ICDAR2005 Page Segmentation Competition

A. Antonacopoulos[1], B. Gatos[2] and D. Bridson[1]

[1]*Pattern Recognition and Image Analysis (PRImA) Research Lab*
School of Computing, Science and Engineering, University of Salford, Manchester, M5 4WT, United Kingdom
http://www.primaresearch.org

[2]*Computational Intelligence Laboratory, Institute of Informatics and Telecommunications,*
National Center for Scientific Research "Demokritos", GR-153 10 Agia Paraskevi, Athens, Greece
http://www.iit.demokritos.gr/cil

Abstract

There is an established need for objective evaluation of layout analysis methods, in realistic circumstances. This paper describes the Page Segmentation Competition (modus operandi, dataset and evaluation criteria) held in the context of ICDAR2005 and presents the results of the evaluation of four candidate methods. The main objective of the competition was to compare the performance of such methods using scanned documents from commonly-occurring publications. The results indicate that although methods seem to be maturing, there is still a considerable need to develop robust methods that deal with everyday documents.

1 Introduction

Layout analysis methods—page segmentation in particular—continue to be reported in the literature on a frequent basis, despite this being one of the most mature sub fields of Document Image Analysis. It is not difficult to see that the reason for this is that the problem is far from being solved. Successful methods have certainly been reported but, frequently, those are devised with a specific application in mind and are fine-tuned to the test image data set used by its authors. The wider gamut of documents encountered in real-life situations is far wider than the target applications of most methods.

There is no doubt that, for a given application, or for a generic selection of real-life documents, it would be desirable to obtain an objective evaluation of the performance of different layout analysis methods. Such a direct comparison between algorithms is not straightforward as it requires both the creation of suitable ground truth (a relatively laborious and precise task) as well as the definition of a set of objective evaluation criteria (and a method to analyse them).

This competition focuses on the evaluation of page segmentation and region classification subsystems. To the best of the Authors' knowledge, this is only the second instance of an international generic layout analysis competition (the first being the ICDAR2003 Page Segmentation Competition [1]). It should be mentioned that a relatively close previous instance, focussing on a specific application domain, was the First International Newspaper Page Segmentation Contest [2] held by the Authors in the context of ICDAR2001. Prior to that, an evaluation of page segmentation (as part of OCR systems) was performed at UNLV [3], based on the results of OCR. That approach, however, cannot not be strictly considered to evaluate layout analysis methods since the OCR-based evaluation does not give sufficient information on the performance of page segmentation and region classification and is only applicable to regions of text (or text-only documents).

The motivation for this competition was the evaluation of page segmentation and region classification methods in *realistic* circumstances. By realistic it is meant that the participating methods are applied to scanned documents from a variety of sources, occurring in real life. This is in contrast to the majority of datasets and reports of results using mostly structured documents (e.g., technical articles).

The competition and its modus operandi is described next. In Section 3, an overview of the dataset and the ground-truthing process is given. The performance evaluation method and metrics are described in Section 4, while each of the participating methods is summarised in Section 5. Finally, the results of the competition are presented and the paper is concluded in Sections 6 and 7, respectively.

2 The competition

The objective of the competition was to evaluate layout analysis (page segmentation and region classification) methods using scanned documents from commonly-occurring publications. While there is a comparative assessment element involved, the real advantage is an initial look in the performance of different classes of methods (e.g., connected component analysis, morphological processing, analysis of background etc.) in identifying different types of regions in a variety of documents.

Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR'05)
1520-5263/05 $20.00 © 2005 IEEE

