Text Line Segmentation by Gaussian Kernel Extended with Binary Morphology

Darko Brodić[1] and Branko Dokić[2]

***Abstract* – In this paper, new approach to the text segmentation by Gaussian kernel is presented. As a result of basic algorithm, defined area exploited for text segmentation and text parameter extraction. To improve text segmentation process, basic method is extended by binary morphological operations. Basic and extended algorithm is examined and evaluated under different text samples. Results are examined, analyzed and discussed.**

***Keywords* – OCR, Document image processing, Text Line Segmentation, Gaussian kernel, Morphological operation.**

I. INTRODUCTION

Printed and handwritten text is characterized by its attributes and features diversity. Hence, text parameters extraction procedure can be quite dissimilar one. But, such algorithm should be valid for printed as well as for handwritten text. Text line segmentation is the major step in document processing procedure. Although some text line detection techniques are successful for printed documents, processing of handwritten documents has remained a key problem in OCR [1,2]. Most text line segmentation methods are based on the assumptions that distance between neighboring text lines is significant as well as that text lines are reasonably straight. However, these assumptions are not always valid for handwritten documents. Hence, text line segmentation is a leading challenge in document processing.

Related work on text line segmentation can be categorized in few directions [3]: projection based methods, Hough transform methods, smearing methods, grouping methods, methods for processing overlapping and touching components, stochastic methods, others method.

Projection base methods have been primarily used for printed document segmentation, but it can be adapted for handwritten documents as well. It uses the vertical projection profile (VPP), which is obtained by summing pixel values along the horizontal axis for each y value. This is accomplished by finding its maximum and minimum value [4]. Because of method drawbacks, short lines will provide low peaks, and very narrow lines. Hence, method failed to be efficient for multi-skewed text lines.

The Hough transform [5] is a widespread technique for finding straight lines in the images. Consequently, image is transformed in the Hough domain. Potential alignments are hypothesized in Hough domain and validated in the image domain. The direction for the maximum variation is determined by a cost function. The "voting" function in Hough domain determine slope of the straight line [6].

In smearing methods the consecutive black pixels along the horizontal direction are smeared [7]. This way, enlarged area of black pixels is formed. It is so-called boundary growing area. Consequently, the white space between black pixels is filled with black pixels. It the valid only if their distance is within a predefined threshold.

Grouping methods is based on building alignments by aggregating them [8]. The units may be pixels or connected components, blocks or other features such as salient points. These units are joined together to form alignments. The joining scheme is based on both local and global criteria used for checking consistency. If the nearest neighbour belongs to another line, then the nearest-neighbour joining scheme will fail to group complex handwritten units.

Method for overlapping and touching components detects such components during the grouping process when a conflict occurs between two alignments [9]. Further, it applies a set of rules to label overlapping or touching components. The rules use as features the density of black pixels of the component in each alignment region, alignment proximity and positions of both alignments around the component. The frontier segment position is decided by analyzing the component VPP. If VPP includes two peaks, the cut will be done in the middle way from them. Otherwise, component will cut in two equal parts.

Stochastic method is based on probabilistic algorithm, which accomplished non-linear paths between overlapping text lines. These lines are extracted through hidden Markov modelling (HMM) [10]. This way, the image is divided into little cells. Each one them correspond to the state of the HMM. The best segmentation paths are searched from left to right. In the case of touching components, the path of highest probability will cross the touching component at points with as less black pixels as possible. However, the method may fail in the case that contact point contains a lot of black pixels.

In this paper, modification of the base method proposed in [12] are implemented, analyzed, examined and compared. It is simple and efficient method in terms of accuracy and computations. Its primary role is to perform text segmentation and to estimate the skew angle of document image. The proposed method is implemented and "measured" in different sample text examples and evaluated as well.

Organization of this paper is as follows. Section II includes brief description and information on proposed algorithm. In Section III text experiments are defined. Further, in Section IV given results are compared and discussed. In Section V conclusion is made and further investigation is pointed out.

[1]Darko Brodić is with the University of Belgrade, Technical Faculty Bor, Vojske Jugoslavije 12, 19210 Bor, Serbia, E-mail: dbrodic@tf.bor.ac.rs

[2]Branko Dokić is with the University of Banja Luka, Faculty of Electrical Engineering, Patre 5, 51000 Banja Luka, Republika Srpska, BIH, E-mail: bdokic@etfbl.net

II. PROPOSED ALGORITHM

The principal stages in document processing system are scanning, binarization, text segmentation, text parameter extraction, text recognition and conversion to ASCII. However, that procedure can be represented with three main stages as shown in Fig.1.

