Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, USA, September 2001, IEEE Computer Society Press, pp. 1190-1194.

First International Newspaper Segmentation Contest

B. Gatos[1], S. L. Mantzaris[1], A. Antonacopoulos[2]

[1] Department of Digital Technologies, Lambrakis Press Archives,
8 Heyden Str,104 34 Athens, Greece
bgatos@lpa.gr

[2] Department of Computer Science,
The University of Liverpool,
Liverpool, L69 7ZF, U.K.
aa@csc.liv.ac.uk

Abstract

This paper presents the results of the First International Newspaper Segmentation contest that was organized on the frame of ICDAR'2001 conference. The aim of this contest was to evaluate all existing algorithms for document image segmentation that can be applied to Newspaper page segmentation. We evaluated the performance of three different newspaper segmentation algorithms on tracing all basic entities that appear in newspaper pages from the beginning of the previous century up to the present. The selected entities are text regions, lines and images/drawings. Both training and test sets come from Greek and English newspapers. The performance evaluation method is based on counting the number of matches between the entities detected by the algorithms and the entities of the ground truth. In order to rank the global performance of each participant, we employed a metric that combines the average values of detection rate and recognition accuracy.

1 Introduction

The conversion of newspaper pages into digital resources is an important task that greatly contributes to the preservation of and access to newspaper archives. A number of techniques have been proposed in the literature aiming to facilitate automatic page decomposition [1,2]. However, many of these algorithms are not directly applicable to newspaper images, which present special problems. The most significant problems include the complex layout of newspaper pages, where text columns are located very close to each other in a haphazard way, as well as the poor scanning results derived from paper material that was originally of low print quality or has deteriorated through time. Another important problem for newspaper page segmentation is that layout habits seem to have changed repeatedly. The aim of this contest was to evaluate all existing algorithms for document image segmentation that can be applied to Newspaper page segmentation.

All participants of the contest downloaded both training and test data. Training data include 4 newspaper images and the corresponding ground truth files, while test data include 20 newspaper images. Both training and test images come from Greek and English newspapers. The Greek newspapers are "TO BHMA" ("THE TRIBUNE") and "ΕΛΕΥΘΕΡΟΝ ΒΗΜΑ" ("FREE TRIBUNE") and we have selected front pages from years 1922, 1940 and 1968 (see Fig.1a). Both newspapers are published by Lambrakis Press S.A. [3]. The English speaking newspaper is the "International Herald Tribune" [4] and we have selected front pages from years 1900, 1925 and 1950 (see Fig. 1b). The experiments were run by the participants themselves and the final results were submitted to the contest organizers. The ground truth and the result files are coded according to a predefined format given by the organizers.

(a) (b)

Fig. 1. Image samples from the Greek newspaper "TO BHMA" ("THE TRIBUNE") (a) and from the "International Herald Tribune" (b).

2 Definition of Newspaper entities

We have selected to evaluate the tracing of all basic entities that appear in most newspaper pages and cover a time period from the beginning of the previous century till today. We selected 7 entities that belong to text regions, lines and images/drawings:

- *TEXT*: An area consisting of letters whose height is approximately equal to or less than the dominant letter height in the newspaper page. The text region does not split as long as a) the vertical distance between successive text lines remains the same, and b) text style and layout remains the same. Possible drop letters are embodied in the text areas (see Fig. 2).

- *TITLE*: A text area consisting of letters whose height is greater than the dominant letter height in the newspaper page. The title region does not split as long as a) the vertical distance between successive title lines remains the same, b) text style and layout remains the same (see Fig. 3).

- *INVERSE TITLE*: Text or title in black background (see Fig. 4).

- *HORIZONTAL LINE*: Horizontal continuous or broken line.

- *VERTICAL LINE*: Vertical continuous or broken line.

- *PHOTO*: A digitized photo (see Fig. 5a).

- *GRAPHIC/DRAWING*: A graphic or drawing. If there is any text inside a graphic region, then it is embodied in the region (see Fig. 5b).

