An Approach to the Estimation of Global and Local Text Skew in Historical Printed Documents

Darko Brodić * **Zoran N. Milivojević** **

* *Technical Faculty in Bor, University of Belgrade, Vojske Jugoslavije 12, 19210 Bor, Serbia, (e-mail: dbrodic@tf.bor.ac.rs, cmaluckov@tf.bor.ac.rs)*
** *College of Applied Technical Sciences, Aleksandra Medvedeva 20, 18000 Niš, Serbia (e-mail: zoran.milivojevic@vtsnis.edu.rs)*

Abstract: Historical printed documents represent an important part of our heritage. In order to preserve their content, digitization process is mandatory. As a possible result of this process, the document image can be generated with some degree of inclination that can affect the text by creating skewed text lines. This paper proposes a multi step approach to the estimation of the global and local text skew for historical printed documents. It is based on the analysis of the connected components created by the filled convex hulls around each text element. Then, the connected components are enlarged by oriented morphological erosion implemented in the complementary image. To identify the initial skew rate, the largest connected components are determined. Furthermore, the process of erosion and largest component determination is repeated with the oriented morphological erosion obtained by the initial skew rate. The orientation of the largest connected components estimates the global text skew of the document. Then, the original document is rotated according to the estimated angle. Furthermore, the text line segmentation is established with striking through each text line as the result of horizontal projection profile's calculation. This way, it creates a connected component, which orientation represents the local text skew identified by the least square method. Efficiency and correctness of the algorithm are examined by testing on the dataset. The results prove the robustness of the algorithm.

Keywords: Document image analysis, Connected components analysis, Least square method, Projection profiles method, Skew estimation.

1. INTRODUCTION

The historical printed documents represent an important part of our heritage. Due to their age, it is quite common that they suffer from different types of degradation. In an optical character recognition (OCR) system, the quality of the input image is a key element which affects the final performance. A variety of interfering effects such as noise and skewing can appear during the scanning process. These distortions decrease the performance of the system. Furthermore, OCR system is very sensitive to any skew appearance in a document image. Hence, skew detection is the key element in document image analysis and processing (Rehman and Saba (2011)).

In this paper, historical typewritten documents are in focus. They are characterized with regularity in shape as any other printed documents (Amin and Wu (2005)). Accordingly, it follows that the letters are characterized with similar sizes. Furthermore, the distance between text lines is decent. It means that the orientation of text lines is almost equal. All aforementioned constitutes the text skew as a uniform, which is commonly called global text skew. However, the old documents include paper stretching. It contributes to non-uniformity of text orientation. Hence, each text line has a slightly different skew, which represents the local text skew.

In this paper, we proposed the method for the estimation of the global text skew in historical documents made by typewriter. Due to the age, these documents incorporate the stretching of the paper. Hence, we extended the method to estimate the local text skew as well.

Organization of this paper is as follows. Section 2 briefly reviews previous works. Section 3 describes a proposed algorithm. Section 4 defines text experiments. Section 5 gives the results and discusses them. Section 6 makes conclusions.

2. PREVIOUS WORKS

Text skew occurrence is unavoidable in the process of document scanning. To detect inherited skew, a large amount of techniques has developed. They are classified as Amin and Wu (2005), Rehman and Saba (2011):

⋆ This work was partially supported by the Grant of the Ministry of Education, Science and Technological Development of the Republic Serbia, as a part of the project TR33037 within the framework of Technological development program.

1. Projection profile's methods,
2. Nearest neighbor clustering methods,
3. Hough transforms methods,
4. Fourier's transformation methods,
5. Cross-correlation methods,
6. Other methods.

Many of the aforementioned methods have some strong points. However, they incorporate the weaknesses as well.

Projection profile method is a straightforward method. It identifies the peaks in the histogram, which represent horizontal black pixel density (Postl (1986)). The profile with a maximum variation refers the best alignment to the text lines. However, this method is suitable for text with uniform skew only (Postl (1986)).

