Optimization of the Anisotropic Gaussian Kernel for Text Segmentation and Parameter Extraction

Darko Brodić

University of Belgrade, Technical Faculty Bor, V.J. 12, 19210 Bor, Serbia
dbrodic@tf.bor.ac.rs

Abstract. In this paper, extended approach to Gaussian kernel algorithm for text segmentation, reference text line and skew rate extractions is presented. It assumes creation of boundary growing area around text based on Gaussian kernel algorithm extended by anisotropic approach. Those boundary growing areas form control image with distinct objects that are prerequisite for text segmentation. After text segmentation, text parameters such as reference text line and skew rate are calculated based on numerical method. Algorithm quality is examined by experiments. Results are evaluated by RMS method. Obtained results are compared with isotropic Gaussian kernel method. All results are examined, analyzed and summarized. Furthermore, optimal parameter values are suggested leading to anisotropic kernel optimization.

Keywords: Image processing, Document image processing, OCR, Text segmentation, Text parameters extraction, Isotropic Gaussian kernel

1 Introduction

Printed and handwritten text is characterized by its attributes and features diversity. Hence, text parameter extraction procedure can be quite dissimilar one. However, its algorithm should be valid for printed as well as for handwritten text. To finish the task efficiently algorithm should be robust enough as well.

Prior to text parameters extraction, text line segmentation should be done. It is an important step in document processing. Although some text line detection techniques are successful in printed documents, processing of handwritten documents has remained a key problem in optical character recognition (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 image analysis [3].

Later, text parameters extraction from scanned documents is primary OCR goal. Reference text line and skew rate identification is mandatory. Their validity is of major importance for OCR process. Various reasons exist for appearance of multi-skewed lines in text, but two are most common [1]. Firstly, during scanning a misalignment degrees of the document made is unavoidable. All printed

text lines in the scanned document are uniformly skewed. Secondly, text lines in original document are differently skewed due to individual handwriting. Hence, text lines are made under different orientation. To enhance the ability of document analysis system, robust algorithm for text segmentation and parameters extraction is needed.

Previous work on text line segmentation can be categorized in few directions [2]: histogram analysis, k-nearest neighbor clustering, projection profile, Fourier transform, cross-correlation and other models.

In [1] is mentioned previously proposed and accepted technique of reference line extraction based on identifying valleys of horizontal pixel density histogram. It failed due to multi skewed text lines. Hence, it is not suitable for handwriting text.

K-nearest neighbor clustering method [2] is by product of a larger page layout analysis system, which assumed only text is being processed. The connected components formed by the nearest neighbors clustering are characters based only. Method is suitable for finding skew angle, but it is limited to Roman languages [2].

Method in [4] deals with "simple" multi skewed text. It uses as a basis Hough transform for straight lines. But, it is too specific and computationally expensive.

Method of identifying words contour area as a start of detecting baseline is proposed in [5]. The assumptions made on the word elements definition are specific.

Method [1] hypothetically assumed a flow of water in a particular direction across image frame in a way that it faces obstruction from the characters of the text lines. This method is adopted in [6]. To be totally robust it needs some further adaptation.

Algorithm proposed by [7] model text line detection as an image segmentation problem by enhancing text line structure using a Gaussian window and adopting the level set method to evolve text line boundaries. Method is specified as robust, but rotating text by an angle of (from -10° to 10°) has significant impact on it. In the paper, modification of this method is proposed, analyzed [8] and compared. Algorithm is evaluated by different sample text examples. Furthermore, optimal parameter values are suggested leading to anisotropic Gaussian kernel optimization.

Organization of the paper is as follows. In Section 2 information on proposed Gaussian kernel algorithm is given. In Section 3 experiments are described. Obtained results are analyzed, examined and discussed as well. In Section 4 conclusion is made.

2 Proposed Algorithm

Although document conversion system incorporates scanning, binarization, region segmentation, text recognition and document analysis, its procedure can be divided into three main stages as shown in Fig. 1.