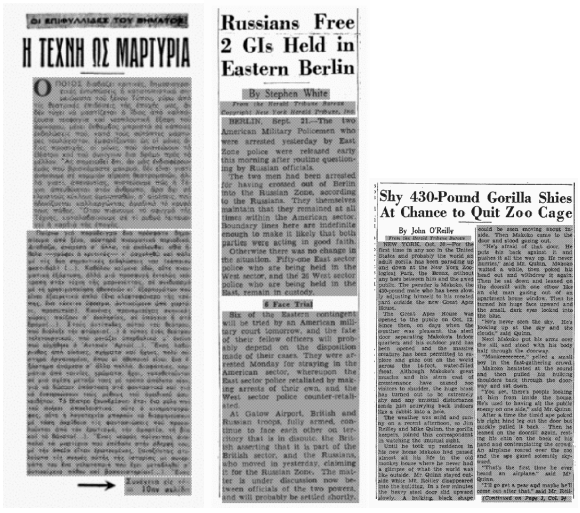


Fig. 2. Text regions.

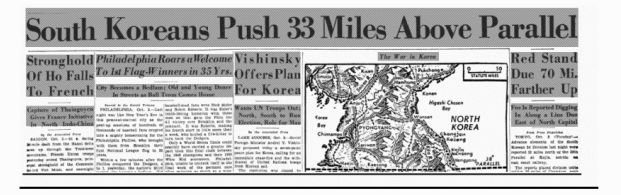
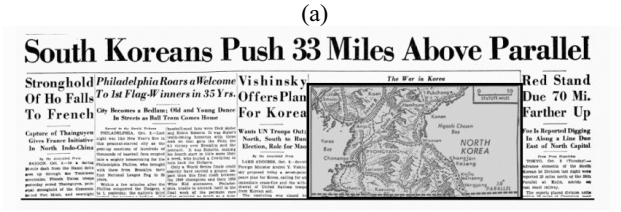


Fig. 3. Title region.

Fig. 4. Inverse title region

(a)



(b)

Fig. 5. Photo (a) and graphic/drawing region (b).

3 Segmentation algorithms

The following segmentation algorithms were evaluated in this contest:

- **Liu, Luo, Yoshikawa and Hu Algorithm**, A New Component based Algorithm for Newspaper Layout Analysis [5]: A components-based bottom-up algorithm for detecting horizontal lines, vertical lines, photos, drawings, titles, inverse titles and texts. Lines are important to separate the text blocks and include both solid line and dashed line. First, all lines are detected, and then the merging progress starts in order

to get the final segmentation result. In the merging progress, the neighboring components are considered for merging. The most acceptable component pair is selected to merge. After the merging progress, further attributes setting progress is used to remove small components and separate title from text, separate drawing and photo from graph.

- **Mitchell and Yan Algorithm**, Newspaper Document Analysis featuring Connected Line Segmentation [6]: The image is initially segmented using a bottom-up approach. Rectangular regions that contain mostly foreground pixels are first located, and then patterns are formed from these regions. Patterns defined this way are larger and fewer in number than connected components, but are guaranteed to segment components separated by more than 3 pixels. The next step is to perform classification based on pattern size, shape and run-length characteristics. The classified patterns are then put through a series of modifications, which include forming complete lines from separate patterns. Finally, the patterns are grouped together to form blocks of patterns all of the same class.

- **Hadjar, Hitz and Ingold Algorithm**: Newspaper Page Decomposition using a Split and Merge Approach [7]: The algorithm is based on the detection of horizontal and vertical line segments, which segment the page into small regions. Neighboring regions are then merged if they meet certain criteria. The algorithm does neither take into account inverted titles nor does differentiate between photos and drawings. It also works without a font recognition, which results in mistakes for the distinction of titles and text regions.

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 [8 -10]. 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 [11]).

Let I 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 j ground truth region and the i result region. Based on a pixel based approach of [8], 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:

$$DetectRate_i = w_1\frac{one2one_i}{N_i} + w_2\frac{g\_one2many_i}{N_i} + w_3\frac{g\_many2one_i}{N_i} \quad (2)$$

$$\text{Re}cognAccuracy_i = w_4\frac{one2one_i}{M_i} + w_5\frac{d\_one2many_i}{M_i} + w_6\frac{d\_many2one_i}{M_i} \quad (3)$$

where $one2one_i$, $g\_one2many_i$, $g\_many2one_i$, $d\_one2many_i$ and $d\_many2one_i$ are calculated from MatchScore table (1) following the steps of [8] for every entity i.

