ICDAR2007 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/~bgat/, bgat@iit.demokritos.gr

Abstract

This paper continues the authors' attempt to address the need for objective comparative evaluation of layout analysis methods in realistic circumstances. It describes the Page Segmentation Competition (modus operandi, dataset and evaluation criteria) held in the context of ICDAR2007 and presents the results of the evaluation of three 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 continue to mature, 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 researched 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 dataset used by their authors. The variety 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. However, 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 third instance of an international generic layout analysis competition (the previous two being the ICDAR2003 and ICDAR2005 Page Segmentation Competitions [1–2]). It should be mentioned that a relatively close previous instance, focusing on a specific application domain, was the First International Newspaper Page Segmentation Contest [3] 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 [4], 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 existing datasets and reports of method results using mostly structured documents (e.g., technical articles).

The competition 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. In addition to the comparative assessment, another objective was to obtain a broad look at the performance of different classes of methods (e.g., connected component analysis, morphological processing, analysis of background etc. as submitted for evaluation) in identifying different types of regions in a variety of documents.

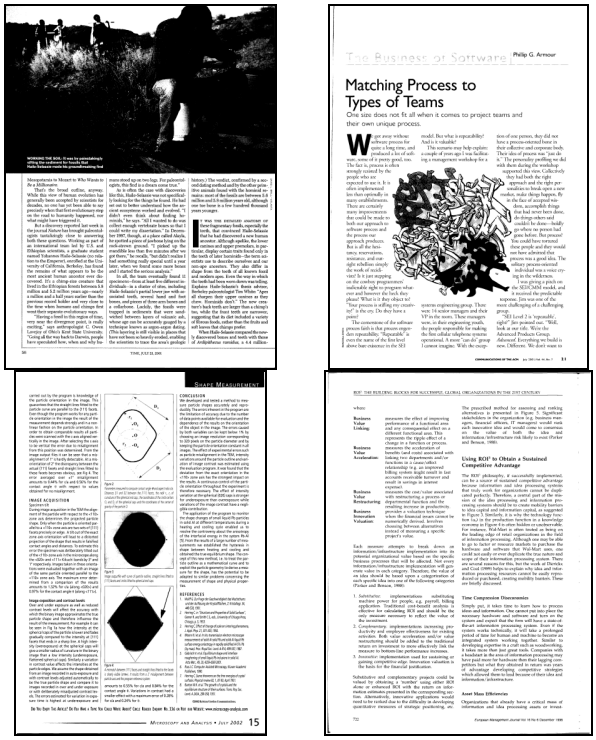

Figure 1. Sample page images from the training dataset.

The competition ran 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 *ground truth*). 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 made available (only the principles were publicised) and above all, the organisers have faith in the authors' scientific integrity.

3 The dataset

It should be noted that there has been 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 and ICDAR2005 competitions [1–2]. 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 of an exhaustive list of document types, the competition subset focuses (for meaningful evaluation purposes) on the most heavily used (in terms of information content and need to analyse) types of documents, such as magazine pages and technical articles.

It should be noted that, as the competition is on page segmentation, the images in the dataset have been processed to remove skew and other artefacts that would affect pre-processing and therefore implicitly also evaluate the pre-processing capabilities of the candidate methods.

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

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 32 page images, comprising 47% technical articles (not necessarily with Manhattan layouts) and 53% magazine pages. It should be noted that also for reasons of tractability, the competition images were bi-level (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: (i) *text*, (ii) *graphics*, (iii) *line art*, (iv) *separator*—graphical line segments between regions, and (v) *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 in [8]).

Let I be the set of all the ON 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 [5], and using a global MatchScore table for all entities, we can define that:

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

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):

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

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):

$$\text{SM} = \frac{\sum_I N_i \text{EDM}_i}{\sum_I N_i} \quad (5)$$

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.

5.1 The Tsinghua methods

D. Wen and X. Ding, of Tsinghua University (State Key Laboratory of Intelligent Technology and Systems), in Beijing, China submitted two methods they developed as part of their effort to build a multi-language page segmentation method. Both methods are improved versions of the methods submitted to the ICDAR2005 competition [2].

Both methods are based on the same kernel, which is called the *Text Line Extraction (TLE)* module. The TLE is designed to solve the (common to both approaches) problem of extracting text lines in various types of document, whether magazines or newspapers, with regular or irregular layouts, English or Chinese (or any other language). It is a bottom-up aggregating method, which starts from connected components and merges them incrementally to obtain hierarchical layout structures. The first step of TLE is *Candidate Line Merging*, where connected components are merged according to their 4-direction Nearest Neighbour Connecting Strength [9] Then in the second step, *Text Line Fitting*, candidate line segments are further merged into integrated text lines by comprehensive consultation of three factors: background separators, single line consistency and neighbouring lines consistency. That is, each pair of neighbouring candidate lines is merged when: 1) there is no background column separator between them; 2) the merged line has good consistency in

character sizes, alignments and spacing; 3) at least one of their common neighbouring lines in the vertical direction suggests them to be merged.

It is based on the results from TLE that different regions are formed. In this subsequent step, the first Tsinghua method (TH1) is different from the second (TH2) with respect to the region shape it supports. TH1 only supports rectangular regions. That is, each region is only represented by its bounding rectangle. For the non-rectangular (isothetic) textual regions, it tends to split them into several rectangular sub-regions. As for irregular graphics and image regions, it will output their bounding boxes only, even if they may overlap with other regions.

On the other hand, TH2 can support irregular regions. It takes the results from TH1 in terms of foreground information and uses a background analysis method to trace the contours of textual regions [10]. Neighbouring textual regions are glued and output as isothetic polygonal regions. However, for the graphics and image regions, the process is still inherited from TH1 so they are still output as bounding boxes.

