A Ground-Truthing Tool for Layout Analysis Performance Evaluation

A. Antonacopoulos and H. Meng

PRImA Group, Department of Computer Science, University of Liverpool
Peach Street, Liverpool, L69 7ZF, United Kingdom
http://www.csc.liv.ac.uk/~prima

Abstract. There is a significant need for performance evaluation of Layout Analysis methods. The greatest stumbling block is the lack of sufficient ground truth. In particular, there is currently no ground-truth for the evaluation of the performance of page segmentation methods dealing with complex-shaped regions and documents with non-uniformly oriented regions.

This paper describes a new, flexible, ground-truthing tool. It is fast and easy to use as it performs page segmentation to obtain a first description of regions. The ground-truthing system allows for the editing (merging, splitting and shape alteration) of each of the region outlines obtained from page segmentation. The resulting ground-truth regions are described in terms of isothetic polygons to ensure flexibility and wide applicability. The system also provides for the labelling of each of the ground truth regions according to the type of their content and their logical function. The former can be used to evaluate page classification, while the latter can be used in assessing logical layout structure extraction.

1 Introduction

Layout Analysis is a key phase in any document image analysis and recognition system. Layout Analysis comprises three main stages: *page segmentation*, *page classification* and *layout structure extraction*. Page segmentation identifies the regions of interest in the document image, typically coherent printed regions such as text paragraphs or columns, graphics, images, and line art. Page classification determines the type of the content of the identified regions of interest. The goal of the third stage is to describe the structure of the layout in terms of geometric and topological properties of regions (physical layout structure) and, possibly, also in terms of the function of each region (logical layout structure). The latter may be deduced from the physical layout structure but more often than not it requires additional information about the fonts used and the recognised content of each region.

Over the last two decades, a plethora of layout analysis—page segmentation in particular—methods have been reported in the literature. It can be argued that the field is now beginning to mature and yet new methods are being proposed claiming to outperform existing ones. Frequently, each algorithm is devised with a specific application in mind and is fine-tuned to the test image data set used by its authors,

D. Lopresti, J. Hu, and R. Kashi (Eds.): DAS 2002, LNCS 2423, pp. 236–244, 2002.

thus making a direct comparison with other algorithms difficult. The need for objective performance evaluation of Layout Analysis algorithms is evident.

In the wider field of Document Image Analysis, significant activity has concentrated on evaluating OCR results [1][2]. In the case of OCR the comparison of experimental results with ground truth is straightforward (ASCII characters) and lends itself to more elaborate analysis using string-matching theory to calculate errors and associated costs. Consequently, it is possible to automate OCR evaluation using large-scale test-databases [3].

A page segmentation evaluation system based on OCR results was proposed as a result of extensive experience in OCR evaluation at UNLV [4]. Although the OCR-based approach has the benefit of allowing for black box testing of complete (OCR-oriented) systems, it does not provide enough detailed information for researchers in Layout Analysis. In addition, there is not always a direct correspondence between segmentation performance and errors in the OCR result. Finally, this method ignores the non-textual entities on the page.

The other category of page segmentation performance evaluation approaches comprises methods that compare *regions* (segmentation result and ground-truth). There are two kinds of region-based approaches: *pixel-based* and *geometric description-based*. A flexible approach that deals with non-rectangular regions has been developed at Xerox [5]. This approach circumvents the problem of comparing regions when different geometric representation schemes are used, by performing a pixel-level comparison of regions (result and ground truth). The pixel-based comparison, however, is considerably slower than if a description-based comparison were to be used. Furthermore, although halftones are taken into account there is no provision for other non-textual components on a page.

A new layout analysis performance evaluation framework based on *geometric* comparison of regions is being developed at the University of Liverpool [6]. The regions are represented by their contours (as isothetic polygons), enabling fast and efficient comparison of segmentation results with ground truth (there is no need for image accesses). The main benefit of that system is that it can describe complex layouts and compare them (using an interval-based description [8]) with efficiency very close to that of comparing rectangles.

For any performance evaluation approach, the Achilles' heel is the availability of ground truth. As ground-truthing cannot (by definition) be fully automated, it remains a laborious and, therefore, expensive process. One approach is to use synthetic data [3]. It is the authors' opinion, however, that for the realistic evaluation of layout analysis methods, 'real' scanned documents give a better insight. Furthermore, it should be noted that there is currently no ground truth available for the evaluation of methods analysing complex layouts having non-rectangular regions.

