MVA'94 IAPR Workshop on Machine Vision Applications Dec. 13-15, 1994, Kawasaki

A Local-to-Global Approach to Complex Document Layout Analysis

Stephen W. Lam
Center of Excellence for Document Analysis and Recognition (CEDAR)
State University of New York at Buffalo

The UB Commons, Suite 202
520 Lee Entrance
Amherst, NY 14228-2567, USA

Abstract

Document layout analysis is concerned about the decomposition of raster representation of a document into several regions which contain homogeneous entities. This paper describes a new approach to segment documents with complex layout and degraded image quality. The approach uses a local-to-global strategy which can be adapted to a variety of documents. The system was tested on different English and Japanese documents and the experiments had shown promising results.

1. Introduction

Document layout analysis (DLA) is the automatic process of decomposing the raster representation of a document into several regions which contain homogeneous entities, such as text, graphics, table, half-tone, *etc.* Since most documents are printed with black (or dark color) ink on white (or light color) paper, the DLA process always utilizes bi-level raster information. The general approaches to segment an image into regions is based on the detection of large horizontal and vertical streams of white consecutive pixels in the image (see Figure 1).

However, difficulties arise in the segmentation process when:

1. documents have complex layout – the white gaps between regions are usually small and some of the regions are non-rectangular (see Figure 2a).

2. no prior knowledge about the documents to be processed – knowledge-based approach is generally used if the type of documents is known. This will help to determine the segmentation parameters in advance. Without the apriori information, the setting of segmentation parameters has to rely on the runtime information derived from the image analysis.

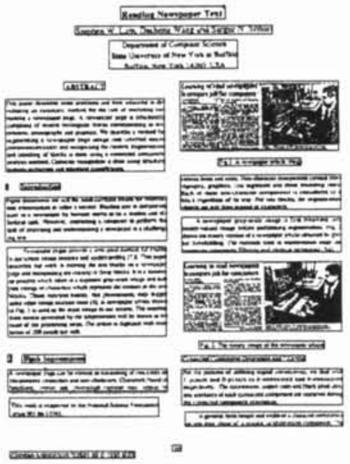

Figure 1: Page segmentation is based on the detection of white space between regions.

3. documents are degraded – detection of regions becomes very difficult when the white gaps are corrupted by noise (see Figure 2b).

This paper describes a new approach to decompose documents with complex layout and degraded quality. The approach uses a *local-to-global* strategy which first partitions the image into several equal-sized sub-images. If white gaps can be found in a sub-image, then the sub-image is segmented into regions. Otherwise, the sub-image will be further partitioned into smaller ones until segmentation is feasible or the image is too small. Once the segmentation has been performed on each of these sub-images, the regions located in the sub-images will be combined progressively to form regions of the whole document. This approach is very robust since it can adapt to a variety of documents. The complexity of the layout analysis directly depends on the document quality and layout complexity.

This approach is being tested on different English

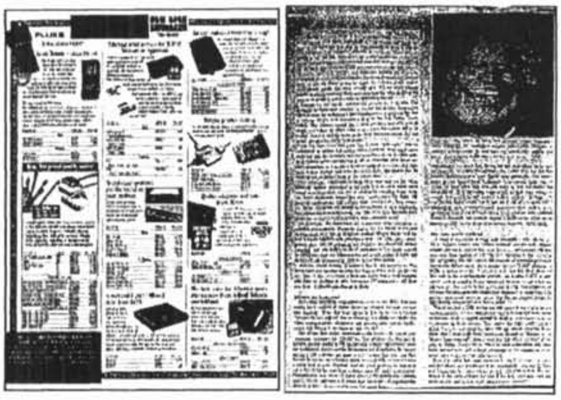

Figure 2: Difficult cases in page segmentation. (a) Complex document layout. (b) Degraded document.

and Japanese documents, such as books, facsimile transmissions, journals, magazines and newspapers. The experiments showed promising results in segmenting complex documents.

2. Background

A lot of work were published in document layout analysis in last 10 years [1, 2, 3, 4, 5, 6, 8, 7, 9]. Most proposed approaches can be divided into two major categories: *top-down* and *bottom-up*. A top-down strategy typically starts by partitioning the document image into horizontal and vertical zones independently by projection profile analysis. The white gaps between regions are shown as valleys in the profile distributions. However, the partition becomes difficult and unreliable when noise is found in the white gaps. It is because the valleys will not be clearly shown in the profile distributions. Setting the thresholds to determine the zones sometimes becomes improbable.

A typical bottom-up strategy starts by merging primitive document components, such as run lengths and connected components, and then merging characters into words, words into lines, *etc.*, until all the components are completed grouped.

All these approaches have difficulties in processing noisy images. In top-down analysis, decision making must depend on the statistical information. Bottom-up strategy is forced to make decisions using the least amount of information (local information), can suffer from the propagation of mistakes.

3. Proposed Algorithm

The proposed approach utilizes the strength of both the top-down and bottom-up strategies and tries to compensate each other weaknesses. Although top-down analysis does not provide reliable segmentation on nosiy and complex layout document at the page level, it can hypothesize probable candidates when it only focus on a small area of the page. As long as the correct zones are among the candidates, they can be picked from the candidates at the later stages. In other words, the top-down analysis in this approach tends to *over hypothesize* the probable zone locations.

Since the candidates are derived from only a small area of the page, candidates from other areas will also be created using the same strategy. Once the candidates have been derived from all the areas, it is required to progressively combine candidates of neighboring areas to form regions of the document. Candidate combination can be best described in form of a relaxation algorithm. Candidates from different areas that satisfy certain spatial constraints are combined to form a larger zone. The candidates of an area that can not combine with those of the other areas will be removed.