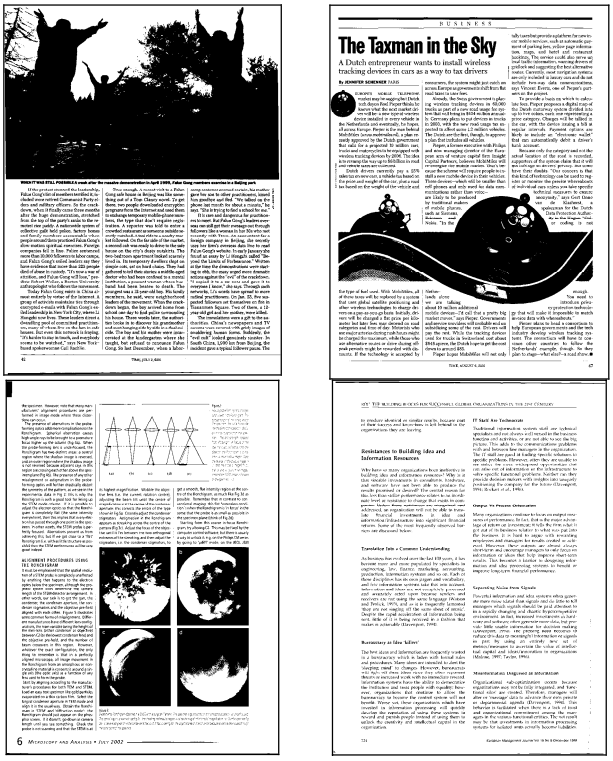

Figure 1. Sample page images from the training dataset.

The competition run in an off-line mode. The authors of candidate methods registered their interest in the competition and downloaded the *training* dataset (document *images* and associated *groundtruth*). One week before the competition closing date, registered authors of candidate methods were able to download the document *images* of the *evaluation* dataset. At the closing date, the organisers received the results of the candidate methods, submitted by their authors in a pre-defined format. The organisers then evaluated the submitted results.

It should be noted that the off-line mode is based on trust that the results submitted by the methods' authors are genuine. This trust is even more necessary if the evaluation system is publicly available. In this case, the evaluation system was not published (only the principles) and above all, the organisers have faith in the authors' scientific integrity.

3 The dataset

For any performance evaluation approach, the Achilles' heel is the availability of realistic and accurate ground truth. As ground-truthing cannot (by definition) be fully automated, it remains a laborious and, therefore, expensive process. One approach would be to use synthetic data [4]. It is the authors' opinion, however, that for the realistic evaluation of layout analysis methods, 'real' scanned documents give a better insight.

It should be noted that ground truth there is scarce availability of ground truth for the evaluation of methods analysing complex layouts (e.g., having non-rectangular regions). Such a dataset was created for the ICDAR2003 competition [1]. However, the current competition was based on a subset of a significantly updated dataset. This dataset, which will shortly be released by the PRImA research lab, contains richer ground truth (in a correspondingly updated XML format) that provides a very wide range of information on region attributes (physical and logical).

Although the dataset contains instances (images and ground truth) of an exhaustive list of document types it does focus, however, (for meaningful evaluation purposes) on the most heavily used (in terms of information content and need to analyse) types of documents, such as office documents, magazine pages, advertisements and technical articles.

For the competition, a subset of documents was selected that reflected both realism in their frequent occurrence and, at the same time, the existence of sufficiently general interest to analyse them.

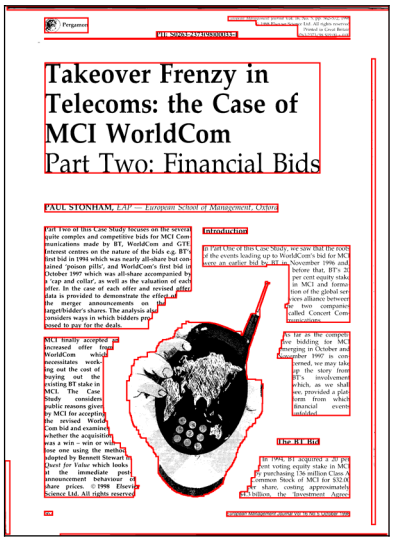

Figure 2. Sample page image from the training dataset showing superimposed description of region contours.

Furthermore, a balance had to be achieved between logistics (a manageable number of document images) and tractability for current methods. The decision was, therefore, made to focus on a cross section of 26 page images, comprising 30% technical articles (not necessarily with Manhattan layouts) and 70% magazine

pages. It should be noted that also for reasons of tractability, the competition images were bilevel (in the general dataset the original images are in colour). A sample of page images given as part of the *training* dataset can be seen in Fig. 1.