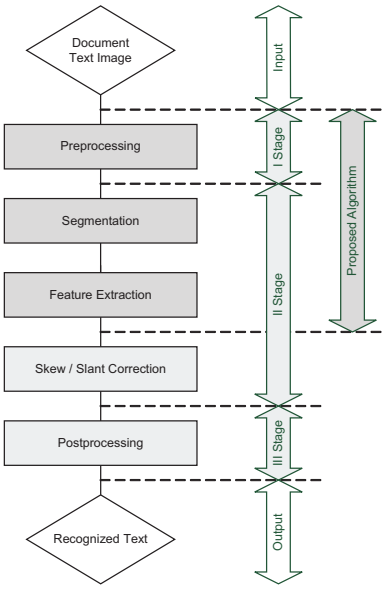

Fig. 1. Document text image identification procedure

In the first stage, algorithm for document text image binarization and normalization is applied. After preprocessing stage, text is prepared for segmentation, feature extraction and character recognition. During the second stage, algorithm for text segmentation as well as for skew and reference text line identification is enforced. Also, reference text based on skew and stroke angle, is straightened and repaired. Finally, in third stage character recognition process is applied. As a result of scanning process, document text image is obtained. It is an input of text grayscale image described by following intensity matrix [9]:

$$D(i,j) \in [0,\ldots,255] \ , \tag{1}$$

where $i \in [1,\ldots,M]$ and $j \in [1,\ldots,N]$. After applying intensity segmentation with binarization, intensity function is converted into binary form given by:

$$X(i,j) = \begin{cases} 1 & \text{for } \ D(i,j) \geq D_{\text{th}} \\ 0 & \text{for } \ D(i,j) < D_{\text{th}} \ , \end{cases} \tag{2}$$

where D_{th} is given by Otsu algorithm [10]. It represents the threshold sensitivity decision value.

Document text image is black and white image represented by matrix $\mathbf{X}$. Each character or word consists of the only black pixels. Hence, every point $X(i,j)$ i.e. $X_{i,j}$ is represented by number of coordinate pairs $(0,1)$, where $i =$

$1, \ldots, M$, and $j = 1, \ldots, N$ of matrix $\mathbf{X}$ [9]. It is represented by document text image fragment as in Fig. 2.

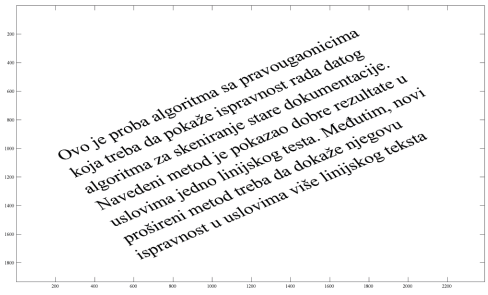

Fig. 2. Document text image fragment represented by matrix $\mathbf{X}$

2.1 Morphological Preprocessing

Prior to processing stage, document text image needs additional preparation. It is assumed text area is extracted by any appropriate method. Further, morphological preprocessing is performed to make document text image "noiseless". The morphological preprocessing was defined in [11] by following steps:

- Document image erosion by $\mathbf{X} \ominus \mathbf{S}_1$,
- Document image opening by $\mathbf{X} \circ \mathbf{S}_1$,
- Dilatation of the opening the document image by $(\mathbf{X} \circ \mathbf{S}_1) \ominus \mathbf{S}_1$,
- Closing of the opening the document image by $(\mathbf{X} \circ \mathbf{S}_1) \bullet \mathbf{S}_1$.

For the above operations, structuring element $\mathbf{S}_1$ dimension 3x3 is used [9].

2.2 Linear Bounding Containers

All text parameters extraction algorithms more or less depend on resolution and size of text letters. Consequently, algorithm's parameters are closely related to it. To be efficient, algorithm should choose optimal parameters from the entire set. Linear containers or its modifications are one of the tools for letter size estimation [11].

