Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR'99), Bangalore, India, August 1999, IEEE Computer Society Press, pp. 451-454.

Methodology for Flexible and Efficient Analysis of the Performance of Page Segmentation Algorithms

A. Antonacopoulos[1] and A. Brough
Department of Computer Science, University of Liverpool,
Peach Street, Liverpool, L69 7ZF, United Kingdom

Abstract

This paper presents part of a new DIA performance analysis framework aimed at Layout Analysis algorithm developers. A new region-representation scheme (an interval-based description of isothetic polygons) and a corresponding comparison approach are introduced. These enable fast and accurate geometric comparison of ground-truth with results of page segmentation, improving on current evaluation methods. Complex layouts are accurately described and Layout Analysis methods that handle them can be effectively evaluated. A further benefit of the new approach is that it measures the accuracy of the description of regions, an issue which is important for complex-layouts involving non-text regions.

1. Introduction

The need for objective evaluation of the performance of Document Image Analysis algorithms is evident as algorithms mature and application areas become diverse.

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

Large-scale testing and evaluation is essential not only for OCR but for each of the subsystems involved in DIA also. The identification of regions of interest in the document page image (*page segmentation*) and the type of their content (*page classification*) are significant stages that seriously affect the performance of subsequent DIA stages (e.g., OCR, Document Image Understanding etc.).

The work described in this paper is part of a new framework being developed for analysing the performance of Layout Analysis subsystems. In this paper, the focus is on methods and issues involved in performance analysis of Page Segmentation algorithms.

It should be noted that there is a distinction between *comparative benchmarking* of a group of algorithms [3][4] (aimed at end-users) and *performance analysis* of individual algorithms [6] (more useful to algorithm developers). The work in this paper is concerned with performance analysis.

Past approaches to the evaluation of page segmentation and classification methods fall into two broad categories: *OCR-based* and *region-based*. In the first category, an 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 DIA (OCR-oriented) systems, it does not provide enough detailed information for (page segmentation) algorithm developers.

The second category of evaluation approaches comprises methods that perform either a *geometric* or a *pixel-based* comparison of regions. The developers of the University of Washington (UWASH) document database [2] have made provisions for ground-truth region-description using bounding rectangles. For each text region an 'annulus' is formed by two rectangles, one containing the other. However, to the best of the authors' knowledge, a method using these 'annuli' has not been reported in the literature. It should also be noted that, while many types of documents have rectangular regions, any approach based on this database will not be applicable to methods dealing with complex layouts [5].

The geometric comparison approach is not straightforward as, to be successful, there must be a very accurate description of regions without excess background and the region-representation schemes of the result and the ground-truth must be directly comparable. To circumvent these problems, a pixel-based approach has been developed at Xerox [6]. This approach performs a black pixel comparison between regions (segmentation result and ground truth). This approach is quite flexible as it deals with non-rectangular regions. The pixel-by-pixel comparison, however, is significantly more time-consuming than if a description-based comparison were to

[1]Corresponding author (e-mail: aa@csc.liv.ac.uk)

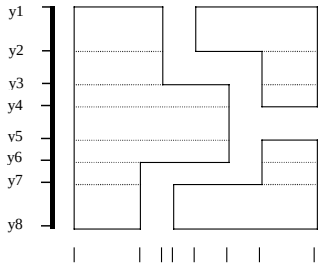

Figure 1. Global interval partitioning.

be used. Finally, as it is very important to ensure that every pixel is correctly labelled, the ground-truthing can be a lengthy and tedious task if additional pixels are introduced by thresholding colour backgrounds or if noise is present [7].

This paper introduces and compares two approaches that perform a fast polygon-based geometric comparison of regions. Both approaches use a versatile interval-based region description. Apart from speed, a significant benefit of both approaches is that they use ground-truth polygons that very accurately describe regions without excess background space. This ground-truth can be relatively easily obtained from the White Tiles page segmentation approach [5].

