

THE METHODOLOGY FOR THE EVALUATION OF THE ALGORITHMS FOR TEXT LINE SEGMENTATION

Darko Brodić

University of Belgrade
Technical Faculty Bor
V.J. 12, 19210 Bor, Serbia

Dragan R. Milivojević

Mining and Metallurgy Institute
Department of Informatics
Z. Bulevar bb, 19210 Bor, Serbia

Abstract

Text line segmentation represents the key for the correct optical character recognition. All previous works deal with different text databases as a reference for the evaluation of the text line segmentation algorithm. Due to inconsistencies in this process, some basic methodology for the experiments is required. In this paper, basic methodology based on the experiments for the evaluation of the algorithm's text line segmentation is proposed. This experiments set consists of a various multi-line text samples primarily linked with the text line segmentation. Although they are mutually independent, the obtained results are cross linked. At the end, its suitability for different types of letters and languages are its main advantages.

Keywords: optical character recognition, document image processing, text line segmentation, testing methodology.

INTRODUCTION

Text line segmentation is an essential stage in off-line optical character recognition (OCR) systems. It is a key because inaccurately segmented text lines will lead to OCR failure [1]. Although text line detection techniques are mainly successful in printed documents, processing of the handwritten documents has remained a key problem [2]. Consequently, text line segmentation of handwritten documents is a complex and diverse problem, complicated by the nature of handwriting. Hence, it is a leading challenge in handwritten document image processing [3].

Many proposed algorithms for text line segmentation have been evaluated by quite different test methods. In fact, these evaluation procedures are usually based on the use of a custom text database as a test sample. Accordingly, testing result interpretation is quite dissimilar [4]. Due to inconsistencies in measurement and evaluation of text segmentation algorithm quality, some basic methodology is required. Currently, there is no commonly accepted one [5]. Hence, the establishment of the basic set of experiments set for the evaluation of the algorithm's text line segmentation is of great importance.

In this paper, a basic methodology for the evaluation of algorithms text line segmentation

is proposed. It consists of a few experiments primarily linked to text line segmentation. Although they are mutually independent, the results obtained are strongly cross linked. In the end, its suitability for different types of letters and languages as well as its adaptability are its main advantages. Thus, the paper presents an efficient evaluation method for text analysis algorithms.

The paper is organized as follows: in Section 2 the basic experiments set for the text line segmentation is presented. Section 3 contains an examination of the test procedure on the example of the region growing algorithm. In Section 4 results are analyzed and discussed. Conclusions are given in Section 5.

BASIC METHODOLOGY

Basic methodology for the evaluation of algorithm's text line segmentation consists of few text experiments. It is based on the following activities [5]:

1. Multi-line text segmentation test,
2. Multi-line waved text segmentation test,
3. Multi-line fractured text segmentation test.

The schematic basic experiments set procedure is shown in Fig. 1.

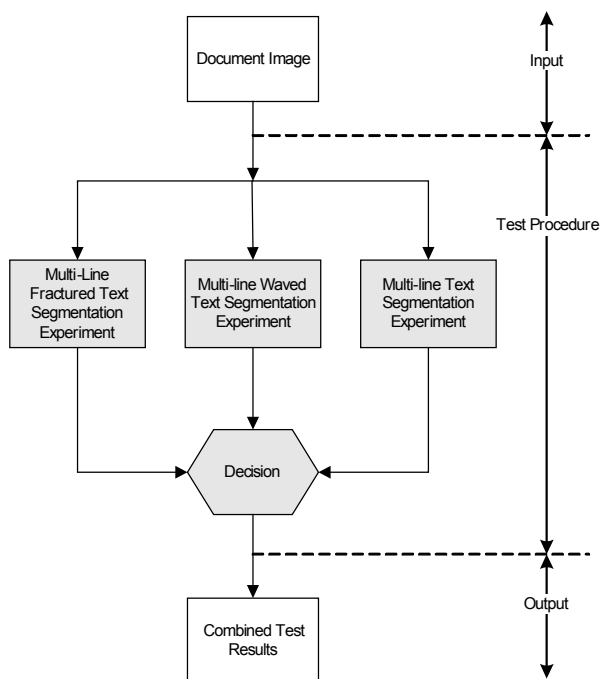


Fig. 1. Schematic test procedure.