Linear container and its interior are specified by a finite number of linear inequalities. In our case linear container is assumed to be special case of the convex polygon i.e. bounding box. Special case of the box is a rectangular region whose edges are parallel to the coordinate axes. Furthermore, it is defined by its

maximum and minimum extents for all axes. Hence, each pixel $X_{i,j}$ belonging to box is given by:

$$x_{\min} \leq i \leq x_{\max} \ , \tag{3}$$

and

$$y_{\min} \leq j \leq y_{\max} \ . \tag{4}$$

Hence, bounding box is defined by its endpoints $x_{\min}$, $x_{\max}$, $y_{\min}$, $y_{\max}$. Inclusion of the point $X_{i,j}$ in a box is tested by verifying these four inequalities. If any one of them fails, then the point is not inside. Bounding box definition is given in Fig. 3.

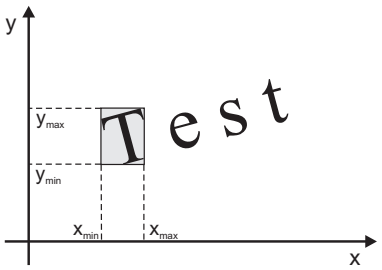

Fig. 3. Bounding box definition

The bounding box is the computationally simplest of all linear containers. Hence, it is one of the most frequently used in many applications due to its simplicity and computationally inexpensiveness [12]. Although the bounding box is not precise method for letter size estimation, it is simple and in many cases adequate to evaluate those values. All text elements like letters, part of words or words are surrounded by bounding boxes. Their heights represent the height of letters. To reduce the error, median height of all bounding boxes is used. Median is the middle value in a set of values. Hence, it is less sensitive to extreme values. Boundary box objects are defined by O_u, where $u = 1, \ldots, V$ and V is the total number of boundary box objects over sample text. If V is the number of objects and u is object index, reorder initial set of values h_u so that $g_1 < g_2 < \ < g_u$ and currently g_u is called u-th order statistic [13]. Hence, following is valid: $g_1 \equiv min(h_u)$ and $g_V \equiv min(h_u)$. Median is defined by [13] as:

$$h_{\text{median}} = \begin{cases} g_{\frac{(u+1)}{2}} & \text{if } u \text{ is odd} \\ \\ \frac{1}{2}(g_{\frac{u}{2}} + g_{1+\frac{u}{2}}) & \text{if } u \text{ is even} \ , \end{cases} \tag{5}$$

After h_{median} of the letter heights set is obtained, typical letter height is annotated. It is prerequisite for algorithm's parameter optimization decision.

2.3 Anisotropic Gaussian Kernel

For the processing stage Gaussian kernel algorithm is used. It is based on two-dimensional Gaussian function given by [14]:

$$f(x,y) = Ae^{-\frac{(x-b_x)^2+(y-b_y)^2}{2\sigma^2}}, \tag{6}$$

where b_x is shift along x-axis, b_y is shift along y-axis, σ is standard deviation defining curve spread parameter and A is the amplitude of the function given as $A = 1/(2\pi\sigma^2)^{\frac{1}{2}}$. From (10) 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. Hence, 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. These pixels have lower intensity of black. Hence, they are grayscale. Their intensity depends on position or distance from original black pixel. Now, document image matrix is represented as grayscale image. Intensity pertains in level region $(0-255)$. Our black pixel of interest has coordinate $X_{i,j}$ and intensity of 255, while neighbor pixels have intensity smaller than 255. So, after applying Gaussian kernel, equal to $2R+1$ 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.

