A NEW FRAMEWORK FOR EFFICIENT AND FLEXIBLE ANALYSIS OF THE PERFORMANCE OF DOCUMENT IMAGE ANALYSIS SUBSYSTEMS

A Antonacopoulos[1] and A Brough

University of Liverpool, United Kingdom

INTRODUCTION

Algorithms in various fields of Image Analysis have matured over the years and yet more new algorithms are being developed 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, thus making a direct comparison with other algorithms difficult. The need for objective evaluation of the performance of Image Analysis algorithms is now widely acknowledged and a number of techniques have been devised for various subsystems, e.g., see Bower and Phillips (1).

In the field of Document Image Analysis (DIA), significant activity has concentrated on evaluating OCR results, e.g., see Nagy (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, e.g., see Philips et al (3).

Large-scale testing and evaluation is essential not only for OCR but for each of the subsystems involved in DIA also. For instance, 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 focuses on subsystems comprising the Layout Analysis stage. The most significant subsystems in this stage are page segmentation and classification.

It should be noted that there is an important distinction between *comparative benchmarking (evaluation)* and *performance analysis*. The latter is aimed primarily at algorithm developers and provides detailed qualitative and quantitative information on the performance of a method in a number of categories. In contrast, benchmarking provides a global score for a method or constituent components and is mainly aimed at end-users of algorithms.

The framework described in this paper is focused mainly on performance analysis. A scoring system is also used to provide developers with a higher-level view of the performance of a method in particular aspects. Furthermore, a global score can be easily produced for benchmarking purposes if required.

In the next section, existing approaches to performance evaluation, relevant to Layout Analysis, are presented. Following that, the proposed performance analysis framework is briefly described. Further details on the central issues of region representation and comparative analysis as well as an overview of the test database are given in subsequent sections. Finally, the paper ends with concluding remarks about the new framework.

PAST APPROACHES

Past approaches to the evaluation of page segmentation and classification methods fall into two broad categories. First, an evaluation system based on OCR results was proposed as a result of extensive experience in OCR evaluation at UNLV, e.g., see Kanai et al (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 researchers in DIA. Furthermore, there is not always a direct correspondence between segmentation performance and errors in the OCR result. Finally, this method does not deal with the non-textual entities on the page.

The second category of approaches comprises methods that compare *regions* resulting from page segmentation with the corresponding ground-truth description of the expected regions. One approach proposed by the University of Washington (3) can use bounding rectangles to describe and compare regions resulting from segmentation with ground-truth rectangles. To the best of the authors' knowledge, such an approach has not been implemented yet. Furthermore, the bounding rectangles provided in the database are mainly aimed at OCR as significant proportion of the information concentrates on the word level. As far as paragraphs and other regions are concerned and while many types of documents have rectangular regions, this approach is not applicable to methods dealing with complex layouts, e.g., see Antonacopoulos (5).

[1] Corresponding author (email: aa@csc.liv.ac.uk)

A more flexible approach that deals with non-rectangular regions has been developed at Xerox, e.g., see Yanikoglu and Vincent (6). This approach circumvents the problem of comparing regions when different 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. This can be an obstacle to large-scale evaluation. Furthermore, although halftones are taken into account there is no provision for other non-textual components on a page.

Finally, none of the above approaches provides for the evaluation of Logical Layout Analysis (functional labelling of regions as heading, body text, footnote etc.). This evaluation category is very useful for assessing methods for use in Indexing and Document Understanding applications.

THE PROPOSED FRAMEWORK

A new performance evaluation framework is under development at the University of Liverpool. It consists of a new test database and new performance analysis methods. A simplified diagram of the system is illustrated in Figure 1.

As mentioned earlier, the focus of the framework is on Layout Analysis subsystems. These include page segmentation, page classification and logical layout analysis. The analysis of the performance of skew detection and correction (pre-processing) methods can be easily added to the framework (having already the test data), however, this aspect is not within the scope of this paper.

The new framework has two aims. The main one is to provide a means for algorithm developers to analyse the performance of their methods using a wide variety of test and ground-truth (known to be correct) data and conditions. This analysis should provide sufficient detail to determine the strong points of a given method and highlight its weaknesses so that they can be improved upon.