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

$$NCDM_i = \frac{2DetectRate_i \text{Re}cognAccuracy_i}{DetectRate_i + \text{Re}cognAccuracy_i} \quad (4)$$

A global performance metric for detecting newspaper components can be extracted if we combine the average values of detection rate and recognition accuracy. If I is the total number of entities we want to detect, then we can define the following Newspaper Segmentation Metric (NSM):

$$NSM = \frac{2\sum_I DetectRate_i \sum_I \text{Re}cognAccuracy_i}{I(\sum_I DetectRate_i + \sum_I \text{Re}cognAccuracy_i)} \quad (5)$$

5 Results

We evaluated the performance of the 3 newspaper segmentation algorithms using equations (1) - (5) with parameters $w_1 = 1$, $w_2 = 0.25$, $w_3 = 0.25$, $w_4 = 1$, $w_5 = 0.25$ and $w_6 = 0.25$. We took into account the following facts: (a) Both Liu et al. and Mitchell et al. did not define polygon-surrounding areas for horizontal and vertical lines. So, we excluded those entities from the total evaluation process since their results did not follow the corresponding guidelines. (b) Mitchell et al. did not give any results for Photos entity. (c) Hadjar et al. did not give any results for Inverse Title entity. (d) Mitchell et al. gave many unconnected segments for some images, so we excluded those images from the evaluation process of their algorithm. All evaluation results for every entity are shown on Fig. 6-12 and the compete data value report is presented at Table 1. Fig. 13 presents the Newspaper Segmentation Metric (NSM) values for all segmentation algorithms and shows that Liu, Luo, Yoshikawa and Hu algorithm has an overall advantage.

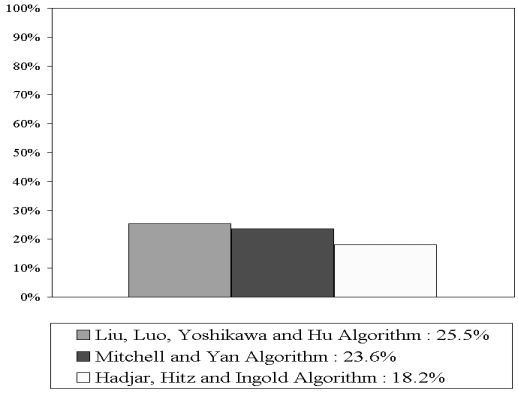


Fig. 6. Evaluation results for text regions ($NCDM_1$).

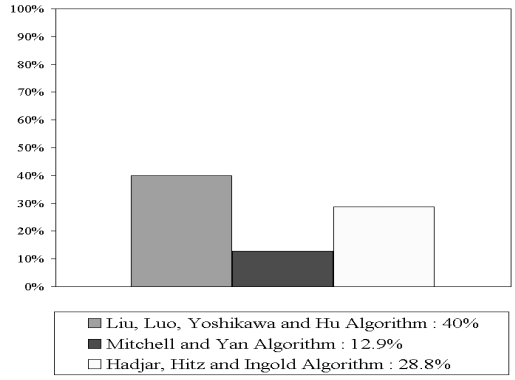


Fig. 7. Evaluation results for title regions ($NCDM_2$).

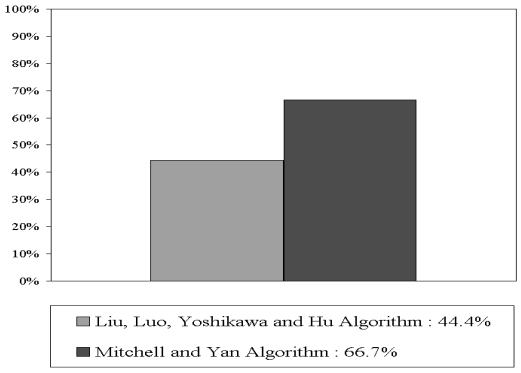


Fig. 8. Evaluation results for inverse title regions ($NCDM_3$).