The nearest neighbor clustering method is based on the page layout analysis (O'Gorman (1993)). It exploits the assumption that characters in a line are close to each other and mutually aligned. However, this method cannot handle incorporation of noisy subparts in text, which leads to reduced accuracy.

The Hough transforms is a widely employed method for detecting lines and curves in digital images (Srihari and Govindaraju (1989)). This approach uses the fact that the highest number of co-linear pixels on lines co-incident with the baseline of the text (Amin and Fischer (2000)). It needs a preprocessing stage, which defines candidate mapping points (Singh et al. (2008)). Still, the method is complex and computer time intensive.

The Fourier transforms method is even more complex (Postl (1986)). In this method, the document image is transformed into the power spectrum of the Fourier domain. Then, the position where the density of Fourier's domain has the maximum value is detected as a text skew. The method can recognize the skew without knowing what is in the analyzed document image (text, image or their combination). Still, it is computationally expensive, and limited to the error of around 0.25° (Lowther et al. (2002)).

The cross-correlation method calculates the cross-correlation between two lines in the image with a fixed distance (Yan (1993)). In this way, the cross-correlation is established between the rotating line with known angle and some part of text representing a text line. Hence, the cross-correlation function is obtained for all pairs of rotating lines, which are accumulated. The position where the accumulated cross-correlation function has a peak determines the text skew. This is a position where a certain line is the best overlapped with the text line. The extension of the method is proposed in (Brodic et al. (2013)). This approach is different because it calculates the cross-correlation between the original image and the reference image in the log-polar domain. The maximum of the cross-correlation function which depends on the angle, determines the text skew. The method gives the correct results with the error just around 0.1°. However, the correctness of the results heavily depends on the choice of the center point of the log-polar transformation. Hence, its application is unstable in some circumstances (Brodic et al. (2013)). In (Makridis et al. (2010)), preprocessing of document image is made by complex decision making. Global text skew is identified with the cross-correlation method applied to remain connected components. At the end, the local text skew is calculated by the least square method.

The other methods combine different techniques for text skew detection. Many of them include morphological processing. It includes the combination of morphological opening and closing (Das and Canda (2001)), and morphological dilation and erosion (Dhandra et al. (2006)). These methods proved to estimate correctly the text skew with the smallest absolute error values.

Furthermore, some of the methods employ the moments based technique. The moments measure the pixel distribution in the image (Shivakumara et al. (2005)). Accordingly, they are sensitive to the rotation. Hence, this approach is useful for text skew identification. However, this technique is suitable only for uniform skew estimation (Flusser et al. (2009); Brodic and Milivojevic (2012)).

Ref. (Chou et al. (2007)) proposes a method based on geometrical transformation. It detects a dominant skew in document images using piecewise covering of document objects by parallelograms. The method is very accurate, but the skew angle range is limited to 15°.

In this paper, the method proposed in (Brodic et al. (2013)) is enlarged by introducing the geometrical preprocessing. This way, the redundant data for the text skew estimation are filtered out. As a consequence, the quantity of the connected components from different text lines and the noise elements in a document image are reduced. It leads to lower errors of estimated global text skew. We found that the proposed method is a quite immune to noise, which appears in the document image. Furthermore, the historical printed documents are exposed to stretching of the paper. Hence, the local text skew represents a small variation in each text line. The proposed algorithm can detect these text line variations. Hence, the main contribution of the proposed algorithm is its ability to correctly estimate the global and local text skew in the historical printed documents made by typewriter.