The results of the analysis will be presented at different levels (test-set / page / region) and will take two forms:

- Qualitative. These will either indicate the correctness of a decision of the method under test or the existence of errors and their types.
- Quantitative. For each of the conditions reported above, a measure of success/failure is reported. These measures are based on region area ratios and are different from the penalties associated with errors (used in the calculation of global scores).

An innovative aspect of the framework, in consistence with the ethos of its main objective, is the provision of links for a developer to visually inspect the results of the method under test directly associated to instances of each error. This feature will facilitate the design of improvements.

A secondary aim of the framework is to provide end-users of Layout Analysis algorithms with a benchmark score corresponding to each method evaluated. Direct comparisons between methods are straightforward in this way. This is also a rapid way of determining the applicability of a method to a given class of documents.

The main benefits of this framework are efficiency (paramount for large-scale evaluation) and flexibility. Significant efficiency is gained from a description-based comparative analysis of regions, which avoids time-consuming pixel-level 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 were 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 provide information at various levels of detail. At the local level, a report is provided for each region where detailed information is available on a number of conditions (some only possible under the new framework) an overview of which is given below. At the global level, information is available for the performance of an algorithm on a whole page or set of pages. Finally, it is possible to select specific conditions (and combinations thereof) for an algorithm to be tested under.

The key parts of the framework are presented below.

Region Representation and Comparative Analysis

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 Antonacopoulos and Ritchings (7). Since the contour of each region can be described by an isothetic (having only horizontal and vertical edges) polygon (5), 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 (7). 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.

For the ground-truth description, each region is described by the closest-fitting isothetic polygon around the region, the *Ground-Truth Polygon (GTP)*. For page segmentation purposes, a region is defined to be the smallest logical entity on the page. For text, this

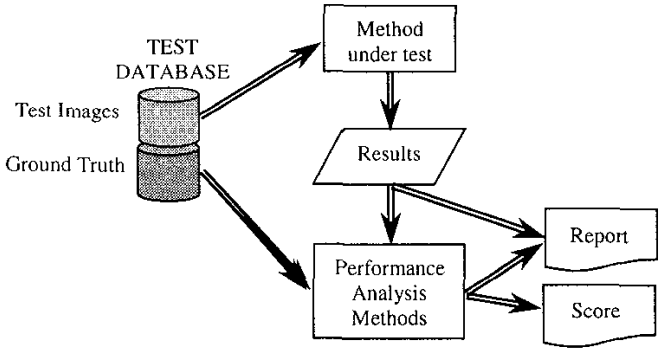

Figure1: A simplified view of the system

corresponds to a single paragraph (body text, header, footnote, caption etc.). A single graphic component on the page (*halftone, line-art, image, horizontal/vertical* ruling etc.) is also considered as a single region. The GTP is obtained by manually correcting segmentation results (5) (contour polygons).

The regions resulting from the application of a page segmentation algorithm are referred to as *Segmentation Polygons (SP)*. The objective of the comparative analysis is to identify a correspondence between SPs and GTPs and determine and report discrepancies in the correctness of description, type of content (page classification) and function (logical layout analysis).

Each region resulting from the page segmentation algorithm to be analysed is compared (having converted its representation to an interval structure, SP) to the ground-truth representation. The following situations may arise:

a) *the SP correctly describes a single GTP* (1:1 correspondence)
b) the SP does not cover the GTP completely (GTP region is split)
c) the SP covers the GTP of one region and part of the GTP of another (GTP regions are merged)
d) no SP describes a given GTP region (GTP region missed)
e) a SP does not cover any GTP (SP region wrongly detected).

In the first of the above cases (correct identification), the accuracy of the description is also measured based on the area of surrounding background space included in the SP. This is possible as the GTP is very tightly wrapped around each region. When part of a GTP is missed the severity of the error is measured based on the area not described by any SP.

