

# Logical Block Labeling for Diverse Types of Document Images

Sabine Bergler, Ching Y. Suen, Christine Nadal, Nicola Nobile, Boulos Waked, and Alan Bloch

Centre for Pattern Recognition and Machine Intelligence  
Concordia University, Suite GM-606  
1455 de Maisonneuve Blvd West, Montréal, Québec H3G 1M8, Canada

Tel: (514) 848 7950, Fax: (514) 848 4522  
email: bergler@cs.concordia.ca, suen@cenparmi.concordia.ca

\*This work was supported by a contract from the Department of National Defense of Canada. We acknowledge the strong support of Mr. Sami Khoury during the tenure of this contract.

## 1. Introduction

Automatic document layout analysis is desirable for two very distinct types of clients, namely those who deal with large amounts of few, potentially standardized layout types (such as forms and letters) for information extraction and the others who deal with a great variety of layout types and are required to archive them (such as government agencies). This paper addresses issues in the automatic segmentation of documents of any type and the assignment of labels from a restricted set to the extracted blocks prior to OCR.

As part of a larger system under development, we describe procedures to deskew and segment pages into blocks based on white runs in [6]. Each block is then classified as a text or non-text block based on the size, distribution and alignment of the bounding boxes of connected components. In the final step, the non-text blocks are assigned a label from one of the following categories: Logo, Photo, ClipSig (clip-art and signatures), Figures, and Dirt. This label assignment is based on the extraction of 28 extracted features which are combined in an expert system.

This paper chooses not to report page segmentation and block labeling results independently, but we report on the decreased performance of a combined system. Errors in segmentation compound with labeling inaccuracies. Human labeling inaccuracies and errors are due to a forced choice for both training and test sets. We illustrate here the difficulties in assessing these problems and in presenting results as well as comparing results with other researchers.

The described procedures are a part of a larger system which, in addition, attempts to identify the script type (ideographic, Roman, Arabic or Russian) and in case of Roman script, also the language (we distinguish only the languages for which dedicated OCR software exists). Our approach to script and language identification has been described in [3, 5, 6].

## 2. Segmentation

After skew detection [6], we segment a document image into blocks separated by white space. Our technique is mainly based on the bounding boxes of connected components. Bounding boxes are first grouped according to size into three categories, small, medium, and large. Typically, graphic

components such as photos and figures occur in large bounding boxes, characters are contained in medium size boxes, and noise in small boxes.

Segmentation proceeds by projecting the medium size boxes first vertically and then horizontally to detect gaps indicating columns and paragraph breaks [2]. In addition, large bounding boxes are further analyzed by projecting both the contained medium size bounding boxes and the pixels contained in the area in order to determine whether it is a framed text area or a table. Also, collinear large boxes of similar height are detected to recognize oversized headlines.

This procedure distinguishes non-text and text regions at the paragraph level. We keep pointers to the coordinates of each block and all contained boxes to facilitate subsequent reading order determination and higher level layout analysis.

The technique of analyzing small, medium, and large size boxes separately yields good segmentation results for overlapping text and non-text regions (such as photographs inserted into a text column). Text contained in oversized text objects such as big titles of articles, however, requires further analysis.

### **3. Non-text Block Labeling**

Labeling relies on extracting features such as block dimension, density, 2x2-grams, 3x3-grams, symmetric pixels, texture [4], and entropy [1]. We use a total of 28 extracted features.

Our training, validation, and test sets are non-text blocks drawn from images scanned at 200 dpi from documents of diverse sources, including business letters, memos, advertisements, journal papers, magazine pages, book pages, etc. The documents span a variety of languages and script types. Using mostly weighted and normalized averages of features on each block type in our training set of 4119 isolated blocks drawn from all 13 categories, we derive 28 heuristics. Using a validation set, we determine which heuristics generate the best results and which have the best consensus results. We report results of an expert system based mainly on best consensus results (of four, three, and two heuristics with corresponding confidence values) combined with a single strong heuristic.

We distinguish thirteen internal non-text block labels (advertisement, clip art, horizontal dirt, vertical dirt, divider, figure, form, graph, logo, map, photo, signature, and table) and five labels for output (logo; photo; "clipsig" including clip-art and signatures; "figure" including figures, advertisements, graphs, maps, tables, and forms; and "dirt", including vertical and horizontal dirt and dividers). Because the statistical features we use do not distinguish all internal types reliably, the output labels group those block types, which have a high rate of confusion.

### **4. Preliminary Results**

We currently test on 1346 non-text blocks which were automatically extracted from 104 documents. The large number of blocks is an artifact of our general segmentation function which produces unwanted blocks. The test set is hand-tagged using the thirteen internal labels to determine accuracy. Human labeling of blocks has proven to be very subjective, in particular for blocks which would not have been extracted by humans, such as small sub-areas of a photo, which have been labeled logo,

clip art, and photo by different people. For reasons of consistency, the training and test sets have been annotated by the same person<sup>1</sup>.