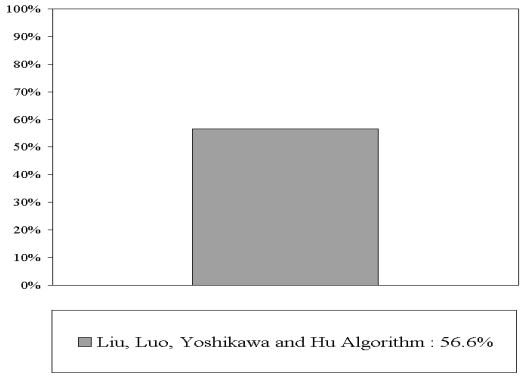


Fig. 9. Evaluation results for photo regions ($NCDM_4$).

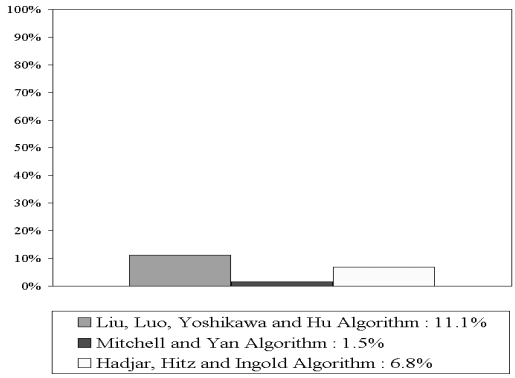


Fig. 10. Evaluation results for graphic regions ($NCDM_5$).

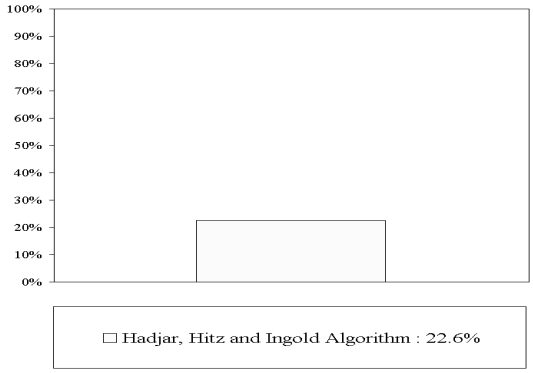


Fig. 11. Evaluation results for vertical line regions ($NCDM_6$).

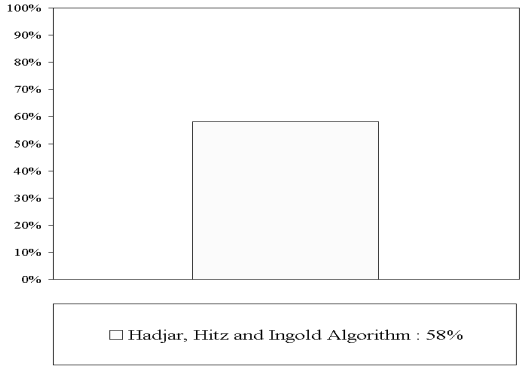


Fig. 12. Evaluation results for horiz. line regions ($NCDM_7$).

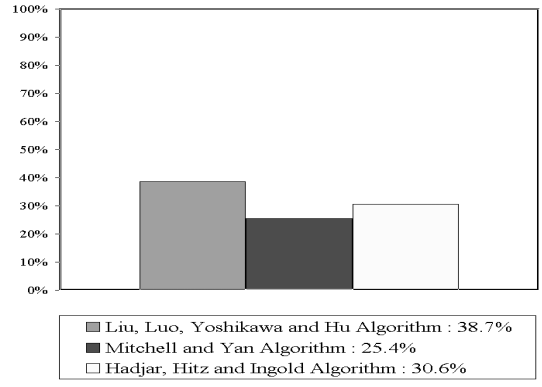


Fig. 13. Newspaper Segmentation Metric (NSM) values.

Concerning text and title region recognition, Mitchell et al. achieved the highest detection rate value (44.2% for text regions) while Liu et al. achieved the highest recognition accuracy value (84.7% for title regions). For inverse title recognition, Mitchell et al. achieved best detection rate and recognition accuracy results (50% and 100% respectively). For photo recognition, we had results only from Liu et al. who achieved a detection rate of 66.7% and recognition accuracy rate of 49.1%. For graphic/drawing recognition all participants got relatively low recognition rates. Finally, for line recognition we had results only from Hadjar et al. who achieved NCDM values of 22.6% and 58% for vertical and horizontal lines correspondingly.

