Performance Analysis Framework for Layout Analysis Methods

A. Antonacopoulos and D. Bridson

PRImA Lab, School of Computing, Science and Engineering, University of Salford,
Greater Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

Abstract

This paper presents a new framework for in-depth analysis of the performance of layout analysis methods. Contrary to existing approaches aimed at evaluation or benchmarking, the proposed framework provides detailed information at various levels that can be used by method developers to identify specific problems and improve their work. Complex layouts are supported as well as the flexible configuration of goal-oriented performance analysis scenarios. The comparison of segmentation results against the ground truth is performed in a very efficient way based on a decomposition of any region shape into an interval-based description. The framework has been validated using the dataset and method results of the ICDAR2005 Page Segmentation Competition.

1 Introduction

Layout Analysis is central to most Document Image Analysis systems and applications. It comprises Page Segmentation (identification of regions of interest), Region Classification (identification of the type of content of each region) and further processes such as Logical Labelling (labelling of regions in terms of their function) and reading order determination.

A considerable amount of effort has been devoted over the past two decades to develop various layout analysis methods (page segmentation, in particular) and new methods continue to be reported in the literature. Most methods were primarily aimed at specific applications and consequently were based on specific assumptions about their target document classes (e.g. text blocks are expected to be rectangular). Typically, each method was evaluated on relatively narrowly-focused application-specific datasets, which more often than not do not reflect the real-world occurrence of documents.

The need for objective and realistic *evaluation* of layout analysis methods is more pressing than ever, as evidenced by the various evaluation approaches proposed so far and the inception of ICDAR competitions in the area [1][2][3].

Past approaches have focused on calculating various error metrics in order to *quantify* the performance of page segmentation methods, mostly for benchmarking or comparative evaluation. Early approaches [4] considered the recognised text inside each region and the corresponding number of edit operations necessary to correct errors. However, such a metric cannot give an accurate indication of page segmentation performance since a number of errors in the text are also due to OCR processes [5][6].

Later approaches focus on calculating discrepancies between ground truth and segmentation *region* characteristics. Such methods can be divided in two main categories: those that examine *geometric* correspondences of regions and those that perform *pixel* comparisons between regions. In almost all methods in the former category [6][7][8], regions (characters, textlines or paragraphs) are described by bounding boxes. Comparisons are efficient and corresponding ground truth straightforward to produce. However, a significant disadvantage is that documents with complex-shaped regions cannot be handled by such approaches although some early ideas of addressing this issue were explored [5][9].

Pixel-based region comparison approaches [10][1][2][3][11] on the other hand are very accurate and can work with complex-shaped regions. However, ground truth creation for such approaches can be more cumbersome [12] and it takes up a lot more storage. Furthermore, pixel-based comparison is much less efficient than geometric comparison.

In addition to the benchmarking goals of past approaches, there is also need for detailed *performance analysis* for each method. Such analysis extends beyond a set of simple scores for each method based on cumulative errors over a whole dataset. While evaluation and benchmarking are useful for a performance overview and direct comparison of methods they do not provide sufficient information for researchers and developers. For them, it is necessary to provide both a more detailed quantitative and a qualitative account of errors. As errors have different significance in different contexts, it is necessary to take this into account during evaluation so that developers may receive the in-depth information necessary to improve their methods.

The proposed framework is designed to provide in-depth information at various levels (dataset/page/region) to assist with method development in addition to goal-oriented performance evaluation and characterisation based on different user-defined scenarios. The correspondences between ground truth and segmentation regions are identified through geometric comparisons of regions represented as polygons achieving, thus, both accuracy in dealing with complex-shaped regions and efficiency (similar to bounding box comparison)

The framework is briefly described in the next section. An overview of ground truth requirements and related issues is given in Section 3. In Section 4, the performance analysis method is presented, with region representation, region correspondence determination and error qualification/quantification explained in separate subsections. The presentation of the analysis results is described in Section 5, while Section 6 discusses the proposed approach and concludes the paper.

