Evaluating SEE - A Benchmarking System for Document Page Segmentation

Stefan Agne, Andreas Dengel, and Bertin Klein

German Research Center for Artificial Intelligence (DFKI GmbH)
P.O. Box 2080, D-67608 Kaiserslautern, Germany
e-mail: {stefan.agne, andreas.dengel, bertin.klein}@dfki.de

Abstract

The decomposition of a document into segments such as text regions and graphics is a significant part of the document analysis process. The basic requirement for rating and improvement of page segmentation algorithms is systematic evaluation. The approaches known from the literature have the disadvantage that manually generated reference data (zoning ground truth) are needed for the evaluation task. The effort and cost of the creation of these data are very high.

This paper describes the evaluation system SEE and presents an assessment of its quality.. The system requires the OCR generated text and the original text of the document in correct reading order (text ground truth) as input. No manually generated zoning ground truth is needed. The implicit structure information that is contained in the text ground truth is used for the evaluation of the automatic zoning. Therefore, an assignment of the corresponding text regions in the text ground truth and those in the OCR generated text (matches) is sought. A fault tolerant string matching algorithm underlies a method, able to tolerate OCR errors in the text. The segmentation errors are determined as a result of the evaluation of the matching. Subsequently, the edit operations which are necessary for the correction of the recognized segmentation errors are computed to estimate the correction costs. Furthermore, SEE provides a version of the OCR generated text, that is corrected from the detected page segmentation errors.

1 Introduction

In the domain of document analysis, document page segmentation is a very significant field of research. The task is to divide documents into separate components such as text regions and graphics. For this purpose, several approaches have been developed.

For development and improvement, as well as for the selection of segmentation algorithms, it is important to evaluate these algorithms objectively, especially in comparison to each other. This process is called benchmarking.

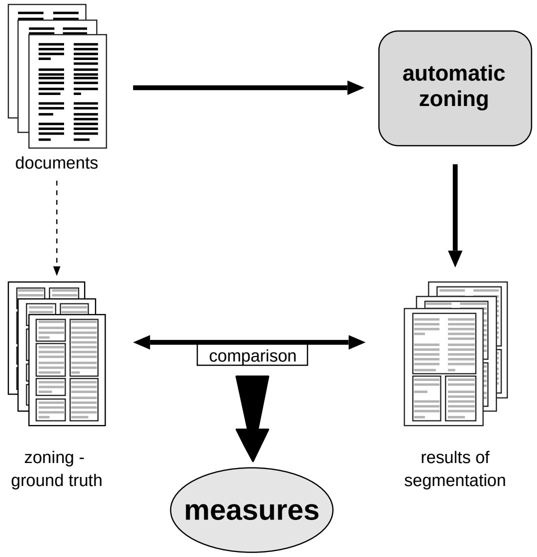

Figure 1. Benchmarking in the field of document analysis

The basic principles of benchmarking in the field of document analysis are shown in Figure 1.

In the first step the zoning ground truth is produced manually for each document. This zoning ground truth is considered the correct decomposition of the document into regions. For instance a region can be specified by a polygon.

During the process of automatic zoning the document is divided automatically into regions. The result of the automatic zoning is then compared with the corresponding zoning ground truth in order to evaluate the quality of the decomposition. Based on this comparison evaluation measures are computed.

The objective of the paper is to present an evaluation method that avoids the need for manual zoning.

2 State of the Art

In the literature two significant classes of approaches for the evaluation of document page segmentation can be found: bitmap based approaches and text based approaches. Bitmap based approaches operate at pixel level on the document bitmap, while text based approaches evaluate the segmentation at character level.