All above tests are suitable for the region growing algorithm as well. For the measurement evaluation of text line segmentation following terms are defined:

- Initial objects,
- Detected objects,
- Referent objects.

Initial objects O_{in} represent the starting number of text objects in referent sample text. After applying the algorithm over sample text counted number of objects is given as detected objects O_{dt} . Further, the desired number of objects O_{ref} represents the number of text lines in referent sample text. It is called referent number of objects. By comparing the referent and detected number of objects the algorithm efficiency is evaluated. This evaluation is accomplished by $RMSE$ method. Hence, it is defined by [6]:

$$RMSE_{seg} = \sqrt{\frac{1}{Q} \sum_{i=1}^Q (O_{ref}^k - O_{dt}^k)^2}, \quad (1)$$

where $k = 1, \dots, Q$ is number of referent text samples, O_{ref}^k is number of referent, while O_{dt}^k is number of detected text objects [5]. Further, SHR (Segmentation Hit Rate) value has been introduced. It represents number of

correct text line segmentation hits. It is counted when relation $O_{ref}^k = O_{dt}^k$ is valid.

Although all test experiments are diverse, their results are inter-related. Hence, those results are combined and conclusions are made.

MULTI-LINE STRAIGHT TEXT SEGMENTATION

In the first experiment, multi-line straight text sample is used. This sample text is defined by parameter β representing the skew angle of the straight text. Inter-line space is set to 20% of the biggest character size [7]. Sample multi-line straight text is shown in Fig. 2.

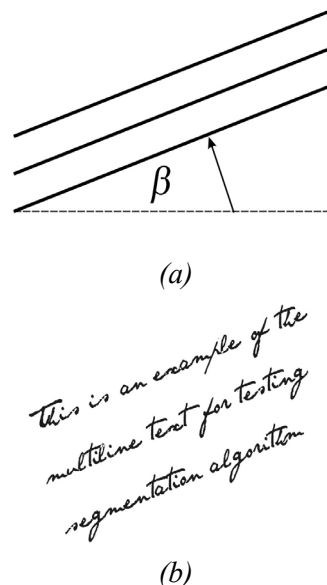


Fig. 2. Multi-line straight text: (a) definition, and (b) example.

As a measure of quality, the root mean square error $RMSE_{seg}$ according to Eq. (1) has been used.

MULTI-LINE WAVED TEXT SEGMENTATION

Second experiment is with multi-line waved text one. Sample text is formed as a group of text lines using the waved referent line as a basis. Referent line is defined by the parameter $\varepsilon = h / l$. Typically, ε is used from the set $\{1/8, 1/6, 1/4, 1/3, \dots\}$ [5]. Sample multi-line waved text for the experiment is shown in Fig. 3.

Similarly, as a previous segmentation test, the number of existing text objects after applied algorithm relate to the text segmentation quality success. Again, for the quality measure, the root mean square error $RMSE_{seg_wav}$ according to Eq. (1) has been used.

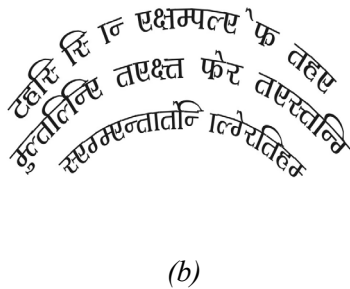
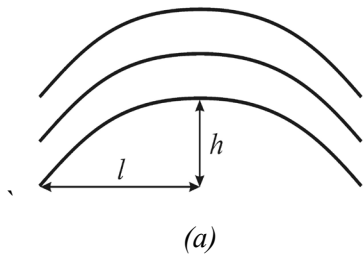


Fig. 3. Multi-line wavy text: (a) definition, and (b) example.

MULTI-LINE FRACTURED TEXT SEGMENTATION

In the last experiment multi-line fractured text is used. Currently, sample text is formed by using referent fractured line as a basis. Fractured text referent line is defined by the slope angle ϕ , as a parameter. Typically, ϕ is used from the set $\{5^\circ, 10^\circ, 15^\circ, 20^\circ\}$ [5]. Sample multi-line fractured text is shown in Fig.4.