In our case isotropic approach is less efficient. Alternatively in some cases, it is not suitable due to its possibility of merging different text lines. Using different curve spread parameter σ in x and y direction i.e. σ_x and σ_y respectively, extends (10) as follows [13,14]:

$$f(x,y) = Ae^{-\left[\frac{(x-b_x)^2}{2\sigma_x^2}+\frac{(y-b_y)^2}{2\sigma_y^2}\right]}. \tag{7}$$

This extension of the previous i.e. isotropic Gaussian function lead to image kernel equal to $2S+1$ in x-direction and $2R+1$ in y-direction. Due to relations $R \neq S$, Gaussian kernel is anisotropic [8]. Additionally, ratio parameter $\lambda = S/R$ completely defines Gaussian anisotropic condition. Example of the isotropic and anisotropic Gaussian kernel for $R = 10$, $\lambda = 1$ and $\lambda = 5$ is given in Fig. 4.

Created boundary growing areas form control image with objects that are prerequisite for document image text segmentation [8]. These black objects represent different text lines needful for text segmentation. Hence, their basic task is text lines splitting. Example of the boundary growing areas is given in Fig. 5.

2.4 Reference Text Line

After text segmentation, primary task is reference text line and skew rate extraction. Their identification is based on information obtained from black pixel contained in boundary growing areas. Reference text line estimation is average

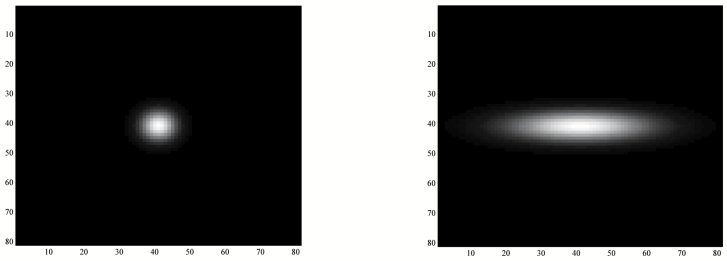

Fig. 4. Gaussian kernel: isotropic (left), anisotropic $\Rightarrow \lambda > 1$ (right)

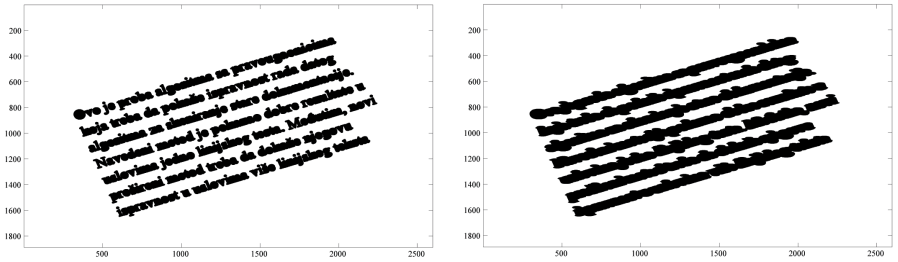

Fig. 5. Boundary growing areas over text: isotropic (left), anisotropic (right)

position calculation of black pixels in every column of document image. It is calculated as [1,9]:

$$x_i = \frac{\sum_{j=1}^{L} y_j}{L} \quad {}_{i=1,\ldots,K} \ , \tag{8}$$

where x_i is point position of calculated reference text line, i is number of column position of calculated reference text, y_j is position of black pixel in column j and L is sum of black pixel number in specified column j of an image.

After calculation, sub-image matrix with only one black pixel per column is obtained. This black pixel defines estimated reference text line. Such reference text line forms continuous or discontinuous line partly "representing" reference text line. To form continuous reference text line, some numerical method could be used.

3 Experiments, Results and Discussion

Algorithm quality evaluation consists of two text experiments. First experiment represents text segmentation estimation. It is inevitable in algorithm segmentation quality assessment. Consequently, it is prerequisite for obtaining other text parameters. If segmentation experiment miscarry, other examination process will

be meaningless. Hence, its importance is critical. Second experiment is mainly concerned with skew rate identification. Its task is algorithm performance evaluation of the skew rate tracking succeed. This experiment is primarily based on printed text, but it is good prerequisite for testing handwritten text. Obtained results are linked.