2 Framework overview

The proposed performance analysis framework comprises two main components. First, a user interface through which batches of ground truth and segmentation results are selected, evaluation scenarios defined and interactive presentation of performance analysis results takes place.

Second, the performance analysis system itself which performs the following steps:

1. *Region representation:* Ground truth and segmentation regions are transformed into an interval-based representation.
2. *Region correspondence determination:* Using the interval-based representation, correspondence between parts of ground truth, segmentation and background regions is established.
3. *Error qualification and quantification:* Errors in correspondence between ground truth and segmentation regions are examined in the context of application scenario and their significance is established.

3 Ground truth

To take advantage of the full power of the framework there must be suitable ground truth with enough information about the regions and a sufficiently flexible description of the region outlines.

In developing the method, we have used the dataset which was also used for the ICDAR2005 Page Segmentation Competition [3]. Its ground truth contains rich information about the content and function of each region as well as about the corresponding page and document [13]. Regions are described in terms of isothetic polygons (polygons having horizontal and vertical edges only).

4 Performance analysis

This is the most important framework component both in terms of technical issues and in terms of achieving the resulting information richness and accuracy.

The key challenge is the effective and efficient analysis and identification of correspondence of polygons instead of bounding boxes or pixel representations of regions.

Each of the steps in the process is described below.

4.1 Region representation

Region representation is key to both efficiency and accuracy of performance analysis. The proposed approach accepts both segmentation results and ground truth regions having practically any shape. However, it should be noted that, as printed regions on documents are mostly polygonal in shape with many of their edges being horizontal or vertical, it is naturally more efficient to represent them as isothetic polygons wherever possible.

Given a set of region contours (segmentation or ground truth), the first step is to create a representation of them in terms of *intervals*. An interval is defined as a maximal rectangle that can be fitted horizontally inside a region (starting at a given point on a vertical edge), spanning the whole width of the region [14]. This process can be thought of as a decomposition of a shape into a set of vertically adjacent horizontally-oriented rectangles. A simple decomposition of a region along these lines is illustrated in Fig 1(a).

The polygons of less complex regions will, more often than not, be decomposed into a set of taller intervals than more complex-shaped regions. In the representation of more complex shapes, certain intervals may be collapsed to horizontal lines. In the simple case of regions represented by bounding boxes (in Manhattan layouts, for instance) a single region will consist of a single interval.

Given a whole document page, the interval representation takes into account the existence of more than one region in the horizontal direction. Intervals are therefore fitted across regions as shown in the simplified (for clarity) example of Fig. 1(b).

For each document page in a dataset, the interval representation of the ground truth regions can be created in advance. The corresponding segmentation result regions are then also represented in a similar interval structure. The two interval structures are subsequently merged to form a *combined interval representation.* It is that representation which is used to determine the correspondence between ground truth and segmentation regions. A simplified example of this representation is given in Fig. 2.

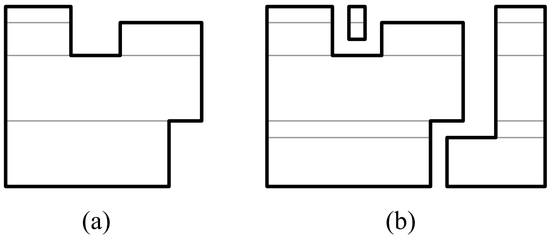


Figure 1. Interval representation of (a) a single region and (b) multiple regions.

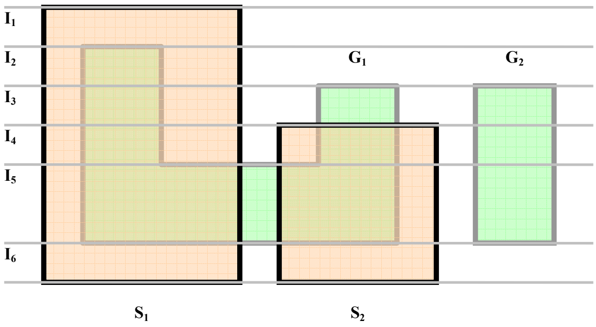


Figure 2. Combined (segmentation and ground truth interval representation.