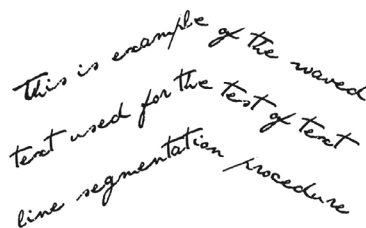
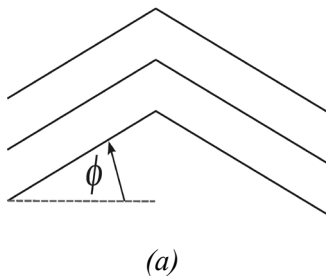


Fig. 4. Multi-line fractured text: (a) definition, and (b) example.

Again, the number of detected text objects relate to the text segmentation quality success.

The root mean square error $RMSE_{seg_frac}$ according to Eq. (1) has been used, as a measure of quality.

TESTING THE ALGORITHMS

For the illustration, proposed testing method will be examined by the example of the basic Gaussian circular algorithm as well as its morphological extension counterpart.

The basic algorithm will be just briefly explained. Its main task is expanding black pixel areas of text by scattering every black pixel in its neighborhood. This way, control areas that separate text lines are established. Its primary purpose is joining only text elements from the same text line into the same distinct control areas. Gaussian probability function is taken as template that gives the probability of the random function. It represents probability of the hypothetical expansion of every black pixel that represents a text element. Hence, around every black pixel, new pixels are non-uniformly dispersed. These new pixels have lower black intensity. Because the level of probability expansion relates to distance from black pixel, their intensity depends completely on their position i.e., the distance from the original black pixel. These newly formed pixels are grayscale. Currently, document text image is represented by a grayscale image matrix. Thus, intensity pertains in level region $\{0, \dots, 255\}$. Hence, after applying Gaussian circular kernel, equal to window $2K+1$ in x -direction as well as in y -direction, text is scattered forming an enlarged area around it. Now, inside the kernel a “probability” sub area is formed using the radius 3σ , where σ represents standard deviation defining curve spread parameter. Converting all these pixels into black pixels as well as inverting image, forms the new black pixel expanded areas [8]. These areas are named boundary-growing areas (BGA).

To extend BGA made by circular Gaussian kernel morphological dilatation is proposed. It is given as [9]:

$$\mathbf{X} \oplus \mathbf{S}_1 \quad (2)$$

$\mathbf{X}$ represents document image matrix after applying Gaussian circular algorithm. Dilatation structuring element $\mathbf{S}_1$ represents line. This line is defined by its height $h_s=1$,

width $w_s=2(R-K)+1$ and parameter $\lambda=R/K$. Now, extended circular Gaussian kernel is equal to $R \times K$ instead of just $K \times K$.

The main purpose of the testing is the optimization of the algorithm parameters. In our example, the parameters of interest are K and λ . K defines kernel size, while λ extends kernel size in x -direction [10].

COMPARATIVE ANALYSIS

In the first experiment, text with typical character height $H_{ch} \approx 100$ px is used. From [8,10] parameter K value may not exceed 20% of H_{ch} . In fact, bigger K could lead to text lines merging. $RMSE_{seg}$ inspection is shown in Fig. 5.

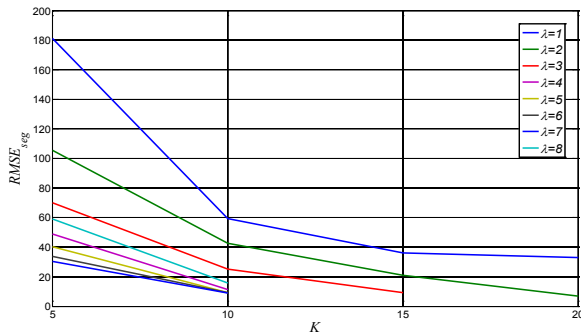


Fig. 5. $RMSE_{seg}$ vs. K (λ as parameter) for multi-line straight text segmentation experiment.