3.1 Text Segmentation Experiment and Results

In the first experiment text segmentation quality is examined. For this purpose multi line text is used. It is given in Fig. 6.

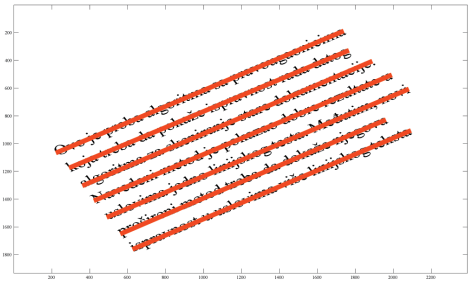

Fig. 6. Sample text for text segmentation experiment

Number of existing text objects 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. The quality is measured by RMS_{seg} calculated as [13,14]:

$$RMS_{\text{seg}} = \sqrt{\frac{1}{Q}\sum_{i=1}^{Q}(O_{i,\text{ref}} - O_{i,\text{est}})^2} \ , \tag{9}$$

where $i = 1, \ldots, Q$ is the number of examined text samples, $O_{i,\text{ref}}$ is number of referent objects in text i.e. number of text lines, and $O_{i,\text{est}}$ is number of obtained objects in text by the applied algorithm for each examined text sample.

Character height $H_{\text{ch}} \approx 100$ pixels (px) obtained as h_{median} is assumed. From [11] parameter R value may not exceed 20% of H_{ch}. In fact, bigger R could lead to text lines merging. Hence, algorithm is examined for $R = (5, 10, 15, 20)$. First part of experiment examined text segmentation through RMS_{seg} for angle range $(0° - 30°)$, while further angle range $(0° - 80°)$ is used. Results are given in Table 1 and Table 2.

From Table 1 and 2, it is obvious that anisotropic approach is advantageous. Nevertheless, it is true for $\lambda > 1$. Under this condition, kernel is stretched in x-direction by parameter λ. Furthermore, segmentation experiment proved to be

Table 1. Text segmentation results: RMS_{seg1} for angle range $(0^\circ - 30^\circ)$ (less values are better ones)

Angle Range	λ	R			
$0^\circ - 30^\circ$	S/R	5	10	15	20
Isotropic	1	180.36	65.77	36.00	32.82
Anisotropic	2	79.86	48.15	3.63	0.00
	3	63.01	14.68	0.44	-
	4	55.40	9.21	-	-
	5	40.73	8.70	-	-

Table 2. Text segmentation results: RMS_{seg2} for angle range $(0^\circ - 80^\circ)$ (less values are better ones)

Angle Range	λ	R			
$0^\circ - 80^\circ$	S/R	5	10	15	20
Isotropic	1	181.05	59.22	36.00	33.01
Anisotropic	2	105.34	42.52	20.87	6.90
	3	70.07	25.18	9.20	-
	4	59.18	15.70	-	-
	5	48.94	11.34	-	-

eligible for optimal parameters selection such as R and λ. This way, each parameter R can be paired with optimal parameter λ. From the above tables those pairs (R, λ) are following: $(5, 5), (10, 5), (15, 3), (20, 2)$ as well as $(5, 4), (10, 4), (10, 3)$, and $(15, 2)$. However, the best choice is $(20, 2)$ from Table 1. All listed paired values are invaluable for further examination process i.e. for other text experiments.

3.2 Skew Rate Text Experiment and Results

Second experiment is mainly concerned with text skew rate. It examines algorithm quality to follow text skewing. In this case, sample printed text rotated by angle β up to 80° by step of 5° around x-axis is used. This is given in Fig. 7.