The ground-truth of each page image is an XML file (defined as part of the general dataset) that contains image and layout-specific information as well as the description of the regions in terms of isothetic (having only horizontal and vertical edges) polygons. The ground-truth for the competition was produced using a semi-automated tool developed by the authors. An XML viewer was developed for examining the images and the corresponding ground-truth XML, and was distributed to the competition participants. Another sample page image with the corresponding description of regions superimposed as isothetic polygons can be seen in Fig. 2.

The types of regions defined for the competition (simplified from the total number of different types in the general dataset) are:

- text,
- graphics,
- line-art,
- separator, and
- noise.

4 Performance evaluation

The performance evaluation method used is based on counting the number of matches between the entities detected by the algorithm and the entities in the ground truth [5-7]. We use a global MatchScore table for all entities whose values are calculated according to the intersection of the ON pixel sets of the result and the ground truth (a similar technique is used at [8]).

Let I be the set of all image points, G_j the set of all points inside the j ground truth region, R_i the set of all points inside the i result region, g_j the entity of j ground truth, r_i the entity of i result, T(s) a function that counts the elements of set s. Table MatchScore(i,j) represents the matching results of the j ground truth region and the i result region. Based on a pixel based approach of [5], and using a global MatchScore table for all entities, we can define that:

$$\text{MatchScore}(i,j) = a\frac{T(G_j \cap R_j \cap I)}{T((G_j \cup R_j) \cap I)}, \text{where } a = \begin{cases} 1, \text{if } g_j = r_i \\ 0, \text{otherwise} \end{cases} \quad (1)$$

If N_i is the count of ground-truth elements belonging to entity i, M_i is the count of result elements belonging to entity i, and w_1, w_2, w_3, w_4, w_5, w_6 are pre-determined weights, we can calculate the detection rate and recognition accuracy for i entity as follows:

$$\text{DetectRate}_i = w_1\frac{\text{one2one}_i}{N_i} + w_2\frac{\text{g\_one2many}_i}{N_i} + w_3\frac{\text{g\_many2one}_i}{N_i} \quad (2)$$

$$\text{RecognAccuracy}_i = w_4\frac{\text{one2one}_i}{M_i} + w_5\frac{\text{d\_one2many}_i}{M_i} + w_6\frac{\text{d\_many2one}_i}{M_i} \quad (3)$$

where the entities one2one_i, g_one2many_i, g_many2one_i, d_one2many_i and d_many2one_i are calculated from MatchScore table (1) following the steps of [5] for every entity i.

A performance metric for detecting each entity can be extracted if we combine the values of the entity's detection rate and recognition accuracy. We can define the following Entity Detection Metric (EDM_i):

$$EDM_i = \frac{2\text{DetectRate}_i\text{RecognAccuracy}_i}{\text{DetectRate}_i + \text{RecognAccuracy}_i} \quad (4)$$

A global performance metric for detecting all entities can be extracted if we combine all values of detection rate and recognition accuracy. If I is the total number of entities and N_i is the count of ground-truth elements belonging to entity i, then by using the weighted average for all EDM_i values we can define the following Segmentation Metric (SM):

$$SM = \frac{\sum_I N_i EDM_i}{\sum_I N_i} \quad (5)$$

5 Participating methods

Brief descriptions of the methods whose results were submitted to the competition are given next. Each account has been provided by the method's authors and edited (summarised) by the competition organisers. The descriptions vary in length according to the level of detail in the source information provided.

5.1 The BESUS method

This method—BESUS stands for Bengal Engineering and Science University, Shibpur (India)—was submitted by S.P. Chowdhury, S. Mandal and A.K. Das (of that university) in association with B. Chanda of the Indian Statistical Institute (ISI) in Calcutta. Similarly to the method submitted by the authors to the ICDAR2003 Page Segmentation Competition [1], this is a system constructed using a number of morphology-based modules.

In a pre-processing step that information is gathered and skew is corrected. Horizontal and vertical separators are extracted next by opening the bilevel image with a

horizontal or vertical (respectively) structuring element and connected component analysis [9]. Text is segmented based on the spatial relationship between pairs of textlines (identified based on the similarity and distribution of connected components) [10]. Graphics regions are extracted from a greyscale image (created from the original bilevel one) based on the analysis of a co-occurrence matrix in relation to the result of opening and closing operations on the whole image [11]. Line art regions (components) are identified based on topological features and a density ratio. Remaining regions are classified as noise.