In the following section, the general DIA performance analysis framework is briefly described. Details about the region-representation scheme used and the approaches for comparative analysis can be found in Section 3. The issues surrounding the analysis of the performance of page segmentation and the proposed methods are discussed in Section 4, which concludes the paper.

2. The proposed framework

A new performance analysis and evaluation framework is under development at the University of Liverpool. It will consist of new performance analysis methods and a new test-image database. Layout Analysis subsystems are the main focus, while the evaluation of Logical Layout Analysis is also of significant interest (for Information Retrieval and Document Image Understanding).

With respect to Page Segmentation and Classification, the main benefits of the new framework are efficiency (paramount for large-scale evaluation) and flexibility. Significant efficiency gains result from a description-based comparative analysis of regions, which avoids time-consuming image accesses. The flexibility of the system is evident in different respects. First, it enables the evaluation of algorithms under an increased number of significant conditions that was not possible under past approaches. Such conditions include complex layouts with non-rectangular regions, colour and textured backgrounds and non-uniform region orientation. Secondly, the evaluation methods can provide information at various levels of detail. At the local level, detailed information is available for each region on a number of conditions. This detailed information is aimed at the algorithm developers. At the global level, information is available for the performance of an algorithm on a whole page or set of pages. A global score is also given at this point for end-users to compare and contrast different algorithms.

3. Region representation and comparative analysis

A region is defined here to be the smallest logical entity on the page. For the purpose of assessing Page Segmentation and Classification, a region is a single 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 region-representation scheme plays a critical role in the efficiency and accuracy of the performance analysis strategy. The proposed scheme is an interval-based description, which has its origins in [8]. Since the contour of each region can be described by an isothetic (having only horizontal and vertical edges) polygon [5], a region is represented by a number of rectangular horizontal intervals whose height is determined by the corners of its 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 [8].

The White Tiles page segmentation method [5], which can identify and describe regions very accurately even in the presence of complex layouts and severe skew, is used as a first stage of the ground-truthing process. With a small number of point-and-click operations to correct the results the final ground-truth polygons are obtained. The given description of each region resulting from a page segmentation method under consideration (e.g., set of bounding boxes) is converted into a minimum-enclosing isothetic polygon and represented in the same way as the ground-truth regions.

The regions on a whole page can be described by a *global interval structure*. In this structure, intervals extend across the page, in the horizontal direction. As a result, all horizontally adjacent region polygons are broken into intervals having the same start and end in the vertical direction. An example of such a description can be seen in Fig. 1.

For ground-truth description, a global interval structure represents all regions, each described by the closest-fitting isothetic polygon around that region, a *ground-truth polygon* (*GTP*), split into intervals. The regions resulting

from the application of a page segmentation method are referred to as *segmentation polygons* (*SP*) and are also described in a global interval structure.

The goal of the comparative analysis that follows is to identify, given the GTP and SP structures, the following situations (or combinations thereof, see Fig. 2):

1. *A SP correctly describes a GTP.*
2. *A GTP is split. More than one SP is involved in the description of that GTP.*
3. *A GTP is merged with one or more other GTPs. A SP describes more than one GTP (or parts of GTPs).*
4. *A GTP is partially missed. Part of a GTP is not described by any SP.*
5. *A GTP is totally missed. No SP describes that GTP.*
6. *A SP does not describe any GTP (or part of one). The SP has been wrongly introduced by page segmentation (possibly due to noise in the image).*

In the first case, when a SP has correctly matched a GTP, the analysis method assesses the *accuracy of description* by calculating the extraneous background space included in the SP. This is a new and very useful feature (especially for complex layouts) introduced here which is possible to calculate due to the accurate region-description of the GTP.