From Fig. 5 optimal K is 15 or 20. Hence, optimal λ is from set $\{2, 3\}$. Bigger λ values will join different text lines causing faulty text line segmentation. Further, SHR value for this experiment is given in Table I.

Table I. SHR for multi-line straight text segmentation experiment.

K	$\square$	SHR							
		1	2	3	4	5	6	7	8
5		0	0	0	0	0	0	0	0
10		0	0	0	1	3	1	0	0
15		0	4	7	4	1	1	1	1
20		0	7	3	1	1	1	1	1

Optimal $\{K, \lambda\}$ pairs from Table I are: $\{15, 3\}$ and $\{20, 2\}$.

$RMSE_{seg_wav}$ obtained from the multi-line waved text segmentation experiment is shown in Fig. 6.

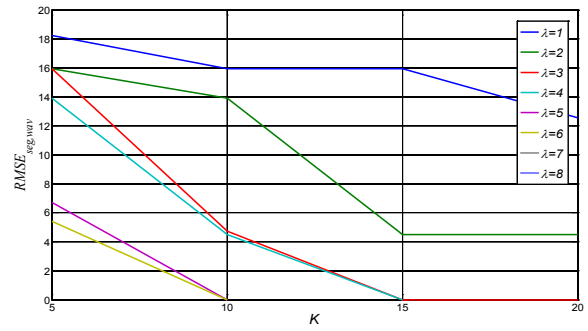


Fig. 6. $RMSE_{seg_wav}$ vs. K (λ as parameter) for multi-line waved text segmentation experiment

In this experiment, using K of 15 or 20 is the good choice. This leads to λ of at least 3. It can be noticed that bigger λ will merge different text lines. Further, SHR value for this experiment is given in Table II.

Table II. SHR for multi-line waved text segmentation experiment.

K	$\square$	SHR							
		1	2	3	4	5	6	7	8
5		0	0	0	0	0	0	0	0
10		0	0	2	3	4	4	3	3
15		0	3	4	4	3	2	2	2
20		2	3	4	2	2	1	1	0

Optimal $\{K, \lambda\}$ pairs from Table II are: $\{10, 5-6\}$, $\{15, 3-4\}$, and $\{20, 3\}$.

In the last text segmentation experiment, $RMSE_{seg_frac}$ is evaluated. It is shown in Fig. 7.

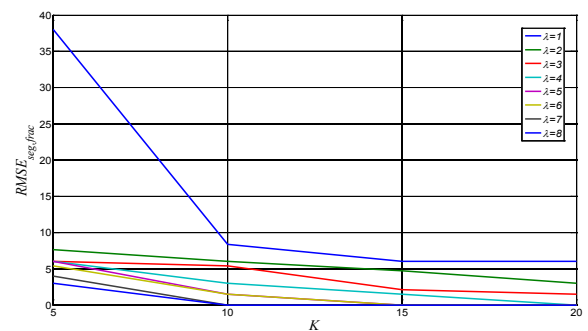


Fig. 7. $RMSE_{seg_frac}$ vs. K (λ as parameter) for multi-line fractured text segmentation experiment

Again, K of 15 or 20 is prerequisite. Still, using λ of 3 or bigger is the best choice. Further, SHR value for this experiment is given in Table III.

Table III. *SHR* for multi-line fractured text segmentation experiment.

K	λ	<i>SHR</i>							
		1	2	3	4	5	6	7	8
5		0	0	0	0	0	0	0	0
10		0	0	0	0	3	3	4	4
15		0	0	2	3	4	4	4	4
20		0	0	3	4	4	4	4	4

Optimal $\{K, \lambda\}$ pairs from Table III are: $\{10, 7-8\}$, $\{15, 5-8\}$, and $\{20, 4-8\}$.

At the end, summary results for the original Gaussian kernel algorithm are not so convincing due to its faulty results obtained from segmentation experiment. All three experiments proved strong inter-relation by their obtained results. Decision making procedure for optimal parameter value is shown in Fig. 8.