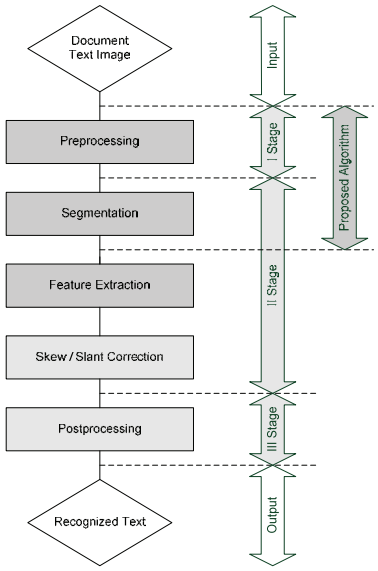


Fig.1. Document processing procedure

In preprocessing stage, algorithms for document text image binarization and normalization are applied. During the processing stage, algorithms for text segmentation as well as for reference text line estimation and skew rate identification are enforced. After that, skew angle is corrected. At the end, in postprocessing stage character recognition process is applied. Final sub stage is the conversion to ASCII characters.

A few assumptions should be made before algorithm description. In this paper, there is an element of preprocessing. After that, document text image is prepared for feature extraction. Main tasks are text segmentation as well as text parameter extraction, specifically reference text line identification and skew rate estimation.

Document text image is an input of text grayscale image described by following intensity function:

$$D\,(m,n) \in [0,\ldots,255] \quad , \tag{1}$$

where m ∈ [0, M-1] and n ∈ [0, N-1].

After applying intensity segmentation with binarization, intensity function is converted into binary intensity function given by:

$$D_{bin}(m,n) = \begin{cases} 1 & for \quad D(m,n) \geq D_{th} \\ 0 & for \quad D(m,n) < D_{th} \end{cases} \quad , \tag{2}$$

where D_{th} is given by Otsu algorithm [13].

Now, extracted text lines are represented as digitized document image by matrix **X** featuring M rows by N columns. Further, document text image is represented as black and white image. Hence, it consists of the only black and white pixels. Each character or word consists of the only black pixels. Every pixel $X(i,j)$ is represented by number of coordinate pairs such as:

$$X\,(i,\,j) \in [0,255] \quad , \tag{3}$$

where i = 1, …, M, j = 1, …, N of matrix **X** [14]. In addition, value 0 represents black pixels, while value 255 represents white pixels.

Prior to processing stage, document text image should be "prepared" for it. It is assumed text area is extracted by some appropriate method. Further, morphological preprocessing is performed to make document text image "noiseless". The morphological preprocessing was defined in [14-15] by following steps: document image erosion: $\mathbf{X} \ominus \mathbf{S}_1$, document image opening: $\mathbf{X} \circ \mathbf{S}_1$, dilatation of the opening the document image: $(\mathbf{X} \circ \mathbf{S}_1) \oplus \mathbf{S}_1$ and closing of the opening the document image: $(\mathbf{X} \circ \mathbf{S}_1) \bullet \mathbf{S}_1$. For these morphological operations, structuring element $\mathbf{S}_1$ dimension 3x3 is used [14-16].

For the processing stage Gaussian kernel algorithm is used. It is based on 2D Gaussian function given by [12]:

$$f(x,y) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-b_x)^2+(y-b_y)^2}{2\sigma^2}} \quad , \tag{4}$$

where b_x is shift along x-axis, b_y is shift along y-axis, σ is curve spread parameter and A is the amplitude.

From (4) it is obvious that curve spread parameter σ is equal for x as well for y-axis. This way, Gaussian function is isotropic. Converting Gaussian function into point spread function, Gaussian kernel is obtained.

Algorithm using Gaussian kernel expands black pixel area by scattering every black pixel in its neighborhood. Around every black pixel new pixels are non-uniformly dispersed. Those pixels have lower intensity of black i.e. level of grayscale. Its intensity depends on their position i.e. distance from original center black pixel. Now, document image matrix is represented as grayscale image. Hence, intensity pertains in level region {0-255}. Black pixel of interest has coordinate $X(i,j)$ and intensity of 255, while neighbor pixels have around coordinates and intensity smaller than 255 i.e. grayscale level. So, after applying Gaussian kernel, equal to 2K+1 in x-direction as well as in y-direction, text is scattered forming enlarged area around it. Converting all non black pixels in the same area, as well as inverting image, forms the black pixel expanded areas. Those areas named boundary growing areas.