For OCR evaluation, definitive ground truth can be relatively easily generated by typing (albeit still time-consuming). In the case of region-based evaluation approaches, however, ground-truthing is not as straightforward. In the pixel-based approach, every pixel of each region has to be correctly labelled, a potentially difficult and very laborious task in the case of complex layouts. In the geometric comparison approach, a flexible and accurate description of regions is essential.

This paper presents a tool that generates ground truth using a flexible page segmentation approach [7] as a first step. This tool facilitates the editing (correction)

of region contours (isothetic polygons) and also enables the specification of the type and function of each region (to evaluate page classification and logical labelling).

A brief description of the new performance evaluation framework and the description of regions is given in the next section, The required ground truth is specified in Section 3. Each of the aspects of the ground-truthing system is described in Section 4 and its subsections. The paper concludes with a discussion in Section 5.

2 Performance Evaluation Framework

The motivation for the new performance evaluation framework is to provide *detailed information for developers* on both the local (page) and the global (whole data set) levels. This in contrast to benchmarking where one is only interested in comparative analysis where a final performance figure suffices. The new framework enables the evaluation of algorithms under an increased number of significant conditions that were not possible under past approaches. Such conditions include complex layouts with non-rectangular regions and regions with non-uniform orientations. The description of each region (and of the page as a whole) is based on interval structures [6] readily obtained from isothetic polygon contours. In this description, the area of a region is represented by a number of rectangular horizontal intervals whose height is determined by the corners of the contour polygon [8]. This (interval structure) representation of regions is very accurate and flexible since each region can have any size, shape and orientation without affecting the analysis method. Furthermore, the interval structure makes checking for inclusion and overlaps, and calculation of area, possible with very few operations.

3 Ground Truth

Region representation is of fundamental importance in any performance evaluation system. A region is defined here to be the smallest logical entity on the page. For Layout Analysis performance evaluation, a region is a paragraph in terms of text (body text, header, footnote, page number, caption etc.), or a graphic region (halftone, line-art, images, horizontal/vertical ruling etc.). Composite elements of a document, such as tables or figures with embedded text, are considered each as a single (composite) region.

The choice of a region representation scheme is crucial for efficiency and accuracy. While rectangles (bounding boxes) enable the simplest region comparisons, they are not suitable for complex-shaped regions. In the performance evaluation framework mentioned above, any region can be represented by an interval structure derived from an isothetic polygon. Isothetic polygons describe regions flexibly and they can be easily used in the context of other performance evaluation applications [9] as well as in the system mentioned above. Simplicity is also retained as for rectangular regions the isothetic polygon would be in essence a bounding box.

To ensure simplicity and wide applicability as outlined above, the chosen ground truth representation of a region is an isothetic polygon.

The ground truth generated by the tool described in this paper includes the following for each region: its description in terms of an isothetic polygon (a list of vertices in anti-clockwise sequence), the type of its contents, and its functional (logical) label.

4 The System

The input to the system is a page image. At the moment, only binary images are supported. Regions of interest are identified using a page segmentation method. The objective is to identify regions in the image as close as possible to the target regions. Naturally, no page segmentation method would produce perfect results. Therefore, it is desirable that the inevitable errors are as straightforward as possible.

The page segmentation method used here is the white tiles method [7]. Apart from its ability to identify and describe complex-shaped regions in different orientations, it is also fast and produces isothetic polygons. If required, another page segmentation method can be used, either in addition or as a replacement.

Having an initial description of the regions, the user has the option to edit individual polygons, merge polygons that should be part of the same region, and split polygons that should represent different regions.

When the user is satisfied with the representation of the regions, further information (type of region contents and functional label) can be specified for each of the polygons in the description of the page.

The following sections describe these processes in more detail.

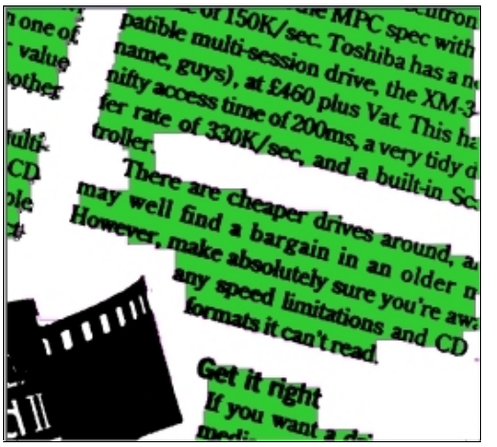

Fig. 1. An example of region description after page segmentation.