Bitmap based evaluation: Bitmap based evaluation uses the document image (e.g. in TIFF format) and the zoning ground truth in which the regions are described by polygons. Furthermore, the result of the automatic zoning is needed in the same format. The evaluation performs a geometrical comparison between the segmentation results and the zoning ground truth by testing the affiliation of each black pixel to corresponding regions [5, 6, 7, 10]. The bitmap based method classifies the errors into 19 different types [9]. The quality of the segmentation is determined by the number of pixels or characters in the wrongly segmented regions of the document. Yet another bitmap based approach is described and compared to others in [4].

Text based evaluation: Text based evaluation operates on the text output of an OCR system. First the OCR system is applied only to the document image. The resulting output contains segmentation errors and OCR errors. Then the OCR system processes the same document image again, additionally provided with the manually generated zoning ground truth. The resulting text output of the second run contains only OCR errors. For both texts the error correction costs are computed by string matching algorithms (e.g. based on the Levenshtein edit distance). The difference then denotes the costs of correcting the segmentation errors [2, 3].

The bitmap based approaches can be used only when the output of the automatic zoning is available, but many OCR systems do not provide this information.

The text based method assumes that segmentation errors and OCR errors are independent, which generally is not correct. Moreover, this technique does not allow any classification of the segmentation errors.

Also, both approaches need manually generated zoning ground truth for each document. This implies considerable effort and cost. Here we present a method that combines the advantages of these methods and eliminates some of their individual drawbacks.

3 Recognized Errors

Document page segmentation involves decomposing a page into its structural units such as graphics or text regions. An incorrect demarcation of the document page regions by the segmentation algorithm causes segmentation errors. The error model is based on the text regions occuring in the document. In the following, coherent text segments which are not interrupted by empty text-lines are denoted as text regions.

Below we give an overview of possible segmentation errors:

1. **Horizontal merge of text regions**
   In this case text regions of the document are merged horizontally and recognized as a single text region by the OCR-System. Generally, this leads to a spoiling of the reading order as can be seen in Figure 2. In this example the correct reading order is: 1, 2, 3, 4. The segmentation component wrongly merges text line 1 and line 3 into one line. Then the reading order of the text is: 1, 3, 2, 4. This results from merging text segments which belong to different text regions in a horizontal direction, and its manual correction is very time-consuming.

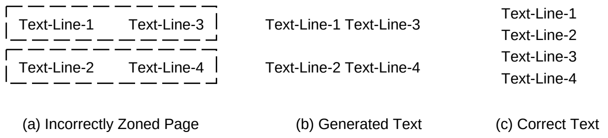

Figure 2. Horizontal merge of text regions

2. **Vertical merge of text regions**
   Here, a single text region is produced by vertically merging text regions of the document. In most cases this error does not lead to a spoiling of the reading order. Thus, this error is practically insignificant, unless the correct segmentation is required for subsequent functional classification of the text.

3. **Horizontal split of text regions**
   This error occurs when the segmentation component recognizes two text regions instead of one. This splitting of a text region is done horizontally. As in the case of horizontal merge this leads to a wrong reading order (see also Figure 2), but it is easier to correct.

4. **Vertical split of text regions**
   This case is similar to the horizontal split except that the split occurs in vertical direction. Generally this error is not critical, since the reading order is not necessarily changed (see vertical merge).

5. **Undetected text region**
   Here, a text region of the document is not recognized as such. It may be wrongly classified as a noise/graphic region. Thus, no text output is created for this region.

6. **Graphic/Noise mistaken for text**
   This error denotes the case that the OCR system interprets graphic or noise as text. Generally, this leads to chaotic character sequences in the generated text.

7. **Horizontal merge with graphic/noise**
   As in the case of the horizontal merge of text regions two separate regions of the document are horizontally merged into a single text region, but here one of these regions is a graphic/noise region. Hence, chaotic character sequences are produced in the generated text.

8. **Vertical merge with graphic/noise**
   Similar to the horizontal merge with graphic/noise, a text region and a graphic/noise region are merged into a single text region, but vertically. This results in the same effects as in the horizontal case.