Boundary growing areas form control image with distinct objects that are prerequisite for document image text segmentation. These objects represent different text lines needful for text segmentation i.e. for disjoining text lines.

To further extend boundary growing area made some additional method is needed. Morphological dilatation is one of the possible solutions. It is given as [16]:

$$X \oplus S_2 \quad . \tag{5}$$

Dilatation structuring element $\mathbf{S}_2$ is used. It is a line defined by its height h=1, width w=2(R-K)+1 and parameter λ=R/K.

Main difference between original algorithm [12] and our approach are in text segmentation domain. In our approach, morphological dilatation extends boundary growing area leading to better text segmentation. Morphologically expanded boundary growing area is given in Fig.2.

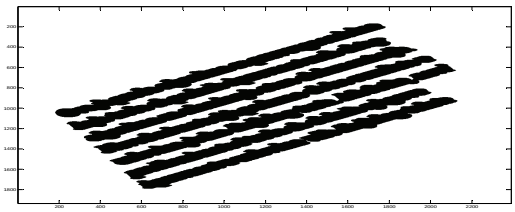


Fig.2. Morphologically expanded boundary growing areas (λ>1)

After segmenting text into separate text lines, primary task is text parameters extraction. Reference text line and skew

III. EXPERIMENTS

Algorithm quality examination consists of few text experiments representing test procedure. Basic and extended approach to algorithm is evaluated by combined text experiments framework. In this paper, only text line segmentation quality is examined. It is based on the following tests: multi-line text segmentation test, multi-line waved text segmentation test, and multi-line fractured text segmentation text. The schematic test procedure is shown in Fig.3.

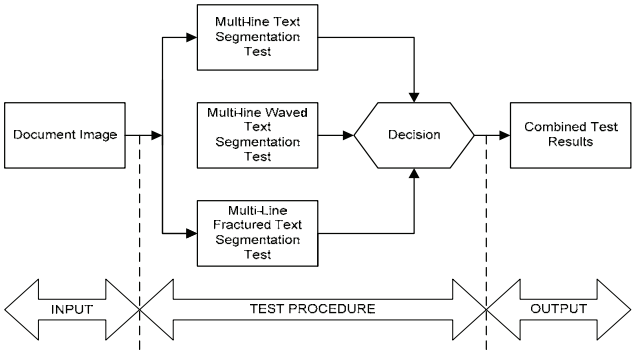


Fig.3. Schematic test procedure

The first experiment in combined text experiments framework is multi-line printed text sample [17]. This text with its skew angle parameter α is shown in Fig.4.

(a) (b)

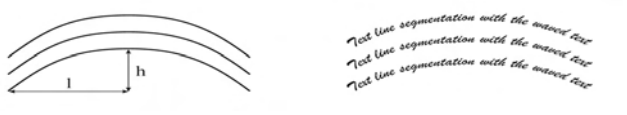
Fig.4. (a) Printed multi-text skew definition (α is parameter), (b) Printed multi-line text sample

Consequently, the number of existing text objects in multi-line text image relate to text segmentation quality success. Hence, the less objects the better segmentation process, except the number may not be less than text lines number. As a quality measure, the root mean square error $RMSE_{seg}$ has been used. It is calculated as [17,18]:

$$RMSE_{seg} = \sqrt{\frac{1}{P}\sum_{k=1}^{P}(O_{k,ref} - O_{k,est})^2} \quad , \qquad (6)$$

where k=1,…,P is the number of examined text samples, $O_{k,ref}$ is the number of referent objects in text i.e. number of text lines, and $O_{k,est}$ is the number of obtained objects in text by the applied algorithm.

The second text line segmentation experiment is based on multi-line waved text [17]. 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 ε = h/l. Typically, ε is used from the set {1/8,1/6,1/4,1/3,…}. Multi-line waved text sample for this experiment is shown in Fig.5.