Detailed reports at different levels of abstraction can be provided based on a number of performance metrics and combinations. Some of the primary metrics used are as follows:

1) number of correctly identified regions, as a percentage of the total number of regions (GTPs) on the document page
2) number of regions wholly missed, as a percentage of the total number of regions (GTPs) on the document page
3) number of regions partially missed, as a percentage of the total number of regions (GTPs) on the document page
4) number of non-existent regions (e.g. noise) that have been detected, as a percentage of the total number of regions (GTPs) on the document page
5) total area of correctly identified regions, as a percentage of the total area of regions (GTPs) on the document page
6) total area of wrongly detected (non-existent) and undetected regions, as a percentage of the total area of regions (GTPs) on the document page

The above metrics can be evaluated on a page level (as described above) but also on a class-level (e.g. in terms of the total text GTPs). For more detailed information on specific cases, area-based reports in addition to those from metrics 5 and 6 above can also be given on a region-level, as the proportion of the erroneous area against the area of the corresponding GTP. The erroneous area can be either the area of the part of the GTP that is not described by any SP or the area of the superfluous surrounding background space included in the description of a correctly identified GTP. Other region-based metrics are based on the number of SPs describing a split GTP and the number of GTPs merged by a single SP.

If the performance of page classification is also to be analysed, the type of the content of each region as determined by page classification is verified against the ground-truth information. Page classification metrics are based on the number (or area) of regions correctly classified, as a percentage of the total number (or area)

of regions in the same class or on the document page (or whole test-set).

Finally, a very useful and innovative indication of the robustness of a method is given by the *skew tolerance* metric. This is defined as the maximum angle of at which a region or page can be oriented and still be correctly identified (and classified).

For obtaining a benchmarking score, each of the segmentation outcomes (a to e, above) and combinations thereof are associated with detailed penalties depending on the both the topology and content type (class) of the regions (GTPs) involved. For instance, merging text lines horizontally across columns is more severe than merging paragraphs or splitting vertically adjoining lines in the same column of text. If necessary, these penalties can be modified for different performance analysis requirements.

The Test-Image Database

For each document page the ground-truth database holds the image, general image and document attributes, GTP interval structure data and individual region attributes. Although at the moment the performance analysis framework is targeted towards DIA subsystems before OCR and graphics recognition, suitable ground truth data for the content of regions can be added.

When complete, the database will contain test data for the following types of documents: articles (journals, proceedings, books), newspapers, magazines, business letters, memorandums, facsimile documents and advertisements. Additional types that can be included are maps, forms, engineering drawings and handwritten documents.

The significant improvement of the database in relation to previous approaches is the inclusion of pages with complex layouts (in terms of shape of regions, colour and orientation) and with different types of non-textual entities.

CONCLUSIONS

This paper has presented a brief description of a new framework for the analysis of the performance of algorithms used in layout analysis, mainly page segmentation and classification. This framework is primarily aimed at algorithm developers and provides rich information at local and global levels. An overall score is also calculated for each method for benchmarking purposes (for end-users). Preliminary results show that flexible and efficient description-based analysis of segmentation results is possible and comprehensive information (including information not available before) is readily available using new analysis methods. Work continues on database improvements and consideration of various testing suites appropriate to different evaluation scenarios.

REFERENCES

1. Bowyer K W and Phillips P J (Eds.), 1998, "Empirical Evaluation Techniques in Computer Vision", IEEE Computer Society Press, USA.

2. Nagy G, 1995, "Document Image Analysis: Automated Performance Evaluation", in Spitz A L and Dengel A (Eds.), "Document Image Analysis Systems", World Scientific Publishing Co, Singapore, 137-156.

3. Philips I T, Chen S and Haralick R M, 1993, "CD-ROM Document Database Standard", Proc 2nd Int Conf on Document Analysis and Recognition (ICDAR'93), Tsukuba, Japan, 478-483.

4. Kanai J, Rice S V, Nartker T A and Nagy G, 1995, "Automated Evaluation of OCR Zoning", IEEE Trans on PAMI, 17, 86-90.

5. Antonacopoulos A, 1998, "Page Segmentation Using the Description of the Background", Computer Vision and Image Understanding, 70, 350-369.

6. Yanikoglu B A and Vincent L, 1998, "Pink Panther: A Complete Environment for Ground-Truthing and Benchmarking Document Page Segmentation", Pattern Recognition, 31, 1191-1204.

7. Antonacopoulos A and Ritchings R T, 1995, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles", Proc 3rd Int Conf on Document Analysis and Recognition (ICDAR'95), Montreal, Canada, 1132-1135.