For each of the erroneous cases 2–4 above, or combinations thereof, there can be different degrees of severity, due to both the *extent* (e.g. the percentage of GTP missed) and the *nature* of the error (topology and type of region). To express the latter quantitatively, detailed information is given in the form of area and number ratios (e.g., correct/total) for all regions and for regions of a specific type. In terms of reporting a global score, specific penalties (user tunable) are associated with each situation and combinations.

Before performing any comparison, correspondences between SPs and GTPs must be identified. This is acknowledged as a non-trivial problem for region-based systems [6]. The approach of Yanikoglu and Vincent [6] creates region-maps (two reduced-resolution images) to overcome this problem. Although this is a flexible approach, it requires two more instances of an image (albeit reduced) and building region maps at run-time requires pixel-level image accesses.

In keeping with the efficiency ethos, alternative ways to identify region correspondence using the polygonal descriptions and global interval structures have been investigated. These are examined below.

3.1. Maximal polygons approach

This approach is inspired by the 'annuli' mentioned in the specification of the UWASH database [2]. For each GTP, a *maximal description polygon* (*MDP*) is created. A MDP is an isothetic polygon, which expands the corresponding GTP in all directions until it either meets other GTPs or reaches the edges of the image. A threshold is applied to avoid very narrow extensions of a MDP being created between adjacent GTPs. Automated MDP creation is straightforward using the GTP global interval structure.

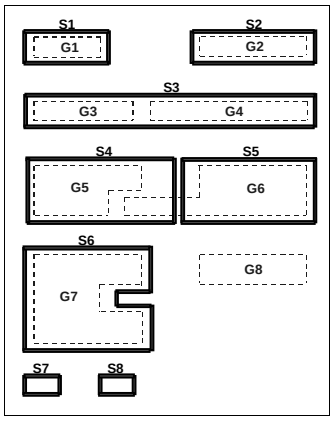

Figure 2. Example of GTPs and given SPs.

In essence, the area between a given GTP and its corresponding MDP is the legal area within which all edges of a SP should fall if that SP is correctly describing the GTP. If a SP (or part of it) falls outside this legal area then a merge or miss (partial or total) has occurred.

The MDPs of all GTPs on the page are described in a global interval structure. This MDP structure can, in the first instance, serve as a map to identify one or more GTPs that may be linked to a given SP (the MDPs of adjacent GTPs overlap). Furthermore, when a given SP is compared to a GTP, the MDP of that GTP can be used to quickly determine whether the SP has merged any GTPs (if the SP is totally inside an MDP, no merging occurs).

3.2. Reverse problem approach

An alternative approach to checking whether a given SP lies within one or more MDPs (i.e. matching an SP with GTPs eventually), is to reverse the problem and identify which SP (or SPs) describe a given GTP. The process starts with a GTP and assesses whether it is contained with one or more SPs. The rationale is that since a GTP is the smallest polygon fitting around a region, more often than not, GTPs will be inside SPs. This renders the matching more straightforward and requires fewer operations. More importantly, comparisons with MDPs are not required.

The algorithm starts by considering each GTP in turn and identifying overlaps (if any) between its constituent intervals and those of the SP global structure. These overlaps and the number of SPs encountered are then analysed to determine whether the GTP has been correctly identified, split or missed (partially or wholly). After all GTPs have been considered, the number of times each SP was encountered by different GTPs is examined to determine whether the SP has merged GTPs or it has been wrongly introduced (no overlaps).

The main step (computationally) is the identification of SP intervals overlapping with the intervals of a given GTP. This matching of intervals is very efficient as each interval is a rectangle and its coordinates are directly compared against those of a set of other rectangles. An extract from the results of the method for the example of Fig. 2 can be seen in Fig. 3.

4. Discussion and conclusions

The main contribution of the new evaluation approach is its flexibility and efficiency. It does not require regions to be rectangular, enabling it to work with complex-shaped regions (possibly severely skewed) with no extra overhead. The efficiency of the approach is owed mainly to the region-representation scheme used, which approximates the efficiency of rectangle-based geometric comparison, without resorting to time-consuming pixel-based image accesses. Apart from enabling fast comparison of regions, the representation also requires considerably less memory than the pixel-based region-map approach [6].