4.2 Region correspondence determination

Within the combined interval representation, each interval line is examined in turn and overlaps are detected between:

a. Segmentation interval and nothing (see interval in I_1 line in Fig. 2)
b. Segmentation interval and ground truth interval (see interval in I_2 line in Fig. 2)
c. Ground truth interval and nothing (see last two intervals in I_3 line in Fig. 2)

Keeping track of the overlaps detected (as above) for all intervals of a given region it is straightforward to identify the following conditions for each region:

1. A segmentation region that has no overlap with any ground truth region (wrongly detected region)
2. A ground truth region that has been completely overlapped by a segmentation region (correctly detected region)
3. A ground truth region that has been overlapped – completely or partially – by more than one segmentation region (split region)
4. More than one ground truth region has been overlapped – completely or partially – by a single segmentation region (merged regions).
5. A ground truth region that has not been completely overlapped by any number of segmentation regions (partially missed region)
6. A ground truth region that has not been overlapped by any segmentation region (completely missed region)

The actual area of the overlap between individual intervals is calculated when overlaps are detected. Therefore, for each region the total area of overlap with other region(s) is recorded.

4.3 Error qualification and quantification

The degree of success of a layout analysis method directly depends on the *type* as well as on the *quantity* of errors it makes. In terms of page segmentation, the five types of error (as listed above) have different significance depending on

- context (within the document)
- application scenario (user defined)

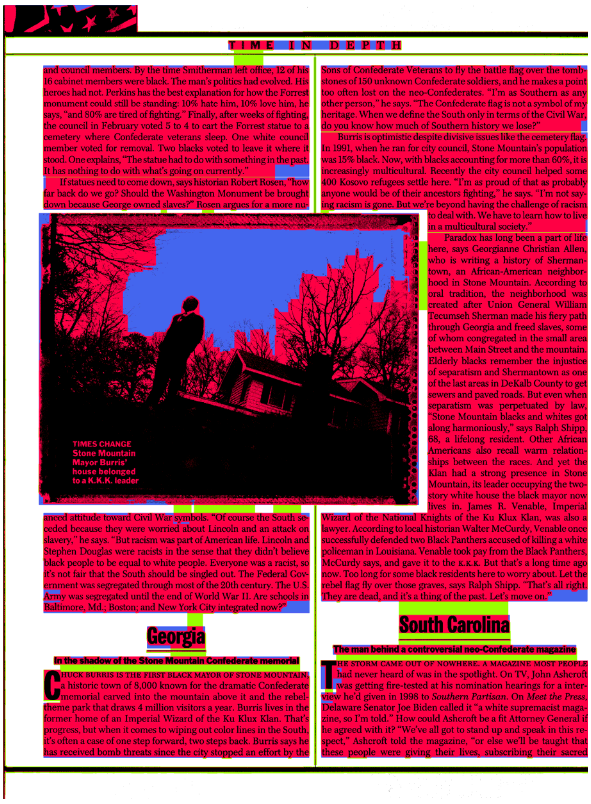
Error significance according to context is in most cases independent of the type of document. Examples include:

- A merger between two adjacent paragraphs within a single column of text is insignificant
- A merger between a paragraph of body text and a figure caption is a significant error
- A merger between two paragraphs across different columns is a significant error
- A merger between a text paragraph and a graphical region is a significant error

Error significance according to application scenario supplements the above, allowing a user to further tailor the performance analysis process. Examples of situations include:

- A merger between two graphical regions may not be significant in an OCR application.
- A merger between a section heading and a body text paragraph may not be significant in a general text processing application but may be significant if a table of contents needs to be constructed using section headings.

The significance of both context and application scenario is expressed by corresponding weights.