The whole process is proceeded in hierarchical form. The top-down analysis first divides the page into four equal-sized sub-images. Each of the sub-images is then further subdivided into four smaller sub-images. The partitioning stops when the sub-images are smaller than a predefined threshold. After candidates have been obtained from all four sub-images, zone candidates are derived by combining the constraint satisfied candidates of these four sub-images. The combination process continues until the four sub-images forms the original document. The result of document segmentation is the regions derived from the candidate combination of the final four sub-images.

4. Zone Candidate Generation

Locating the white gaps in a sub-image is performed by analyzing the horizontal and vertical profile distributions of the sub-image. For a sub-image, R with height h and width w, the profiles are defined as follow:

Horizontal:

$$hpp(i) = \sum_{j=1}^{w} pixval(R_{ij})$$

where

i= 1, ..., h,

R_{ij} is the pixel at ith row, jth column of the subimage R, and

$pixval(R_{ij})$ is the value (1 or 0) of the pixel at ith row, jth column of the subimage R.

Vertical:

$$vpp(j) = \sum_{i=1}^{h} pixval(R_{ij})$$

where

j= 1, ..., w.

Locating the white gap candidates using a profile contribution can be done by finding the profile values lower than a certain threshold instead of zero. Using a non-zero threshold may help to detect noisy white gaps. The calculation of the threshold can based on the statistical estimation of the profile values of the sub-image. The calculation of thresholds for the profiles are as follow:

Threshold t_h for horizontal profile distribution:

$$t_h = c_h * mean(hpp)$$

where c_h is constant between 0.1 and 0.3.

Threshold t_v for vertical profile distribution:

$$t_v = c_v * mean(vpp)$$

where c_v is constant between 0.1 and 0.3.

Once the thresholds are calculated, generation of zone candidates becomes trivial. A horizontal zone candidate HZ at row i if

$$hpp(i) < t_h.$$

Similarly, a vertical zone candidate VZ at column j if

$$vpp(j) < t_v.$$

In order to make such simple generation process work on complex documents and degraded images, it is required to focus only on a small area. It is because the statistical information gathered from a large area always misleads the zone detection process in an imperfect environment.

5. Candidate Combination

Candidate combination is carried out in pair-wise merging operation. For horizontal zone generation, a horizontal zone candidate of a sub-image will try to join with a horizontal zone candidate of its horizontal neighbor sub-image. If the search is successful, the zone dimension in the horizontal direction will be expanded. The process applies each horizontal zone candidate. The similar process also applies vertical zone candidate but the vertical neighboring sub-image is used.

After the new zone dimensions of both directions are derived, it is necessary to check if the new information can be used to form a region. This can be done by aligning the zone dimensions to see if rectangular regions can be formed.

6. Experimental Results

A system which is based on this approach was developed. It was implemented in C and runs on SUN SPARCstations. Figure 3 shows some of the test documents. On a SPARC 10, the system takes 2 seconds to generate zone candidates from all sub-images and takes another 4 seconds to combine the candidates to form regions of the document. The system was tested on 250 images which contain both English and Japanese documents. Currently, the system achieves 95% accuracy rate in locating regions.

Figure 3: Documents used in the experiments.

7. Conclusion

An approach using local-to-global strategy to perform document segmentation has been described. The strength of this approach lies at its simplicity in locating white gap candidates and combining them into regions. The partitioning of a page image into sub-images allows the system fully utilizes local statistical information for more reliable detection of white gaps. The preliminary experiments had shown this approach is feasible for complex document layout analysis.

References

[1] H.S. Baird, S.E. Jones and S.J. Fortune, "Image Segmentation by Shape-Directed Covers", *10th International Conference on Pattern Recognition*, Atlantic City, New Jersey, July 1990.

[2] G. Ciardiello, M.T. Degrandi, M.P. Roccoteli, G. Scafuro and M.R. Spada, "An Experimental System for Office Document Handling and Text Recognition", *9th International Conference on Pattern Recognition*, Rome, Italy, 1988.

[3] F. Esposito, D. Malerba and G. Semeraro, "An Experimental Page Layout Recognition System for Office Document Automatic Classification: An Integrated Approach for Inductive General-

ization", *10th International Conference on Pattern Recognition*, Atlantic City, New Jersey, July 1990.

[4] J.L. Fisher, S.C. Hinda and D.P. D'Amato, "A Rule-based System for Document Image Segmentation", *10th International Conference on Pattern Recognition*, Atlantic City, New Jersey, July 1990.

[5] J. Higashino, H. Fujisawa, Y. Nakano and M. Ejiri, "A Knowledge-based Segmentation Method for Document Understanding", *8th International Conference on Pattern Recognition*, 1986.

[6] D.J. Ittner and H.S. Baird, "Language-Free Analysis", *2nd International Conference on Document Analysis and Recognition*, Tsukuba Science City, Japan, October 1993.

[7] M Krishnamoorthy, G. Nagy, S. Seth and M. Viswanathan, "Syntactic Segmentation and Labeling of Digitized Pages from Technical Journals", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 15, No. 7, July 1993.

[8] T. Pavlidis and J. Zhou, "Page Segmentation by White Streams", *1st International Conference on Document Analysis and Recognition*, Saint Malo, France, September 1991.

[9] K.Y. Wong, R.G. Casey and F.M. Wahl, "Document Analysis System", *IBM Journal Research and Development*, Vol. 26, No. 6, November 1982.