5.2 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 earlier versions of the method submitted by the authors to the ICDAR2003 and ICDAR2005 competitions [1–2], this is a system constructed using a number of morphology-based modules [11]. The segmentation procedure is applicable to both Manhattan and non-Manhattan layouts and it can detect text in any orientation.

The segmentation is carried out through the following phases:

1. Pre-processing. Skew correction is performed (not necessary in the competition dataset). The information zone is also found out of the whole document by omitting boundary noise.

2. Graphics segmentation. A pseudo-greyscale image is first created (the method works in greyscale whereas the test images were bi-level) using a low-pass adaptive filter based on the size of objects and on the frequency of their occurrence. Morphological open and close operations are then used to generate a unique feature known as OCF matrix [12] which is examined to estimate and remove the graphics regions from the image.

3. Line art segmentation. At this stage the page images contain mainly line art and text. The idea is to remove line art regions using the fact that they do not exhibit regular band structures as text lines do. An extended mask region is computed on all components to form groups and the similarity of the components is examined. Line art regions exhibit different characteristics to text and are identified and removed from the image [13].

4. Text segmentation. Text mostly remains in the image at this point, exhibiting a regular structure of textlines and gaps between them. A vertical window of size $2(Text_{ht}+Gap_{ht})$ is created adaptively based on the statistical estimation of the height of the text band ($Text_{ht}$) and the line gap (Gap_{ht}) in between two text lines. Using this window a rough estimation of text lines is obtained. Further refinement is achieved through the use of additional features such as pen width [14].

6 Results

The performance of the 3 segmentation algorithms (BESUS, TH1 and TH2) was evaluated using equations (1)–(5) for all 32 test images with parameters $w_1 = 1$, $w_2 = 0.75$, $w_3 = 0.75$, $w_4 = 1$, $w_5 = 0.75$ and $w_6 = 0.75$. These parameters are set to give maximum score to one-to-one matches and rather generous scores to other (partial) matches. Evaluation results for all types of entities are shown in Fig. 3 where the EDM_i values averaged over all images are depicted ("noise" regions are omitted as their number was not significant enough). Fig. 4 presents the Segmentation Metric (SM) values for all segmentation algorithms averaged over all images. The BESUS method has a slight overall advantage over TH2 and TH1 with SM results of 55.75%, 55.46% and 51.75% respectively.

In more detail, concerning text region segmentation, the BESUS method achieved the highest averaged EDM_i rate value (68.29%) while TH1 and TH2 achieved an averaged EDM_i rate value of 53.82% and 58.56%, respectively. For graphics, TH1 achieved the highest averaged EDM_i rate value (17.32%). For line-art entities, the BESUS method achieved the highest averaged EDM_i rate value (14.52%) while for separator detection, TH1 and TH2 both achieved the highest averaged EDM_i rate value (64.38%). Both Tsinghua methods achieved zero EDM_i rate values for line-art segmentation.

7 Conclusions

The motivation for the ICDAR2007 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 both 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 ran in an off-line mode and evaluated the performance of three segmentation algorithms. The evaluation results show that the BESUS method has an overall advantage (and gives better results for text and line-art). TH1 and TH2 performed better at segmenting separator regions, while the TH1 method performed best on graphics regions.

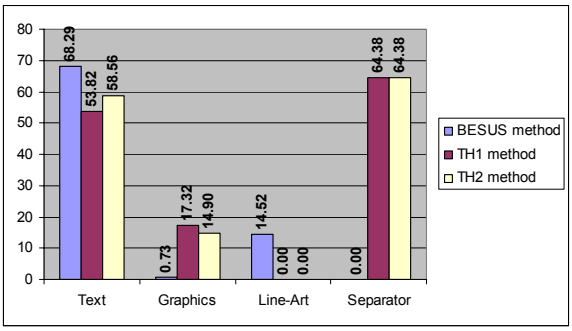

Figure 3. Evaluation results for all entities (EDM_i values averaged over all images).

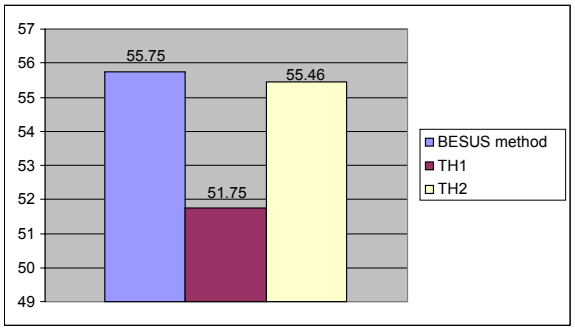

Figure 4. Averaged Segmentation Metric (SM) values.

Acknowledgement

The authors gratefully acknowledge the support of Google in creating the dataset used in this competition.

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] 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 2005, pp. 75–79.

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

[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] 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] 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.

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

[11] 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.

[12] A. K. Das and B. Chanda, "Extraction of half-tones from document images: A morphological approach", *Proceedings of the International Conference on Advances in Computing*, Calicut, India, April 6–8, 1998, pp. 15–19.

[13] S. P. Chowdhury, S. Mandal, A. K. Das, and B. Chandha, "An efficient method for graphics segmentation from document images," *Proceedings of the 6th International Conference on Advances in Pattern Recognition*, Kolkata, India, Jan. 2-4, 2007, pp. 107-111.

[14] S. P. Chowdhury, S. Mandal, A. K. Das, and B. Chandha, "Segmentation of text and graphics from document images," *Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR 2007)*, 23-26 Sept, 2007, Curitiba, Brazil.