TIME IN DEPTH

and council members. By the time Smitherman left office, 12 of his 16 cabinet members were black. The man's politics had evolved. His heroes had not. Perkins has the best explanation for how the Forrest monument could still be standing: 10% hate him, 10% love him, he says, "and 80% are tired of fighting." Finally, after weeks of fighting, the council in February voted 5 to 4 to cart the Forrest statue to a cemetery where Confederate veterans sleep. One white council member voted for removal. Two blacks voted to leave it where it stood. One explains, "The statue had to do with something in the past. It has nothing to do with what's going on currently."

If statues need to come down, says historian Robert Rosen, "how far back do we go? Should the Washington Monument be brought down because George owned slaves?" Rosen argues for a more nuanced attitude toward Civil War symbols. "Of course the South seceded because they were worried about Lincoln and an attack on slavery," he says. "But racism was part of American life. Lincoln and Stephen Douglas were racists in the sense that they didn't believe black people to be equal to white people. Everyone was a racist, so it's not fair that the South should be singled out. The Federal Government was segregated through most of the 20th century. The U.S. Army was segregated until the end of World War II. Are schools in Baltimore, Md.; Boston; and New York City integrated now?"

Sons of Confederate Veterans to fly the battle flag over the tombstones of 150 unknown Confederate soldiers, and he makes a point too often lost on the neo-Confederates. "I'm as Southern as any other person," he says. "The Confederate flag is not a symbol of my heritage. When we define the South only in terms of the Civil War, do you know how much of Southern history we lose?"

Burris is optimistic despite divisive issues like the cemetery flag. In 1991, when he ran for city council, Stone Mountain's population was 15% black. Now, with blacks accounting for more than 60%, it is increasingly multicultural. Recently the city council helped some 400 Kosovo refugees settle here. "I'm as proud of that as probably anyone would be of their ancestors fighting," he says. "I'm not saying racism is gone. But we're beyond having the challenge of racism to deal with. We have to learn how to live in a multicultural society."

Paradox has long been a part of life here, says Georgianne Christian Allen, who is writing a history of Shermantown, an African-American neighborhood in Stone Mountain. According to oral tradition, the neighborhood was created after Union General William Tecumseh Sherman made his fiery path through Georgia and freed slaves, some of whom congregated in the small area between Main Street and the mountain. Elderly blacks remember the injustice of separatism and Shermantown as one of the last areas in DeKalb County to get sewers and paved roads. But even when separatism was perpetuated by law, "Stone Mountain blacks and whites got along harmoniously," says Ralph Shipp, 68, a lifelong resident. Other African Americans also recall warm relationships between the races. And yet the Klan had a strong presence in Stone Mountain, its leader occupying the two-story white house the black mayor now lives in. James R. Venable, Imperial Wizard of the National Knights of the Ku Klux Klan, was also a lawyer. According to local historian Walter McCurdy, Venable once successfully defended two Black Panthers accused of killing a white policeman in Louisiana. Venable took pay from the Black Panthers, McCurdy says, and gave it to the K.K.K. But that's a long time ago now. Too long for some black residents here to worry about. Let the rebel flag fly over those graves, says Ralph Shipp. "That's all right. They are dead, and it's a thing of the past. Let's move on."

TIMES CHANGE
Stone Mountain
Mayor Burris'
house belonged
to a K.K.K. leader

Georgia

In the shadow of the Stone Mountain Confederate memorial

CHUCK BURRIS IS THE FIRST BLACK MAYOR OF STONE MOUNTAIN, a historic town of 8,000 known for the dramatic Confederate memorial carved into the mountain above it and the rebel-theme park that draws 4 million visitors a year. Burris lives in the former home of an Imperial Wizard of the Ku Klux Klan. That's progress, but when it comes to wiping out color lines in the South, it's often a case of one step forward, two steps back. Burris says he has received bomb threats since the city stopped an effort by the