Reference line of the sample text is represented by:

$$y = ax + b \; . \tag{10}$$

After applying algorithm to sample text, reference text line is calculated by (8). To achieve continuous linear reference text line, least square method is used. Function approximation by first degree polynomial is given as:

$$y = ax' + b' \; . \tag{11}$$

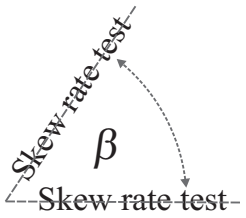

Fig. 7. Sample text rotating for the angle β up to 80°

Then number of data points i.e. ndp is used and the slope a', and the y-intercept b' are calculated as [14]:

$$a' = \frac{\sum y \sum xy - ndp \sum xy}{(\sum x)^2 - ndp \sum x^2} \ , \tag{12}$$

and

$$b' = \frac{\sum x \sum xy - \sum y \sum x^2}{(\sum x)^2 - ndp \sum x^2} \ . \tag{13}$$

Further, referent line hit rate i.e. $RLHR$ is defined as [8]:

$$RLHR = 1 - \frac{\beta_{\text{ref}} - \beta_{\text{est}}}{\beta_{\text{ref}}} \ , \tag{14}$$

where β_{ref} is arctangent from origin (10) i.e. a and β_{est} is arctangent from calculated i.e. estimated (11) i.e. a'. Then, RMS_{skew} values are calculated by [13,14]:

$$RMS_{\text{skew}} = \sqrt{\frac{1}{P} \sum_{i=1}^{P} (x_{i,\text{ref}} - x_{i,\text{est}})^2} \ , \tag{15}$$

where $i = 1, \ldots, P$ is number of examined text rotating angles up to 80°, $x_{i,\text{ref}}$ is $RLHR$ for β_{est} equal to β_{ref}, due to normalization equal to 1, and $x_{i,\text{est}}$ is $RLHR$.

Again, $H_{\text{ch}} \approx 100$ px is assumed. Algorithm is examined for $R = (5, 10, 15, 20)$ and $\lambda = (1, 2, 3, 4, 5)$. Quality of skew rate identification is obtained by $RLHR$ value as in (14). Furthermore, level of spreading results is obtained by RMS_{skew} value given in (15) for two angle ranges: $(0° - 30°)$ and $(0° - 80°)$. These results are given in Fig. 8. Isotropic results are given on the left side of the chart on Fig. 8 - 11. Unlike, anisotropic results are shown in the rest of the each chart.

It can be noted, anisotropic approach leads to quite better results. Still, it should be cautious on interpreting presented results. Namely, high S values should be avoided. These values contribute to faulty text segmentation process. So, it is recommended to match results from this experiment and previous one. This way, parameter pairs (R, λ) obtained from previous experiments is optimized ones. Still, from the optimized parameter pairs, set members

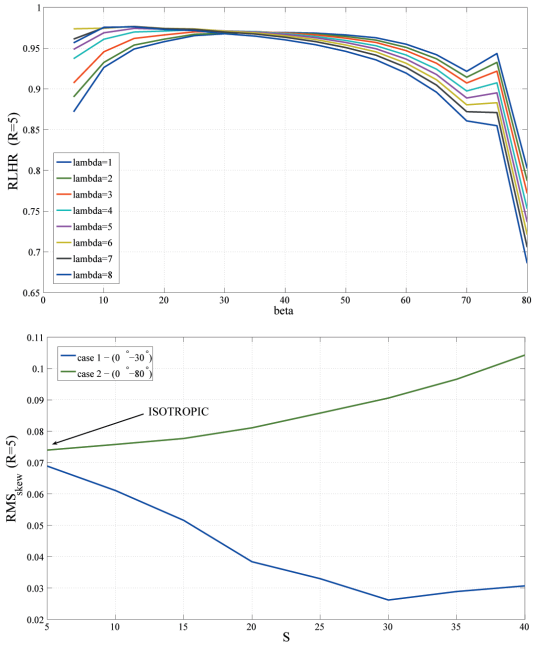

Fig. 8. Skew rate test for R=5, λ=$(1,\ldots,8)$: $RLHR$ (top), RMS_{skew} (bottom)