6 Acknowledgments

The organizers of this contest wish to thank "International Herald Tribune" for permitting the use of their newspapers for this contest. Many thanks are also due to the participants who provided us with their segmentation results coded according to given guidelines.

References

[1] C. Strouthopoulos and N. Papamarkos, "Text identification for document image analysis using a neural network", Image and Vision Computing, Vol. 16, Iss. 12-13, pp. 879-896, 1998.

[2] K. Fan, C. Liu and Y. Wang, "Segmentation and Classification of Mixed Text/Graphics/Image Documents", Pattern Recognition Letters, Vol. 15, pp. 1201-1209, 1994.

[3] http://www.dol.gr/, http://www.lpa.gr/

[4] http://www.iht.com/

[5] Fei Liu, Yupin Luo, Masataka Yoshikawa and Dongcheng Hu, "A New Component based Algorithm for Newspaper Layout Analysis", Sixth International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, USA, September 2001.

[6] P. E. Mitchell and H. Yan, , "Newspaper Document Analysis featuring Connected Line Segmentation", Sixth International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, USA, September 2001.

[7] K. Hadjar and O. Hitz, "Newspaper Page Decomposition using a Split and Merge Approach", Sixth International Conference on Document Analysis and Recognition (ICDAR2001), Seattle, USA, September 2001.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

Table 1: Evaluation results (**L:** Liu, Luo, Yoshikawa and Hu, **M:** Mitchell and Yan**, H:** Hadjar, Hitz and Ingold Algorithms).

	Texts (i=1)			Titles (i=2)			Inv.Titles (i=3)			Photos (i=4)			Graphics (i=5)			Vert. Lines (i=6)			Horiz. Lines (i=7)		
	L	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**	**L**	**M**	**H**
N_i	1100	435	1100	755	329	755	12	12	-	42	-	-	14	6	14	-	-	294	-	-	929
M_i	852	1256	1097	233	859	762	6	6	-	57	-	-	4	127	74	-	-	864	-	-	1311
$one2one_i$	228	175	165	194	29	139	4	6	-	28	-	-	1	1	3	-	-	73	-	-	599
$g\_one2many_i$	4	62	35	4	90	16	0	0	-	0	-	-	0	0	0	-	-	98	-	-	102
$g\_many2one_i$	100	8	112	11	0	12	0	0	-	0	-	-	0	0	0	-	-	0	-	-	6
$d\_one2many_i$	45	4	48	5	0	5	0	0	-	0	-	-	0	0	0	-	-	0	-	-	3
$d\_many2one_i$	11	169	85	8	259	33	0	0	-	0	-	-	0	0	0	-	-	301	-	-	273
$misses_i$	768	190	788	556	210	588	8	6	-	14	-	-	13	5	11	-	-	123	-	-	222
false $alarm_i$	568	1008	799	26	571	85	2	0	-	29	-	-	3	126	71	-	-	490	-	-	436
$DetectRate_i$	23,1%	44,2%	18,3%	26,2%	15,6%	19,3%	33,3%	50%	-	66,7%	-	-	7,1%	16,7%	21,4%	-	-	33,2%	-	-	67,4%
$RecognAccuracy_i$	28,4%	16,1%	18%	84,7%	10,9%	56,7%	66,7%	100%	-	49,1%	-	-	25%	0,8%	4%	-	-	17,2%	-	-	50,9%
$MissingDetectionRate_i$	69,8%	43,7%	71,6%	73,6%	63,8%	77,9%	66,7%	50%	-	33,3%	-	-	92,9%	83,3%	78,6%	-	-	41,8%	-	-	23,9%
$FalseAlarmRate_i$	66,7%	74,3%	72,8%	11,2%	66,5%	32,4%	33,3%	0%	-	50,9%	-	-	75%	99,2%	95,9%	-	-	56,7%	-	-	33,3%
$NCDM_i$	**25,5%**	**23,6%**	**18,2%**	**40%**	**12,9%**	**28,8%**	**44,4%**	**66,7%**	-	**56,6%**	-	-	**11,1%**	**1,5%**	**6,8%**	-	-	**22,6%**	-	-	**58%**