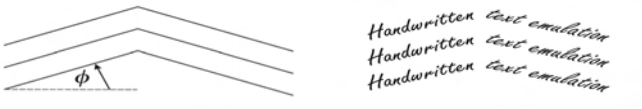


(a) (b)

Fig.5. (a) Waved text referent line shape definition (h and l are parameters), (b) Waved multi-line text sample

Similarly, the number of existing text objects after applied algorithm relate to the text segmentation quality success. Again, for the quality measure, *RMSE* has been used. Currently in (6) instead of RMS_{seg}, k, P, $O_{k,ref}$ and $O_{k,est}$ variables $RMSE_{seg,wav}$, l, R, $O_{l,ref}$ and $O_{l,est}$ are used, respectively. Alternatively, l=1,…,R is the number of examined text samples, $O_{l,ref}$ is the number of referent objects in text i.e. number of text lines, and $O_{l,est}$ is the number of obtained objects in text by the applied algorithm.

The last text line segmentation experiment is based on multi-line fractured text [17]. This text is formed by using the fractured referent 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°,10°,15°,20°}. Multi-line fractured text for the last segmentation experiment is shown in Fig.6.



(a) (b)

Fig.6. (a) Fractured text reference line slope definition (ϕ is parameter), (b) Fractured multi-line text

Again, the number of existing text objects relate to the text segmentation quality success. *RMSE* has been used as a quality measure. Currently in (6) instead of RMS_{seg}, k, P, $O_{k,ref}$ and $O_{k,est}$ variables $RMSE_{seg,frac}$, m, Q, $O_{m,ref}$ and $O_{m,est}$ are used, respectively. Alternatively, m=1,…,Q is the number of examined text samples, $O_{m,ref}$ is the number of referent objects in text i.e. number of text lines, and $O_{m,est}$ is the number of obtained objects in text by the applied algorithm.

In above experiments, printed text could be interchanged by handwritten text written on the defined shape referent line.

IV. RESULTS AND DISCUSSION

In the first experiment, character height $H_{ch} \approx 100$ px is used. From [19] parameter K value may not exceed 1/5 of H_{ch}. In fact, bigger K could lead to text lines merging. Number of objects inspection is given in Fig.7.

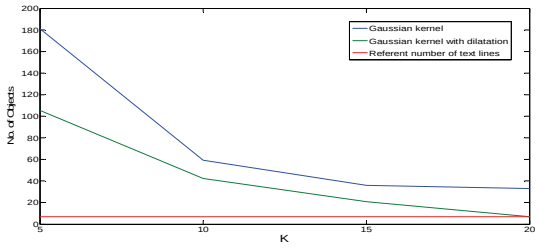


Fig.7. Number of objects from multi-text segmentation experiment

It is obvious that obtained results for extended algorithm are quite better than for original one. Further, in Fig.7 specific case is for K=20. Still, enlarging K above 20 is forbidden. Furthermore, K should be up to 20% of the H_{ch} to significantly close to boundary condition [18].

$RMS_{seg,wav}$ and $RMS_{seg,frac}$ is shown in Fig.8. and 9., respectively.

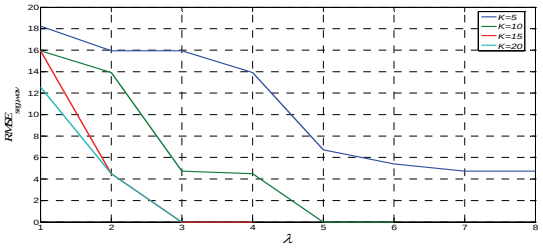


Fig.8. $RMS_{seg,wav}$ from waved text segmentation experiment

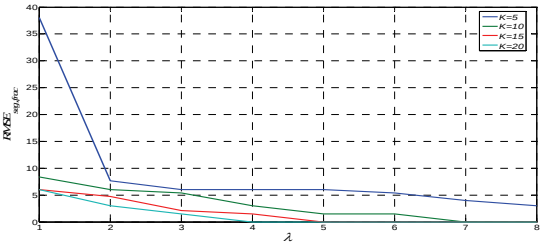


Fig.9. $RMS_{seg,frac}$ from fractured text segmentation experiment

Consequently, summary results for the original Gaussian kernel algorithm are not so convincing. Because of its faulty results from segmentation experiment, Gaussian kernel extended by morphological dilatation is promising.

V. CONCLUSION

In this paper, new approach to Gaussian kernel algorithm for text segmentation is presented. It assumes creation of boundary growing area around text based on Gaussian kernel algorithm extended by morphological dilatation. Those boundary growing areas form control image with distinct objects that are prerequisite for text segmentation. After text segmentation, reference text line and skew rate are calculated based on numerical method. Algorithm quality and robustness is examined by three experiments. Results are evaluated by RMSE method. All obtained results are compared with basic Gaussian kernel method.