4.1 Page Segmentation

The white tiles page segmentation method [7] is part of the white tile approach that also performs page classification [8] and region orientation estimation [10], using the

description of the background space. It is equally applicable to the segmentation of images of document pages having both traditional and complex layouts. The underlining idea is to efficiently produce a flexible description (by means of tiles) of the background space that surrounds the printed regions in the page image under conditions such as the presence of non-rectangular regions and regions with different orientations. Using this description of space, the contours of printed regions are identified with significant accuracy. The white tiles approach is fast as there is no need for skew detection and correction, and only few simple operations are performed on the description of the background (not on the pixel-based data).

In the ground-truthing system, the white tiles page segmentation method is set to slightly over-segment regions. This ensures that the number of inadvertent mergings of regions of different types is kept to a minimum.

At the end of the page segmentation process, all printed regions on the page image are represented by isothetic polygons. An example of the description of part of a page can be seen in Figure 1.

Fig. 2. A logically coherent region (running footer) described as more than one region.

4.2 Region Editing

As with any page segmentation method, there will be cases where a logically coherent printed region has been described by more than one polygon. An example can be seen in Figure 2 where each word in the running footer has been described by a separate polygon. On the other hand, there may be cases where more than one logically distinct regions have been described by a single polygon. This is frequently the case with paragraphs in a single column when there is no extra spacing between the paragraphs (see Figure 1).

Fig. 3. The resulting single region after merging.

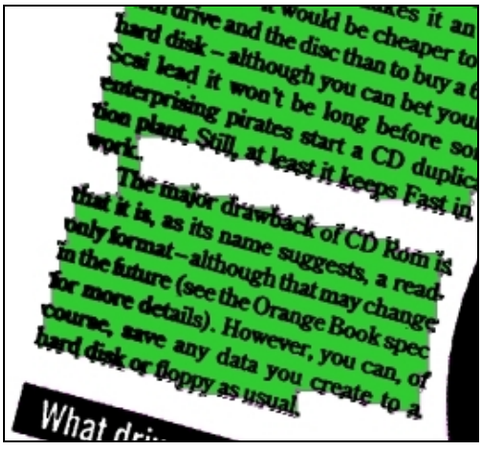

Fig. 4. More than one paragraph in a column described as one region.

The system provides editing facilities for merging and splitting polygonal regions. In addition, the position of each vertex of a polygon can be adjusted (by clicking and dragging) to ensure the regions are accurately described to the user's satisfaction. It is important to mention here that the system makes no assumptions about the orientation of the regions and about their shape. The system can be used to ground-truth skewed images as well.

Merging regions

When two or more regions resulting from page segmentation must be merged, the user selects the corresponding polygons and clicks on the 'merge' button. The merging process sorts all selected polygons according to their position and starts

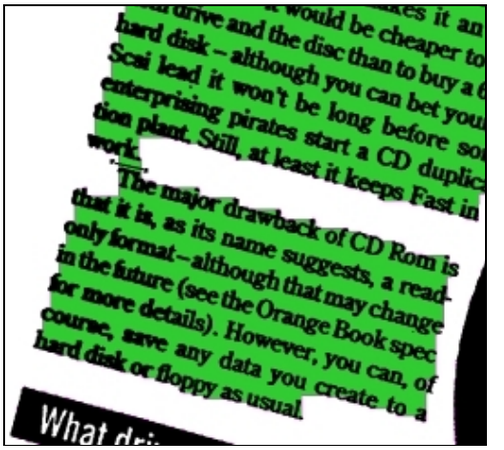

Fig. 5. Placement of a line indicating the direction and position of splitting.

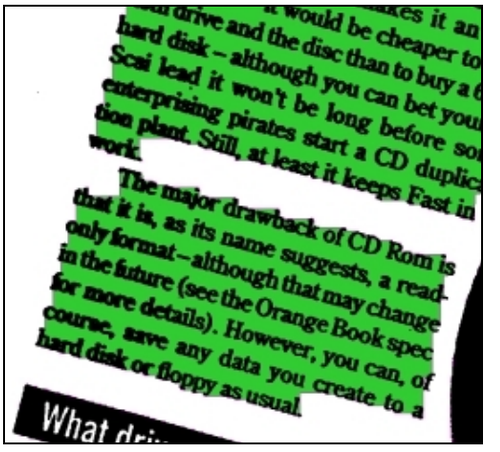

Fig. 6. Result of splitting.

merging them into an aggregate polygon, one at a time, starting from the top-left-most one. The resulting polygon has the same structure (list of vertices ordered anti-clockwise) as the polygons resulting from page segmentation.