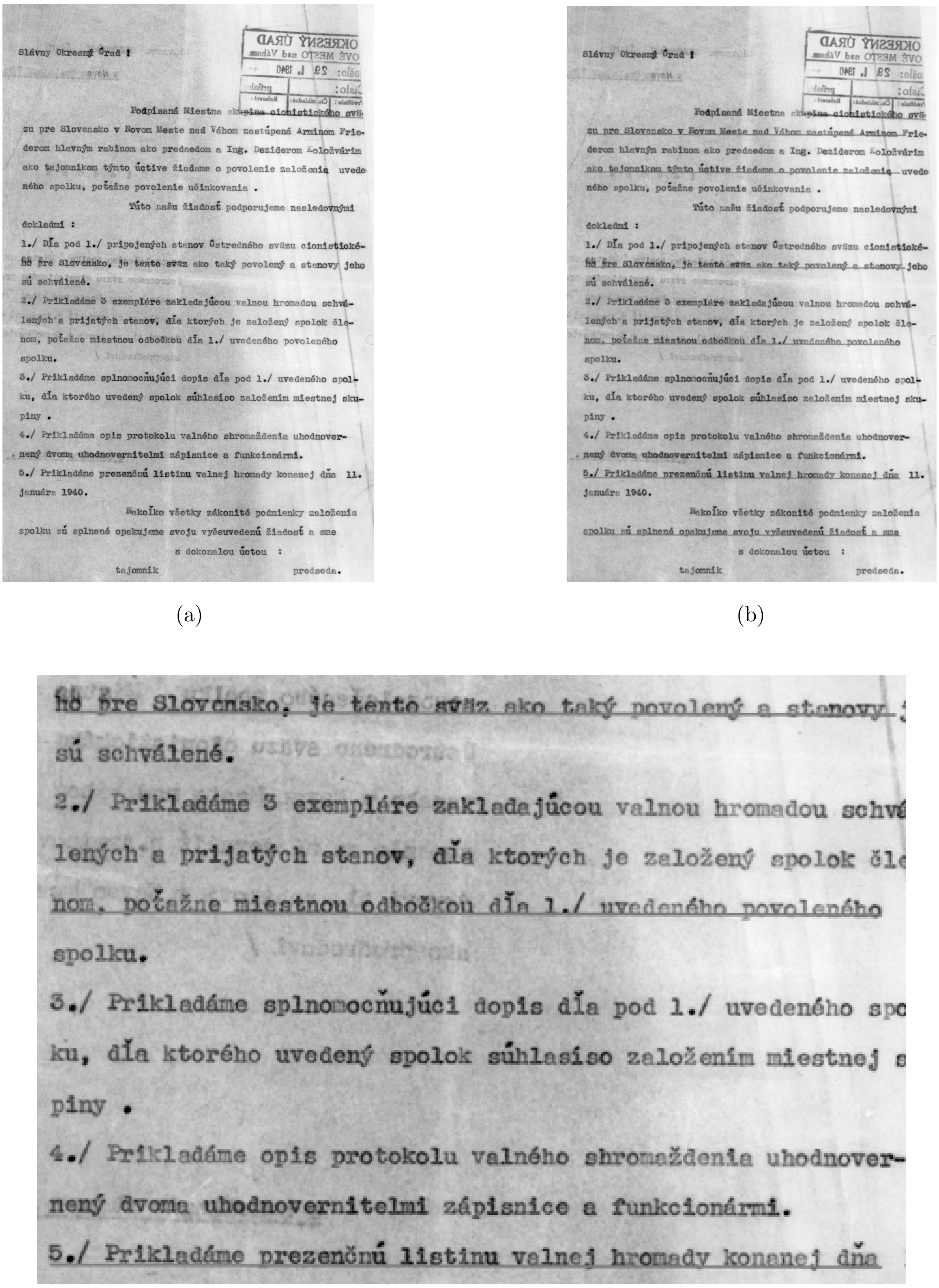
3. THE PROPOSED ALGORITHM

The proposed algorithm identifies the global and local text skew of the printed documents. It consists of the following steps:

1. Bi-level thresholding,
2. Geometrical preprocessing,
3. Convex hulls extraction,
4. Initial skew rate detection,
5. Morphology based object grouping,
6. Extraction of the longest object,
7. Global skew estimation based on the least square method,
8. Global de-skewing of the document,
9. Vertical projection profiles of de-skewed document,
10. Grouping objects by striking them with the line,
11. Local skew estimation of each text line.