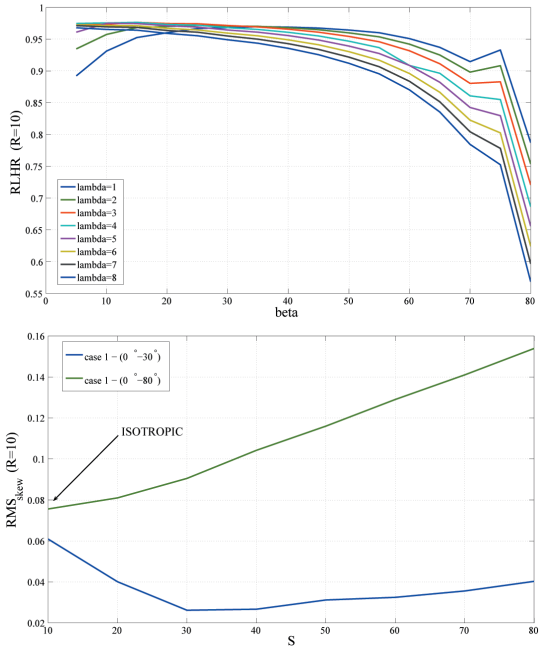

Fig. 9. Skew rate test R=10, λ=$(1,\ldots,8)$: $RLHR$ (top), RMS_{skew} (bottom)

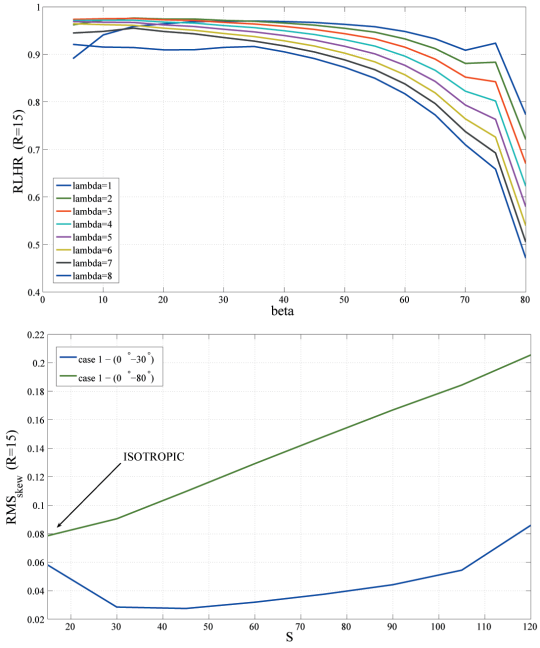

Fig. 10. Skew rate test R=15, λ=$(1,\ldots,8)$: $RLHR$ (top), RMS_{skew} (bottom)

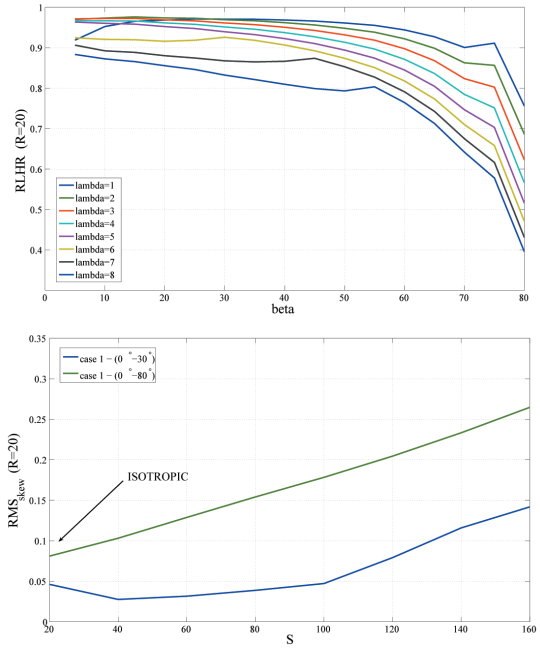

Fig. 11. Skew rate test R=20, λ=$(1,\ldots,8)$: $RLHR$ (top), RMS_{skew} (bottom)