The aggregate polygon resulting from merging the words in the running footer of Figure 2 can be seen in Figure 3.

Splitting Regions

The division (splitting) of an identified region into two separate regions is slightly more complicated than the merging operation described above. The operation involves the examination of the background space along the split direction so that each of the resulting regions does not contain excess space.

First, the user selects the polygon corresponding to the region to be split. For instance, one may wish to split the column in Figure 4 into separate paragraphs. Then they place a line indicating the direction of the split (see Figure 5) and click the 'split' button.

The system first calculates the intersection of the splitting line with the selected polygon. Within the polygon boundary either side of the split line, black pixels are counted along the direction of the split line (there should be no black pixels along the split line itself. As soon as the first black pixels are encountered along parallel lines on either side of the split line, the boundaries of the new polygons (where the split will occur) are fixed to these lines. New vertices are inserted in the original polygon at its intersections with the lines denoting the split positions. To ensure conformity with other polygons, the vertices are ordered so that the resulting two polygons will have their vertices ordered anti-clockwise. The result from the splitting of the region in Figure 4 can be seen in Figure 6.

Region Labelling

Once the regions are correctly described by the polygons (after editing), the user can enter further information about the regions. By right-clicking in a region they can select the option to associate a region with its content type and logical label. A dialog box appears that is filled in with the necessary information and optional comments. For ease of use, regions that have been indicated by the user as completely edited and specified are drawn in a different colour.

5 Concluding Remarks

A ground-truthing system for layout Analysis performance evaluation has been described in its context (the system is currently in the last stages of development and it is anticipated that it will be ready for demonstration at DAS'02). The system addresses a significant need to produce ground truth, especially for complex layouts (where no ground truth exists at the moment).

Flexibility is one of the main advantages of this system. The choice of region representation (isothetic polygons) enables accurate description and the resulting ground truth is not only applicable to the performance evaluation framework described here. Furthermore, the page segmentation method can be changed or enhanced, or even use the results of alternative methods, each better tuned perhaps to different types of documents.

The system is implemented in Visual C++ and works on PCs. This choice was made in order to ensure wide compatibility and enhanced performance. The system will be made available to the document analysis community.

References

[1] G. Nagy, "Document Image Analysis: Automated Performance Evaluation", Document Image Analysis Systems, A.L. Spitz and A. Dengel (eds.), World Scientific, 1995.

[2] C.H. Lee and T. Kanungo, "The architecture of TRUEVIZ: A groundTRUth / metadata Editing and VisualiZing toolkit", *Symposium on Document Image Understanding Technology*, April 23 –25, 2001, Columbia, Maryland,

[3] I.T. Philips, S. Chen and R.M. Haralick, "CD-ROM Document Database Standard", Proceedings of 2nd International Conference on Document Analysis and Recognition (ICDAR'93), Tsukuba, Japan, 1993, pp. 478–483.

[4] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated Evaluation of OCR Zoning", IEEE Transactions on Pattern Recognition and Machine Intelligence, Vol. 17, No. 1, January, 1995, pp. 86–90.

[5] B.A. Yanikoglu and L. Vincent, "Pink Panther: A Complete Environment for Ground-Truthing and Benchmarking Document Page Segmentation", Pattern Recognition, Vol. 31, No. 9, 1998, pp. 1191–1204.

[6] A. Antonacopoulos and A Brough, "Methodology for Flexible and Efficient Analysis of the Performance of Page Segmentation Algorithms", Proceedings of 5th International Conference on Document Analysis and Recognition (ICDAR'99), Bangalore, India, 1999, IEEE-CS Press, pp. 451–454.

[7] A. Antonacopoulos, "Page Segmentation Using the Description of the Background", Computer Vision and Image Understanding, Special issue on Document Analysis and Retrieval, Vol. 70, No. 3, June 1998, pp. 350–369.

[8] A. Antonacopoulos and R.T. Ritchings, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles", Proceedings of 3rd International Conference on Document Analysis and Recognition (ICDAR'95), Montreal, Canada, 1995, Vol. 2, pp. 1132–1135.

[9] B. Gatos, S.L. Mantzaris and A. Antonacopoulos, "First International Newspaper Contest", Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, USA, September 2001, pp. 1190–1194.

[10] A. Antonacopoulos, "Local Skew Angle Estimation from Background Space in Text Regions", Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR'97), Ulm, Germany, August 18–20, 1997, IEEE-CS Press, pp. 684–688.