5.2 The Océ method

This method was submitted by M. Bilderbeek, Z. Goey and R. Audenaerde of Océ Technologies B.V. in the Netherlands. It is a variant of the winning method of the ICDAR2003 Page Segmentation Competition [1]. Its working principles are as follows.

Connected components are identified in the image (after removing a 25-pixel wide border) and classified into small character, normal character, large character, photograph, graphic, vertical line, horizontal line or noise (in terms of the region types used in the competition, photographs are graphics, lines are separators and graphics are line-art) using a manually constructed decision tree based on features such as width, height, number of pixels etc. Using the result of this classification four images are split off:

(a) an image containing photos and noise,
(b) an image containing graphics,
(c) an image containing lines, and
(d) an image containing text.

In the last case, those blocks, in which the majority of connected components are classified as large characters are split off to a separate image. Thus, the image containing text is divided into two images:

(d1) an image containing normal/small text, and
(d2) an image containing headers.

Next, the components in the normal/small text image (d1), in the photo/noise image (a) and in the graphics image (b) are joined into blocks using a run length smearing procedure.

The resulting blocks are then classified by a voting algorithm that takes the connected component class statistics as its input. In the line image (c), each line is considered as a separate block with class label 'separator'. The blocks in the header image (d2) are identified by applying a connected component grouping algorithm, which also applies a post-classification step to assure that the blocks really contain text.

A boundary tracking algorithm [12] is used to trace the outer contours of all blocks (originally represented as rectangles) in the smeared images and represent them as polygons. Finally, a cleaning step removes all polygons that are contained within others, (re)labels all very small polygons as noise and merges polygons that overlap to a certain extent.

5.3 The Tsinghua methods

Di Wen and Ming Chen, of Tsinghua University (State Key Laboratory of Intelligent Technology and Systems), in China submitted two different methods.

The first one (referred to as "Tsinghua method 1" here) is a bottom-up approach that works by progressively merging primitives at different levels (starting from connected components and resulting in text paragraphs etc.) based on the calculation of a quantitative measure (the Multi-Level Confidence – MLC value). This method has been reported in [13] and is adapted to English layouts for this competition. The output of this method is bounding rectangles only (a region may appear split as a result or bounding rectangles may overlap for different regions)

The second method ("Tsinghua method 2") is devised to deal better with irregular regions. It starts with the output of method 1 and text regions are separated from non-text ones. Text regions are identified as isothetic polygons based on a background analysis algorithm similar to [14] but working with connected components. Other types of regions are output as rectangles exactly as in method 1.

6 Results

We evaluated the performance of the 4 segmentation algorithms using equations (1)–(5) for all 26 test images with parameters $w_1 = 1$, $w_2 = 0.75$, $w_3 = 0.75$, $w_4 = 1$, $w_5 = 0.75$ and $w_6 = 0.75$. All evaluation results for all entities are shown in Fig. 3 where the EDM_i values averaged over all images are depicted. Fig. 4 presents the Segmentation Metric (SM) values for all segmentation algorithms averaged over all images. Fig. 4 shows that the second approach of Tsinghua has an overall advantage.

Concerning text region segmentation, the second approach of Tsinghua achieved the highest averaged EDM rate value (53.22%) while the first approach Tsinghua, the Océ method and the BESUS method achieved an averaged EDM rate value of 46,64%, 31,16% and 29,62% respectively. For graphics, the second approach of Tsinghua achieved the highest averaged EDM rate value (42,38%). For line-art and noise entities, the BESUS method achieved the highest averaged EDM rate values (80% and 20,24% respectively) while for

seperator detection, the Océ method achieved the highest averaged EDM rate value (51,13%). The Tsinghua methods achieved zero EDM rate values for line-art, separator and noise entity segmentation.