$(R, \lambda) = (15, 45)$ and $(20, 40)$ are the best ones for $H_{\text{ch}} \approx 100$ px. Hence, from the previous statements and claims, it can be concluded on optimal parameter values of R and λ. These optimal values are $K \approx$ 15-20% of H_{ch} as well as $\lambda \approx$ 2-3, leading to anisotropic approach.

4 Conclusions

Anisotropic Gaussian kernel algorithm proved to be advanced in the domain of text segmentation, which is of primary importance. Due to isotropic approach faulty results from segmentation experiment, all it's slightly better results in skew rate experiments are completely irrelevant. Still, text segmentation is primary goal. It is prerequisite for reference text line and skew rate identification. Slightly weaker results of anisotropic approach in some part of text parameter estimation are in the background. This way, anisotropic Gaussian kernel and its optimized parameter pairs proved to be useful and robust method which is promising. Consequently, for higher angles some modification of the anisotropic approach is recommended.

References

1. Basu, S., Chaudhuri, C., Kundu, M., Nasipuri, M., Basu, D.K.: Text Line Extraction from Multi-Skewed Handwritten Documents. *Pattern Recognition*, Vol.40, pp. 1825–1839, 2006.
2. Amin, A., Wu, S.: Robust Skew Detection in Mixed Text/Graphics Documents. In: *Proceedings of ICDAR'05*, pp. 247–251, Seoul, Korea, 2005.
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, No.2-4, pp. 123–138, 2007.
4. Louloudis, G., Gatos, B., Pratikakis, I., Halatsis, C.: Text Line Detection in Handwritten Documents. *Pattern Recognition*, Vol.41, pp. 3758–3772, 2008.
5. Wang, J., Mazlor, K.H.L., Hui, S.C.: Cursive Word Reference Line Detection. *Pattern Recognition*, Vol.30, No.3, pp. 503–511, 1997.
6. 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, Heidelberg, 2010.
7. Li, Y., Zheng, Y., Doermann, D., Jaeger, S.: A New Algorithm for Detecting Text Line in Handwritten Documents. In: *Proceedings of* 18^{th} *International Conference on Pattern Recognition*, Vol.2, pp. 1030–1033, Hong Kong, China, 2006.
8. Brodić, D., Milivojević, Z.: Using Anisotropic Gaussian Window for Printed and Handwritten Text Parameters Extraction. In: *Proceedings of* 9^{th} *International Scientific Conference UNITECH*, Vol.1, pp. 453–460, Gabrovo, Bulgaria, 2009.
9. Gonzalez, R.C., Woods, R.E.: *Digital Image Procesing,* 2^{nd} *ed.* Prentice-Hall, Eaglewoods Cliffs, New Jersey, 2002.
10. Draganov, I.V, Popova, A.A.: Rotation Angle Estimation of Scanned Handwritten Cursive Text Documents. In: *Proceedings of ICEST '06*, Sofia, Bulgaria, 2006.

11. Brodić, D., Dokić, B.: Initial Skew Rate Detection Using Rectangular Hull Gravity Center. In: *Proceedings of* 14^{th} *International Conference of Electronics*, Kaunas, Lithuania, 2010.
12. Chang, C.M.: Detecting Ellipses via Bounding Boxes. *Asian Journal of Health and Information Sciences*, Vol.1, No.1, pp. 73–84, 2006.
13. Qui, P.: *Image Processing and Jump Regression Analysis.* John Wiley & Sons, New Jersey, 2005.
14. Bolstad, W.M.: *Introduction to Bayesian Statistics.* John Wiley and Sons, New Jersey, 2005.