Fig. 1(a) shows the historical document made by typewriter.

From Fig. 1(a), the orientation of text lines is similar. Hence, it represents the global text skew. Fig. 1(b) illustrates a small variation of the orientation between the



Slávny Okresný Úrad !

Podpisaná Miestna skupina cionistického sväzu pre Slovensko v Novom Meste nad Váhom zastúpená Arminom Friederom hlavným rabinom ako predsedom a Ing. Deziderom Kološvárim ako tajomníkom týmto úctive žiadame o povolenie založenia uvedeného spolku, poťažne povolenie účinkovania .

Túto našu žiadosť podporujeme nasledovnými dokladmi :

1./ Dľa pod 1./ pripojených stanov Ústredného sväzu cionistického pre Slovensko, je tento sväz ako taký povolený a stanovy jeho sú schválené.

2./ Prikladáme 3 exempláre zakladajúcou valnou hromadou schválených a prijatých stanov, dľa ktorých je založený spolok členom, poťažne miestnou odbočkou dľa 1./ uvedeného povoleného spolku.

3./ Prikladáme splnomocňujúci dopis dľa pod 1./ uvedeného spolku, dľa ktorého uvedený spolok súhlasiso založením miestnej skupiny .

4./ Prikladáme opis protokolu valného shromaždenia uhodnovernený dvoma uhodnovernitelmi zápisnice a funkcionármi.

5./ Prikladáme prezenčnú listinu valnej hromady konanej dňa 11. januára 1940.

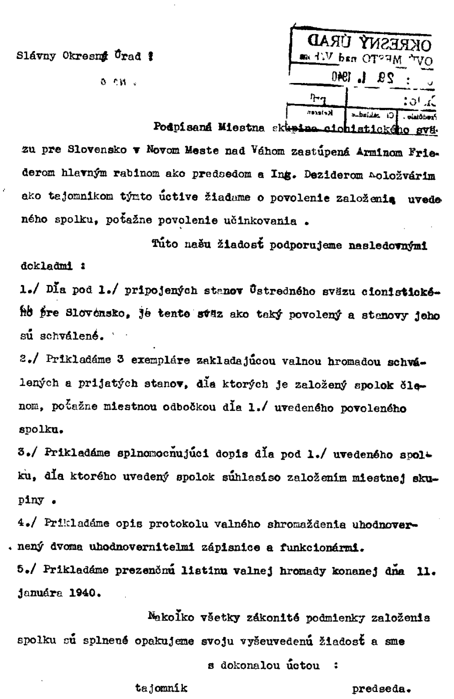
Nakoľko všetky zákonité podmienky založenia spolku sú splnené opakujeme svoju vyšeuvedenú žiadosť a sme

s dokonalou úctou :

tajomník predseda.

(a) (b) (c)

Fig. 1. The historical printed document: (a) Initial document, (b) Initial document with lines that are subsequently drawn showing the local skew variation, (c) Zoomed fragment of document with lines that are subsequently drawn.



Slávny Okresný Úrad !

Podpísaná Miestna skupina cionistického sväzu pre Slovensko v Novom Meste nad Váhom zastúpená Arminom Friederom hlavným rabinom ako predsedom a Ing. Deziderom Aoložvárim ako tajomníkom týmto úctive žiadame o povolenie založenia uvedeného spolku, poťažne povolenie účinkovania .

Túto našu žiadosť podporujeme nasledovnými dokladmi :

1./ Dľa pod 1./ pripojených stanov Ústredného sväzu cionistického pre Slovensko, je tento sväz ako taký povolený a stanovy jeho sú schválené.