4 SEE — an Evaluation System

The input data for the evaluation consists of OCR generated text and the corresponding text ground truth. Neither the manually generated zoning ground truth nor the output of the automatic zoning is needed. The text ground truth contains the original text of the document in the correct reading order. Usually, this means no additional effort, since text ground truth in correct reading order is generated for the evaluation of subsequent steps of a document analysis system, e.g. OCR or text categorization.

The *preprocessing* step normalizes both input texts and removes characters that are irrelevant for the evaluation. For instance, two words within one line should be separated by a maximum of one space. After that, the *matching* step assigns corresponding text regions of the text ground truth and OCR generated text. Then, the segmentation errors specified in Section 3, which are a measure of the quality of a segmentation are computed. Subsequently, the costs for the correction of the detected segmentation errors are estimated by the computation of the required edit operations. In addition SEE provides a version of the OCR generated text, that is corrected from the detected page segmentation errors.

You can find a detailed description of the evaluation system SEE in [1].

5 Output of SEE

The output generated by SEE is an error report. Figure 3 shows the output of a test run example.

```
Number of Text Regions
----------------------
Ground Truth......: 10
Output of OCR.....: 14


Segmentation Errors
-------------------
Horizontal Merge......................: 0
Vertical Merge........................: 1
Horizontal Split......................: 0
Vertical Split........................: 4
Undetected Text Region................: 0
Graphic/Noise mistaken as Text........: 1
Horizontal Merge with Graphic/Noise...: 1
Vertical Merge with Graphic/Noise.....: 1


Edit Operations for Correcting the Segmentation Errors
------------------------------------------------------
Move........: 0
Insertion...: 1
Deletion....: 7
```

Figure 3. Output of the evaluation system SEE

The first section shows the number of text regions in the ground truth text and in the OCR generated text, respectively. In the second section statistics on the detected segmentation errors are displayed. For this, the frequency of occurence is determined for each of the eight possible error types. The last part contains the number of edit operations that are necessary to correct the detected segmentation errors.

Furthermore, SEE provides a version of the OCR text, that is corrected from the detected page segmentation errors.

6 How good is SEE?

In this section we will demonstrate the quality of our evaluation system SEE by using it for the evaluation of OCR systems.

Accuracy is the most widely used software tool for the evaluation of OCR systems, developed by the Information Science Research Institute (ISRI) [2, 8]. The most important measure supplied by *Accuracy* is the *Character-Accuracy* (Acc) [8], which is computed as follows:

$$Acc[\%] = \frac{n - errors}{n} * 100$$

where *n* is the number of characters in the ground truth text and *errors* is the minimal number of required edit operations (insertion, deletion, and substitution of a character) to transform the output of the OCR system into the ground truth text.

Using the experiment described below we will demonstrate how well SEE is able to detect and fully automatically correct page segmentation errors made by a commercial OCR system. Figure 4 illustrates our test approach.

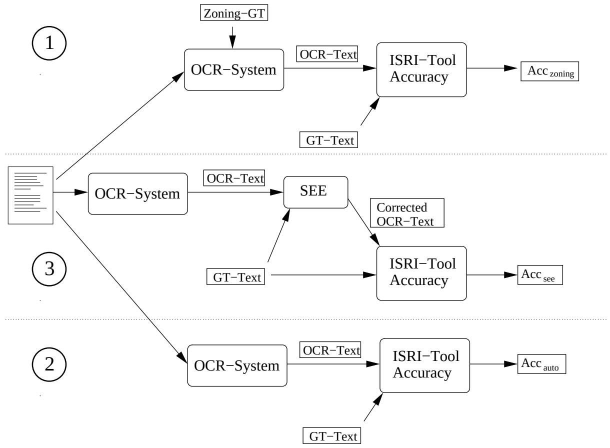

Figure 4. Approach for testing SEE

We used a commercial OCR system, which is not only able to process a scanned document page alone (automatic zoning mode), but also a scanned document page with the according zoning ground truth (manual zoning mode). Then we computed the *Character-Accuracy* for the following three versions of each document:

1. the text produced by the OCR system by using the according zoning ground truth (manual zoning mode)
2. the text produced by the OCR system without using the according zoning ground truth (automatic zoning mode)
3. the text produced as in 2., but subsequently the segmentation errors removed automatically by SEE

The first text includes only character recognition errors, i.e. the character accuracy for this text (Acc_{zoning}) must be the best of all. The second text includes page segmentation errors and character recognition errors, i.e. the characte r accuracy for this text (Acc_{AUTO}) must be the lowest of all three texts.

The third text includes character recognition errors, but most of the page segmentation errors should be corrected by SEE, i.e. the character accuracy for the third text (Acc_{SEE}) should be between these two limits, formally it should satisfy:

$$Acc_{zoning} \geq Acc_{see} \geq Acc_{auto}$$

The smaller the difference between Acc_{see} and Acc_{zoning}, the better the segmentation errors are detected and corrected by SEE.

For this experiment we have used the *German Business Letter Sample*, a sample of 200 German-language business letters collected by the DFKI and used by the ISRI at the fifth annual test of OCR Accuracy [8]. Each document page was scanned by a Fujitsu scanner at resolutions of 200, 300, and 400 dots per inch (dpi) and was twice transmitted locally by a Xerox fax machine to a fax modem to obtain both a standard-mode fax image (204 x 98 dpi) and a fine-mode fax image (204 x 196 dpi).

The results of the experiment described are presented in Table 1.

	Accuracy		
Resolution	Acc_{auto}	Acc_{see}	Acc_{zoning}
200dpi	92.02	94.43	95.14
300dpi	92.58	96.23	96.82
400dpi	92.75	96.66	97.18
Standard-Mode-Fax	82.09	88.46	89.50
Fine-Mode-Fax	88.35	94.84	95.53
Average	89.56	94.12	94.83

Table 1. Results of an evaluation of a commercial OCR system using SEE

The results in Table 1 prove, that SEE is able to detect and correct most page segmentation errors. Acc_{see} is at all resolutions very close to Acc_{zoning}.

To find out more about the evaluation quality of SEE we evaluated two commercial OCR systems, OCR_1 and OCR_2, and compared the evaluation results. Our goal is to decide, which system is better at recognizing the text of facsimiles.

Table 2 shows the evaluation results for documents transmitted by the standard-mode of a fax machine.

	Accuracy		
OCR-System	Acc_{auto}	Acc_{see}	Acc_{zoning}
OCR_1	82.09	88.46	89.50
OCR_2	**82.17**	**88.56**	**89.78**

Table 2. Comparison of two commercial OCR systems with Standard-Mode-Fax

For the Standard-Mode-Fax it's an easy decision for us; the accuracy of OCR_2 is higher for all three methods than the ones of OCR_1, i.e. we would choose system 2. Now, we did a second experiment with the Fine-Mode-Fax. You can see the results in Table 3.

Unfortunately, there is no consistent decision for one of the two OCR systems at the fine-mode-fax. Please remember, there are cases where it's immpossible to compute the correct accuracy Acc_{zoning}, e.g. always when

	Accuracy		
OCR-System	Acc_{auto}	Acc_{see}	Acc_{zoning}
OCR_1	**88.35**	94.84	95.53
OCR_2	87.81	**95.35**	**96.37**

Table 3. Comparison of two commercial OCR systems with Fine-Mode-Fax

- no zoning ground truth exists for the documents
- the OCR system does not have a manual zoning mode

In these cases only the approximation of the correct accuracy, Acc_{auto} and Acc_{see}, is available. But, based on the value of Acc_{auto}, one would make the (wrong) decision for OCR_1.