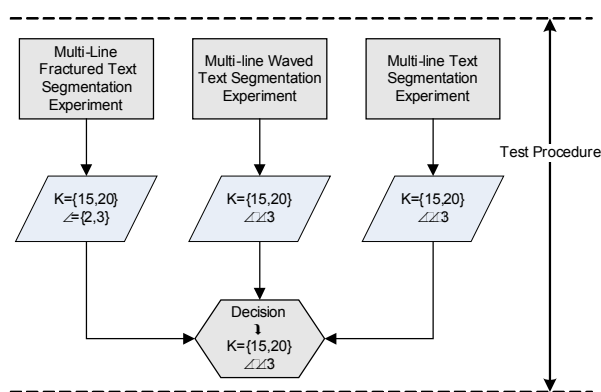


Fig. 8. Decision making procedure for optimal parameter value.

Hence, parameter value of $K = \{15, 20\}$ is mandatory for the efficient text line segmentation process. Also, optimal value of λ is from 2 to 4.

After decision making by intersection of the obtained results Gaussian circular algorithm extended by morphological dilatation is proved to be advanced over the basic Gaussian circular algorithm in the domain of the text line segmentation. Hence, proposed methodology for testing algorithms proved to be efficient.

CONCLUSION

This paper describes the proposal of the methodology for the evaluation of the algorithms text line segmentation. All previous evaluation procedures were custom oriented. However, the proposed test method is the first step toward generalization of the test procedure in the domain of document image proce-

ssing algorithms evaluation. It consists of the test experiments that measure text line segmentation algorithm quality. They incorporate three various multi-line text experiments. Results from all test experiments are inter-related. Hence, they are combined after some decision making by the intersection of obtained results. The benefit of this process is optimized subset of algorithm parameters values. Hence, this process is invaluable for algorithm evaluation. At the end, its suitability for different types of letters and languages as well as its adaptability is a strong advantage.

REFERENCE

- [1] Basu, S.; Chaudhuri, C.; Kundu, M.; Nasipuri, M.; Basu, D.K. Text Line Extraction from Multi-Skewed Handwritten Documents. Pattern Recognition 2006, Vol.40, pp.1825–1839.
- [2] Amin, A.; Wu, S. Robust Skew Detection in mixed Text/Graphics Documents. In Proceedings of International Conference on Document Analysis and Recognition (ICDAR), Seoul, Korea, 2005.
- [3] Likforman Sulem, L.; Zahour, A.; Taconet, B. Text Line Segmentation of Historical Documents: A Survey. IJDAR 2007, Vol.9, pp.123–138.
- [4] Razak, Z.; Zulkiflee, K.; Idris, M.Y.I.; Tamil, E.M.; Noor, M.N.M.; Salleh, R.; Yaakob, M.; Yusof, Z.M.; Yaacob, M. Off-Line Handwriting Text Line Segmentation: A Review. IJCSNS 2008, Vol.8, pp.12–20.
- [5] Brodić, D.; Milivojević, D. R.; Milivojević, Z. Basic Test Framework for the Evaluation of Text Line Segmentation and Text Parameter Extraction. Sensors 2010, Vol.10, No.5, pp. 5263-5279.
- [6] Bolstad, W. M. Introduction to Bayesian Statistics, John Wiley & Sons, NJ, U.S.A.
- [7] Brodić D., Dokić B., Initial Skew Rate Detection Using Rectangular Hull Gravity Center, E2010 - 14th International Conference on Electronics, Vilnius, 2010.
- [8] Brodić, D.; Milivojević, Z. Using Anisotropic Gaussian Window for Printed and Handwritten Text Parameters Extraction. Proceedings of the 9th International Scientific Conference (UNITECH'09), Gabrovo, Bulgaria, 2009, Vol.1, pp. 453–460.
- [9] Gonzalez, R. C.; Woods, R. E. Digital Image Processing. 2nd ed. New Jersey: Prentice-Hall, 2002.
- [10] Brodić, D. Optimization of the Anisotropic Gaussian Kernel for Text Segmentation and Parameter Extraction, Proceedings of WCC, Theoretical Computer Science 2010, Calude C.; Sassone V., Ed., Springer IFIP AICT 323, Brisbane, Australia, 2010, pp. 140-152.

