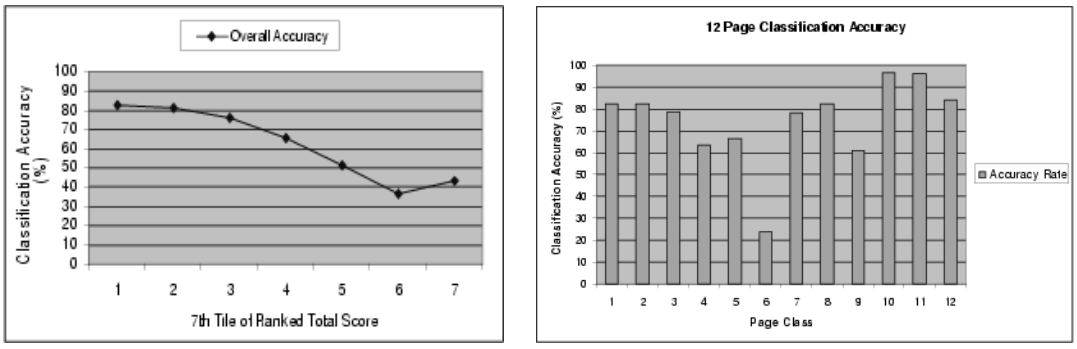
In order to obtain relevance judgments, we conducted a user relevance test using UW-I, a collection of 979 technical article images. These images are single-page images (not full documents), and are treated as individual, independent images.

We first selected 12 representatives of visually different classes from the database (Figure 1). We prepared a survey form that showed the 12 representative thumbnail images, and a

separate package of all UW-I thumbnail images. For each representative image, we asked each of seven subjects to browse through the UW-I thumbnails, and to find images that are visually similar. For each similar image, they were asked to rank each similar image with a degree of relevance ranging from 1 (minimally similar) to 5 (highly similar). From each of our subjects, we received 12 sets of relevance judgment, one set for each representative image. The data obtained from the experiment are summarized in Figure 2. The two key factors are number of judgments (Figure 2a) and relevance ranking (Figure 2b).

4 Classification Results

We have developed a classifier that determines class membership among the 12 classes of layout structure. The relevance judgments obtained in Section 3 are used for determining class memberships of training and test examples. For each representative image, candidate image pair we compute a total score by multiplying the number of judgments made by the maximum similarity rating for that image. For each class, we then ordered test and training images by their total score.



(a) (b)

Figure 3: (a) Page classification accuracy for ordered 1/7th tiles, and (b) page classification accuracy for each document types.

We evaluated performance of the classifier on various training sets based on their strength of the training set determined by the user relevance test. Figure 3a shows classification accuracy for each incremental 7th-tile of the test set using the "leave-one-out" resampling evaluation method. The left most point shows around 83% classification accuracy for the top 1/7th tile of the test set, and the rightmost point is the classification accuracy for the bottom 1/7th tile. This confirms the fact that the classifier performance increases as quality of training set increases. Figure 3b shows overall classification accuracy for each of the 12 document types. We looked at page classes that obtained the four lowest classification accuracies (i.e., page classes 6, 4, 5, and 9), and the similarity experiment (described in Section 3) shows that page

class pairs 5 & 6 and 4 & 9 are perceived as similar to each other.[1] Page classes 4 & 9 are both two column title pages, and page classes 5 & 6 are similar in that they both are two column pages with some graphics.

5 Conclusions and Future Work

In this work, we have proposed and implemented a supervised classification module that is capable of classifying user-defined types in the absence of domain specific models. The classification accuracy we obtained is very promising. We are currently identifying more relevant features, and building an automatic feature extraction module. As discussed in Section 4, classification accuracy is dependent on the quality of training samples. We have plans to develop an effective methodology or mechanism to build efficient document image classifiers by better understanding similarity/distance relationships among the training examples.

References

[1] University of Washington Document Images I CD.

[2] J. Blomberg, L. Suchman, and R. Trigg. Reflections on a work-oriented design project. *PCD '94: Proceedings of the Participatory Design Conference*, pages 99–109, 1994.

[3] A. Dengel, R. Bleisinger, F. Fein, R. Hoch, F. Hones, and M. Malburg. Officemaid - a system for office mail analysis, interpretation and delivery. In *International Workshop on Document Analysis Systems*, pages 253 – 276, 1994.

[4] D. Doermann, C. Shin, A. Rosenfeld, H. Kauniskangas, J. Sauvola, and M. Pietikainen. The development of a general framework for intelligent document image retrieval. In *International Workshop on Document Analysis Systems*, pages 605–632, 1996.

[5] X. Hao, J.T.L. Wang, M.P. Bieber, and P.A. Ng. Heuristic classification of office documents. *International Journal on Artificial Intelligence Tools*, 7:233–265, 1995.

[6] G. Maderlechner, P. Suda, and T. Bruckner. Classification of documents by form and content. In *Pattern Recogniton Letters, 18 (11-13)*, pages 1225–31.

[7] S. Murty, S. Kasif, and S. Salzberg. A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2:1–32, 1994.

[8] S.L. Taylor, M. Lipshutz, and R.W. Nilson. Classification and functional decomposition of business documents. In *Proceedings of the International Conference on Document Analysis and Recognition*, pages 563–566, 1995.

[1] The images belonging to page classes 4 & 9, and 5 & 6 have one of the highest co-identified ratios (91% and 48%, respectively), where subjects identified that an image is, for example, a member of both page classes 4 & 9.