However, SEE detects and corrects most page segmentation errors to get an more exact estimation of the text recognition errors. So Acc_{see} is close enough to the correct accuracy Acc_{zoning} to take the right decision, i.e. one would choose the actually better system 2.

7 CONCLUSIONS

We have presented the evaluation system SEE, which has been developed for benchmarking document page segmentation systems. The input data for the evaluation task are an OCR generated text and the corresponding ground truth text in correct reading order. Mostly this ground truth text is generated anyway for the evaluation of further processing steps like text recognition or text categorization.

Contrary to the bitmap based approaches, SEE is able to evaluate the segmentation of OCR systems which do not provide the results of automatic zoning. Furthermore, the segmentation errors can be classified, which was not possible with the text based evaluation methods. The fact that SEE does not need the manually generated zoning ground truth as input leads to a significant reduction of effort and cost. As a side effect of this SEE can only approximate the number of true occuring segmentation errors. However, our tests have revealed very good results regarding the quality of the evaluation. In addition, SEE enables for the first time the character accuracy evaluation of OCR systems lacking a manual zoning mode.

ACKNOWLEDGMENTS

The authors want to express their appreciation to George Nagy for valuable discussions and helpful comments. We also wish to thank Jane Bensch for proof-reading the document and Marcus Sigmund for his practical assistance.

References

[1] S. Agne, M. Rogger, and J. Rohrschneider. Benchmarking of document page segmentation. In D. P. Lopresti and J. Zhou, editors, *Document and Recognition and Retrieval VII*, volume 3967 of *Proceedings of SPIE*, pages 165 – 171, San Jose, California, USA, 2000.

[2] J. Kanai, S. V. Rice, and T. A. Nartker. A preliminary evaluation of automatic zoning. In K. O. Grover, editor, *Annual Research Report*, pages 35–45, University of Nevada, Las Vegas, 1993. Information Science Research Institute.

[3] J. Kanai, S. V. Rice, T. A. Nartker, and G. Nagy. Automated evaluation of OCR zoning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(1):86–90, January 1995.

[4] L. Peng, M. Chen, C. Liu, X. Ding, and J. Zheng. An automatic performance evaluation method for document page segmentation. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition (ICDAR)*, pages 134 – 137, Seattle, Washington, USA, September 10-13 2001. IEEE Computer Society Press.

[5] S. Randriamasy and L. Vincent. Benchmarking page segmentation algorithms. In *Proceedings of the 1994 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 411–416, Seattle, Washington, USA, June 21–23 1994. IEEE Computer Society Press.

[6] S. Randriamasy and L. Vincent. A region-based system for the automatic evaluation of page segmentation algorithms. In A. Dengel and A. L. Spitz, editors, *Proceedings of the International Association for Pattern Recognition Workshop on Document Analysis Systems DAS94*, pages 29–41, Kaiserslautern, Germany, October 18–20 1994.

[7] S. Randriamasy, L. Vincent, and B. Wittner. An automatic benchmarking scheme for page segmentation. In *Proceedings of the IS&T/SPIE 1994 International Symposium on Electronic Imaging Science and Technology*, volume 2181, pages 217–230, 1994.

[8] S. V. Rice, F. R. Jenkins, and T. A. Nartker. The fifth annual test of OCR accuracy. Technical Report TR-96-01, Information Science Research Institute, University of Nevada, Las Vegas, USA, April 1996.

[9] M. Thulke, V. Märgner, and A. Dengel. A general approach to quality evaluation of document segmentation results. In S.-W. Lee and Y. Nakano, editors, *Proceedings of the International Association for Pattern Recognition Workshop on Document Analysis Systems DAS98*, pages 79–88, Nagano, Japan, November 1998.

[10] B. A. Yanikoglu and L. Vincent. Ground-truthing and benchmarking document page segmentation. In *Proceedings of the Third International Conference on Document Analysis and Recognition*, volume 2, pages 601–604, Montréal, Canada, August 14–16 1995. IEEE Computer Society Press.