This detail indicates an inherent problem with block labeling. Label type definition as well as label assignment are task specific and different annotators will produce different results. For instance, our training and test sets were tagged by the same person. If the system is to be used for another application, where the labeling will most certainly differ, a new training set must be created.

Another limiting factor is the dependence on a particular resolution. We used 200 DPI throughout the training set. The statistical nature of our procedure leads to poorer performance on images scanned at 300 DPI than 200 DPI.

For the five output labels our expert system achieves 68% correctness, labeling 323 blocks incorrectly and rejecting 104 (7.7%), when forced to assign a label to all areas returned by the segmentation procedure<sup>2</sup>.

It seems difficult, however, to justify a correct or incorrect label for a block that in isolation does not fit any of the thirteen labels (such as a meaningless subarea of a photograph). We have thus evaluated the current system on that subset of the test set which only contains blocks which can be justifiably labeled (note, however, that it still contains undesirably many blocks). On these 739 blocks we achieve 81% accuracy with a 9% rejection rate.

The differences between machine and human labeling in the testing phase originate from the fact that is forced to make a decision, no matter how confident. Humans tend to go through a long period of doubt and analysis when faced with a block which can be classified into different types. In this situation, the human tagger is not completely confident in his/her choice since the user usually makes a choice out of frustration or resignation. Figure 1 shows some examples of non-text blocks which can be classified into more than one type.

An advantage with our machine labeling lies with the pre-filtering module. For example, when a non-text block contained within another non-text block is presented to the program, the program will ignore it and only consider the larger image. The human, however, may not know that the smaller block is spurious and may spend some time labeling it.

## **5. Conclusion**

Our approach is unique in the variety of block and document types it considers. Using very general procedures for segmentation and simple statistical features, we report encouraging results. We are further improving several aspects of this system.

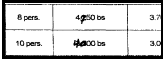


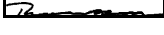

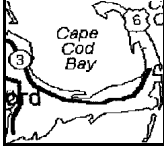
First, we are close to testing a much more refined segmentation procedure which promises to produce fewer blocks more justifiable to the human observer. Such an improvement will have a great

---

<sup>1</sup>We force the system to make a choice from the thirteen labels, even for spurious blocks.

<sup>2</sup>We train and test using imperfect automatic segmentation.

influence on both the training and test sets for the labeling procedure.

<b>Block Image</b>						
<b>Machine Label</b>	Figure	Photo	Logo	Horizontal Dirt	Clip-Art	Clip-Art
<b>List of Human Labels</b>	Table	Logo	Signature Clip-Art	Signature	Signature	Map

**Figure 1:** Sample Discrepancies Between Machine and Human Labeling

Secondly, we are fine-tuning the labeling procedure itself. The inhomogeneous output category "figure", for instance, will be divided. Retraining the program for improved segmentation algorithms and new label categories is a vital part, not only during the development phase of a block labeling system, but also for its deployment, when desired label categories may change, data may differ, or even the assignment of labels to data may change. We are therefore also parametrizing the training procedure and developing fine-tuning tools to make this a portable and useful tool.

## Bibliography

- [1] Castleman, K. R., "Digital Image Processing," Prentice Hall Inc., Inglewood, New Jersey, 1996.
- [2] Liang, J., Ha, J., Haralick, R. M., and Phillips, I. T. , "Document Layout Structure Extraction Using Bounding Boxes of Different Entities," Proc. Workshop on Applications of Computer Vision, Sarasota, Florida, December 1996.
- [3] Nobile, N., Bergler, S., Suen, C. Y., and Khoury, S., "Language Identification of On-line Documents Using Word Shapes," Proc. International Conference of Document Analysis and Recognition (ICDAR), Ulm, Germany, Vol. 1, pp. 258-262, August 1997.
- [4] Shiranita, K., Miyajima, T., Takiyama, R., "Determination of Meat Quality by Texture Analysis," Pattern Recognition Letters 19, pp. 1319 - 1324, 1998.
- [5] Suen, C. Y., S. Bergler, N. Nobile, B. Waked, C. P. Nadal, and A. Bloch, "Categorizing Document Images into Script and Language Classes," Proc. International Conference on Advances in Pattern Recognition, Plymouth, U.K., pp. 297 - 306, November 1998.
- [6] Waked, B., Bergler, S., Suen, C. Y., and Khoury, S., "Skew Detection, Page Segmentation, and Script Classification of Printed Document Images," Proc. Systems, Man, and Cybernetics, San Diego, California, October 1998.