One of the main reasons that geometric comparison is possible in this case because the ground-truth regions (GTPs) are described very accurately without any excess surrounding background space. This accurate description is achieved by using the results of the White Tiles page segmentation method [5] as the basis for deriving the ground truth. An additional benefit is that the creation of ground truth information requires less effort (to edit page segmentation results) than manually segmenting a page image with significant attention to detail (particularly true for complex layouts).

Using the global interval structure representation of regions it is possible to identify correspondence between ground truth and segmentation regions and perform a comparative analysis by simply checking for overlap of intervals. Furthermore, it is straightforward to calculate a meaningful measure of how well a region has been described or how much of it has been missed by the segmentation. Greater accuracy may also be achieved using the interval-based structure since regions can be described in full resolution, rather than the reduced resolution required by the pixel-based method (to reduce comparisons) [6].

Two approaches have been presented in this paper. The first uses information about the legal limits that a segmentation polygon can reach in order to correctly (even if not very accurately) describe a ground-truth region. This method is logically appealing and can be faster in determining whether there are any merges (the assessment utilises the fact that the area of the whole page is described by MDPs). However, this approach requires the extra (but not time-consuming) step of comparing SPs with MDPs (the effort of creating MDPs is not significant and takes place once, when GTPs are created). The second approach reverses the problem of judging how well a segmentation polygon describes regions to the problem of assessing how a ground-truth region has been described. This is arguably a more immediate approach as the bottom-line is the ground-truth. The second approach also requires fewer operations. A decision has therefore been made to use the second approach in the DIA performance analysis framework.

```
GTPs correctly identified:
    GTP1  found by  SP1
    GTP2  found by  SP2
    GTP7  found by  SP6
-----------------------
GTPs missed:
    GTP8
-----------------------
GTPs split and part/s missed:
    GTP6  found by:
                    SP5
                    SP4
-----------------------
SPs have merged GTPs:
    SP3 has merged:
      GTP3
      GTP4
    SP4 has merged:
      GTP5
      GTP6  ** Part Missed **
------------------------
SPs have un-associated GTPs:
    SP7
    SP8
-----------------------
```

Figure 3. Results extract for Fig. 2.

This paper has presented part of a new DIA performance analysis framework aimed at Layout Analysis algorithm developers. The methods have successfully analysed a large number of diverse test cases, including cases that may be rare in reality. Further validation with large numbers of real image data is in progress.

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] I.T. Philips, S. Chen, J. Ha and R.M. Haralick, "English Document Database Design and Implementation Methodology", *Proc. 2nd Annual Symp. on Document Analysis and Retrieval*, UNLV, USA, 1993, pp. 65–104.

[3] S. Nieminen, J. Sauvola, T. Seppänen and M. Pietikänen, "Benchmarking System for Document Analysis Algorithms", *Proc. SPIE Conf. on Document Recognition (V)*, San Jose, CA, USA, 1998, Vol. 3305, pp. 100–111.

[4] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated Evaluation of OCR Zoning", *IEEE Trans. Patt. Anal. Machine Intell.*, Vol. 17, No. 1, January, 1995, pp. 86-90.

[5] A. Antonacopoulos, "Page Segmentation Using the Description of the Background", *Computer Vision and Image Understanding*, Vol. 70, No. 3, 1998, pp. 350-369.

[6] 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.

[7] J. Kanai, "Automated Performance Evaluation of Document Image Analysis Systems: Issues and Practice", *Int. J. Imaging Sys. and Technol.*, Vol. 7, 1996, pp. 363–369.

[8] A. Antonacopoulos and R.T. Ritchings, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles", *Proc. 3rd Int. Conf. on Doc. Anal. and Rec. (ICDAR'95)*, Montreal, Canada, 1995, pp. 1132-1135.