Extended algorithm proved to be advanced in the domain of text segmentation which is of primary importance. Further investigation should be toward creating optimal adjusted and dilated Gaussian kernel rotated by the initial skew step.

REFERENCES

[1] A. Amin, S. Wu, "Robust Skew Detection in mixed Text/Graphics Documents", 8th International Conference on Document Analysis and Recognition (ICDAR 2005), Conference Proceedings, Seoul, Korea, 2005.

[2] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri, D. K. Basu, "Text Line Extraction from Multi-Skewed Handwritten Documents", Pattern Recognition, vol. 40, pp. 1825-1839, 2006.

[3] Likforman-Sulem, L., Zahour, A., Taconet, B., "Text Line Segmentation of Historical Documents: A Survey", International Journal on Document Analysis and Recognition (IJDAR), vol.9(2-4), pp.123-138, 2007.

[4] Silva, L. F., Conci, A., Sanchez, A., "Automatic Discrimination between Printed and Handwritten Text in Documents", In Proceedings of XXII Brazilian Symposium on Computer Graphics and Image Processing, pp. 261-267, Rio de Janeiro, Brazil, 2009.

[5] Ballard, D. H., "Generalizing the Hough Transform to Detect Arbitrary Shapes", Pattern Recognition, vol.13(2), pp. 111-122, 1981.

[6] Amin, A., Fischer, "A Document Skew Detection Method Using the Hough Transform", Pattern Analysis & Applications, vol.3(3), pp. 243-253, 2000.

[7] Shi Z., Govindaraju V., "Line Separation for Complex Document Images Using Fuzzy Runlength", In Proceedings of the International Workshop on Document Image Analysis for Libraries, Palo Alto, U.S.A., 2004.

[8] Likforman-Sulem, L., Faure, C., "Extracting Lines on Handwritten Documents by Perceptual Grouping, in Advances in Handwiting and drawing: a multidisciplinary approach", Faure, C., Keuss, P., Lorette, G., Winter A. (Eds.), pp. 21-38, Europia, Paris, 1994.

[9] Zahour, A., Taconet, B., Mercy, P., Ramdane, S., "Arabic hand-written text-line extraction", In Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR'01), pp. 281-285, Seattle, U.S.A., 2001.

[10] Koshinaka, T., Ken'ichi, I., Akitoshi, O., "An HMM-based Text Segmentation Method using Variational Bayes Approach", IEIC Technical Report, vol.104(87), pp. 19-24, 2004.

[11] Brodić, D., Milivojević, Z., "An Approach to Modification of Water Flow Algorithm for Segmentation and Text Parameters Extraction", In Emerging Trends in Technological Innovation. Camarinha-Matos, L.M., Pereira, P., Ribeiro, L., (Eds.), IFIP AICT, vol.314, pp. 324-331, Springer, Boston, U.S.A., 2010.

[12] Y. Li, Y. Zheng, D. Doermann, S. Jaeger, "A New Algorithm for Detecting Text Line in Handwritten Documents", 18th International Conference on Pattern Recognition (ICPR 2006), Conference Proceedings, vol.2, pp. 1030-1033, Hong Kong, China, 2006.

[13] Otsu, N., "A threshold selection method from gray-level histograms", IEEE Transactions on Systems, Man, and Cybernetics, vol. 9(1), pp. 62–66, 1979.

[14] R. C. Gonzalez, R. E. Woods, Digital Image Processing, 2nd ed., New Jersey, Prentice-Hall, 2002.

[15] M. Sonka, V. Hlavac, R. Boyle, Image Processing, Analysis and Machine Vision, Toronto, Thomson, 2008.

[16] Y. F. Shih, Image Processing and Mathematical Morphology – Fundamentals and Applications, Boca Raton, CRC Press, Taylor & Francis Group, 2009.

[17] Brodić D., Milivojević D. R., Milivojević Z., "Basic Test Framework for the Evaluation of Text Line Segmentation and Text Parameter Extraction", Sensors, vol.10(5), pp. 5263-5279, 2010.

[18] Bolstad, W. M., Introduction to Bayesian Statistics, John Wiley & Sons, NJ, U.S.A

[19] Brodić D., Dokić B., "Initial Skew Rate Detection Using Rectangular Hull Gravity Center", 14th International Conference on Electronics (E2010), Vilnius, Lithuania, 2010.