South Carolina

The man behind a controversial neo-Confederate magazine

THE STORM CAME OUT OF NOWHERE. A MAGAZINE MOST PEOPLE had never heard of was in the spotlight. On TV, John Ashcroft was getting fire-tested at his nomination hearings for a interview he'd given in 1998 to *Southern Partisan*. On *Meet the Press*, Delaware Senator Joe Biden called it "a white supremacist magazine, so I'm told." How could Ashcroft be a fit Attorney General if he agreed with it? "We've all got to stand up and speak in this respect," Ashcroft told the magazine, "or else we'll be taught that these people were giving their lives, subscribing their sacred

Figure 3. An example of visual presentation of results at the page level. Ground truth is in medium-dark (blue) colour while segmentation regions are in lighter (light green) colour. Overlapping regions are in darker (red) colour. Split and merged regions can be seen at a glance.

The proposed approach records each individual error, its context and the general application scenario. Based on this information, it also uses the information on the area of overlap between regions to assess and quantify the severity of the error.

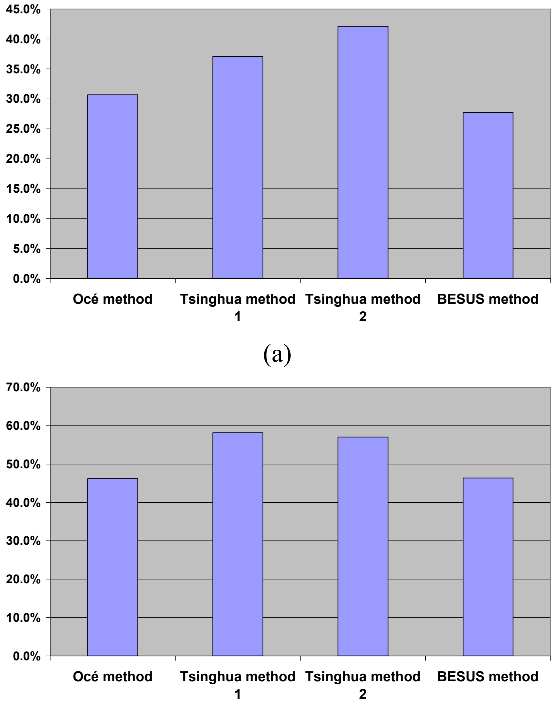
5 Presentation of analysis results

The above performance analysis gives rise to a considerable amount of information from overall task performance down to details of individual errors.

Information is available at dataset, page and region levels. Information is available by region type or error type.

A developer, for instance, can order results by error significance and individual errors can be displayed superimposed on the original page image (see Fig. 3).

A system integrator looking to choose between methods can specify a suitable scenario and a set of scores can be produced to provide a summary of the performance of each method for direct comparison.



(a)

(b)

Figure 4. (a) Results from the ICDAR2005 Page Segmentation Competition, and (b) from the proposed approach.

6 Discussion and conclusions

The new performance analysis method was compared against the published evaluation process of the ICDAR2005 Page Segmentation Competition [3]. Both the competition dataset and the results reported by the individual segmentation methods that took part were used in evaluating the system.

A graph of the overall competition results of the four different segmentation methods is shown in Fig. 4(a). The corresponding graph using the proposed approach is shown in Fig. 4(b).

Overall, the results broadly agree. Detailed results on different types of regions (not shown here) indicate that the main difference between the second and the third candidate methods (variant methods from the same research group) is due to slightly different weighting in application scenario (the ICDAR2005 seems to have been heavily weighted towards the detection of text). The use of the proposed framework has provided detailed information in order to better understand this situation and to suggest a more balanced scenario for future competitions.

In addition to the page segmentation results discussed above, it is of course straightforward to also analyse the performance of region classification and logical layout analysis. As long as suitable information (region type and functional labels) exists in the ground truth it can be utilised. In fact, as evident from above, such information is necessary in order to take full advantage of the error qualification and quantification process of the framework.