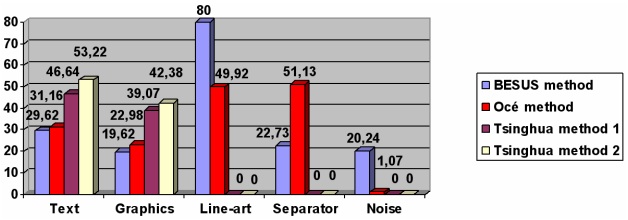

Figure 3. Evaluation results for all entities (EDM_i values averaged over all images).

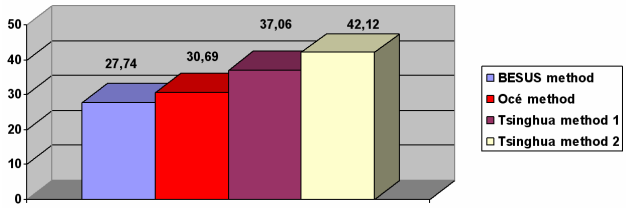

Figure 4. Averaged Segmentation Metric (SM) values.

7 Conclusions

The motivation of the ICDAR2005 Page Segmentation Competition was to evaluate existing approaches for page segmentation and region classification using a realistic dataset and an objective performance analysis system. The image dataset used comprised scanned technical articles and (mostly) magazine pages. The performance evaluation method used is based on counting the number of matches between the entities detected by the algorithm and the entities in the ground truth. The competition run in an off-line mode and evaluated the performance of four segmentation algorithms. The evaluation results show that the second Tsinghua method has an overall advantage (and gives better results for text and graphics). The Océ method is third overall with good consistency (and the best performance on separators). The BESUS method achieved the highest rates for line-art and noise entity segmentation.

References

[1] 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 2003, pp. 688–692.

[2] 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.

[3] 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.

[4] 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.

[5] I. Phillips and A. Chhabra, "Empirical Performance Evaluation of Graphics Recognition Systems," IEEE Transaction of Pattern Analysis and Machine Intelligence, Vol. 21, No. 9, pp. 849-870, September 1999.

[6] A. Chhabra and I. Phillips, "The Second International Graphics Recognition Contest - Raster to Vector Conversion: A Report," in Graphics Recognition: Algorithms and Systems, Lecture Notes in Computer Science, volume 1389, pp. 390-410, Springer, 1998.

[7] I. Phillips, J. Liang, A. Chhabra and R. Haralick, "A Performance Evaluation Protocol for Graphics Recognition Systems" in Graphics Recognition: Algorithms and Systems, Lecture Notes in Computer Science, volume 1389, pp. 372-389, Springer, 1998.

[8] B.A. Yanikoglu, and L Vincent, "Pink Panther: a complete environment for ground-truthing and benchmarking document page segmentation", Pattern Recognition, volume 31, number 9, pp. 1191-1204, 1994.

[9] S. Mandal, S.P. Chowdhuri, A.K. Das and B. Chanda, "Automated Detection and Segmentation of Form Document", *Proceedings of the 5th International Conference on Advances in Pattern Recognition (ICAPR2003),* December 2003, Calcutta, India, pp. 284–288.

[10] A.K. Das and B. Chanda, "Segmentation of Text and Graphics in Document Image: A Morphological Approach", *Proceedings of the International Conference on Computational Linguistics, Speech and Document Processing (ICCLSDP'98),* Calcutta, India, December 1998, pp. A50–A56.

[11] A.K. Das, S.P. Chowdhuri and B. Chanda, "A Complete System for Document Image Segmentation", *Proceedings of national Workshop on Computer Vision, Graphics and Image Processing (WVGIP2002)*, Madurai, India, February 2002, pp. 9–16.

[12] F. Chang, C.J. Chen, and C.J. Lu, "A Linear-Time Component-Labelling Algorithm Using Contour Tracing Technique", *Computer Vision and Image Understanding,* vol. 93, no. 2, 2004, pp. 206–220.

[13] M. Chen, X. Ding, et al. "Analysis, Understanding and Representation of Chinese newspaper with complex layout". Proceedings of 7th IEEE International Conference on Image Processing, 10–13 Sept. 2000, Vancouver, BC, Canada, IEEE..

[14] A. Antonacopoulos, "Page Segmentation Using the Description of the Background" *Computer Vision and Image Understanding*, vol. 70, no. 3, 1998, pp. 350–369.