Concluding, a new performance evaluation framework has been presented. Its novelty lies in two main directions. First, it provides considerably more in-depth information which is useful for developers (as opposed to evaluation or benchmarking only). It also enables goal-oriented performance analysis through a detailed error qualification and quantification scheme. Second, it is efficient and accurate using an interval-based region representation to establish correspondence between ground truth and segmentation regions. This representation closely approaches the efficiency of rectangular representation schemes but with the advantage that it supports the accurate handling of layouts with complex-shaped regions.

Further work continues towards building an on-line system (web service) which will enable researchers to use the framework as a web service.

References

[1] B. Gatos, S.L. Mantzaris and A. Antonacopoulos, "First International Newspaper Page Segmentation Competition", *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR2001)*, Seattle, USA, September 10–13, 2001, pp. 1190–1194.

[2] A. Antonacopoulos, B. Gatos and D. Karatzas, "ICDAR2003 Page Segmentation Competition", *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR2003)*, Edinburgh, UK, August 3–6, 2003, pp. 688–692.

[3] A. Antonacopoulos, B. Gatos and D. Bridson, "ICDAR2005 Page Segmentation Competition", *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005)*, Seoul, South Korea, August 29–September 1, 2005, pp. 75–79.

[4] J. Kanai, S.V. Rice, T.A. Nartker and G. Nagy, "Automated evaluation of OCR zoning" *IEEE Transactions on Pattern Analysis and Machine Intelligences*, Vol. 17 (1995), pp. 86–90.

[5] A. Antonacopoulos and A. Brough, "Methodology for Flexible and Efficient Analysis of the Performance of Page Segmentation Algorithms", *Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR1999)*, Bangalore, India, September 20–22, 1999, pp. 451–454.

[6] M. Thulke, V. Märgner and A. Dengel, "A General Approach to Quality Evaluation of Document Segmentation Results", *Proceedings of the 3rd IAPR Workshop on Document Analysis Systems (DAS98)*, Nagano, Japan, November 4–6, 1998, Springer LNCS (1655), pp 43–57.

[7] S. Mao and T. Kanungo, "Software Architecture of PSET: A Page Segmentation Evaluation Toolkit" *International Journal of Document Analysis and Recognition*, Vol. 4 (2002), pp. 205–217.

[8] A.K. Das, S.K. Saha and B. Chanda, "An empirical measure of the performance of a document image segmentation algorithm" *International Journal of Document Analysis and Recognition*, Vol. 4 (2002), pp. 183–190.

[9] A. Antonacopoulos, F. Coenen, "Region Description and Comparative Analysis using a Tesseral Representation", *Proceedings of the 5th International Conference on Document Analysis and Recognition (ICDAR1999)*, Bangalore, India, September 20–22, 1999, pp. 193–196.

[10] B. Yanikoglu and L. Vincent, "Pink panther: a complete environment for ground-truthing and benchmarking document page segmentation" *Pattern Recognition*, Vol. 31 (1998), pp. 1191–1204.

[11] F. Shafait, D. Keysers and T.M. Breuel, "Pixel-Accurate Representation and Evaluation of Page Segmentation in Document Images", *Proceedings of the 18th International Conference on Pattern Recognition (ICPR2006)*, Hong Kong, China, August 20–24, 2006, pp. 872–875.

[12] J. Kanai, "Automated Performance Evaluation of Document Image-Analysis Systems: Issues and Practice" *International Journal of Imaging Systems and Technology*, Vol. 7 (1996), pp. 363–369.

[13] A. Antonacopoulos, D. Karatzas and D. Bridson, "Ground Truth for Layout Analysis Performance Evaluation", Proceedings of the 7th IAPR Workshop on Document Analysis Systems (DAS2006), Nelson, New Zealand, February 13–15, 2006, Springer LNCS (3872), pp 302–311.

[14] A. Antonacopoulos and R.T. Ritchings, "Representation and Classification of Complex-Shaped Printed Regions Using White Tiles", *Proceedings of the 3rd International Conference on Document Analysis and Recognition (ICDAR95)*, Montreal, Canada, August 14–15, 1995, pp. 1132–1135.