2./ Prikladáme 3 exempláre zakladajúcou valnou hromadou schválených a prijatých stanov, dľa ktorých je založený spolok členom, poťažne miestnou odbočkou dľa 1./ uvedeného povoleného spolku.

3./ Prikladáme splnomocňujúci dopis dľa pod 1./ uvedeného spolku, dľa ktorého uvedený spolok súhlasíso založením miestnej skupiny .

4./ Prikladáme opis protokolu valného shromaždenia uhodnoverený dvoma uhodnovernitelmi zápisnice a funkcionármi.

5./ Prikladáme prezenčnú listinu valnej hromady konanej dňa 11. januára 1940.

Nakoľko všetky zákonité podmienky založenia spolku sú splnené opakujeme svoju vyšeuvedenú žiadosť a sme

s dokonalou úctou :

tajomník predseda.

Fig. 2. Document after segmentation with bi-level thresholding.

text line marked as the gray lines. These variations of the orientation are characterized by paper stretching, which establishes a local text skew.

3.1 Bi-level thresholding

A segmentation algorithm based on visual perception (Palmer (1999)) is proposed in (Mello (2010)). The idea is to simulate a person going far from an object. As the distance increases, we lose visual information about details (in documents, details are the text) although the main colors are still perceived (the paper, for documents). If we simulate the perception of a document image at a certain distance, we will not see the letters, but we will see the colors of the background even if it is non-uniform. This simulation is done first through a pre-processing stage (image equalization followed by the application of two morphological closing operations with two different radiuses disks as structuring elements), image downsize (to simulate the distance), image resize (restoring its original size), absolute difference between the preprocessed image after equalization and the resized image, dark pixels (which means zero difference) are converted into white, non-white pixels are complemented (new color = 255 - old color), bright pixels are converted into white and the image is equalized again. This process removes the most part of the background of the image generating a grayscale image which is thresholded by Johannsen and Billie's algorithm (Johannsen and Bille (1982)) producing the final bi-level image $\mathbf{B}$ from initial image $\mathbf{I}$, featuring M rows and N columns. Fig. 2 shows a document sample after algorithm's application.

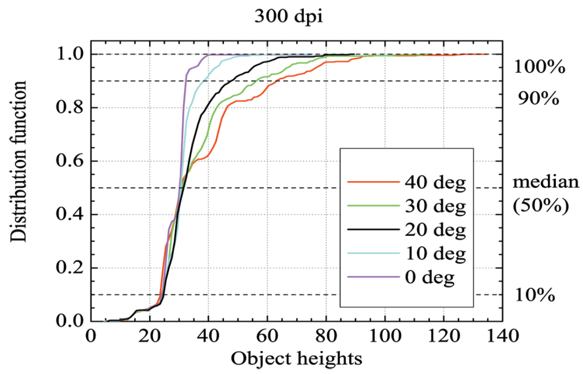


Fig. 3. Distribution function F(x) as a function of object heights for the full range of rotation angles.

3.2 Geometrical preprocessing

The redundant data from document image can be excluded during the preprocessing stage. Hence, the analysis of the object height distributions is performed. We investigate the process of the redundant data exclusion linked to the object heights that do not fall within ranges based on the percentile values of its distribution ($1-99\%$, $5-95\%$, $10-90\%$, $\mu\pm\sigma$, where μ is the mean value and σ is the standard deviation). The percentile of the distribution is the value of a variable below which a certain percentage of observations fall (Wakerly et al. (1996)). Experiment shows that the best results are obtained when $10\%-90\%$ percentile is used (It uses 80% of measured data around the mean value). Fig. 3 shows typical values in the distribution: min, max, 10% percentile, 90% percentile for the document image, which is rotated for the full range of angles, i.e. from $0°$ to $40°$.

The proposed statistical approach, which uses values from the range of $10\%-90\%$ percentile, proved to be successful in the decision-making process of detecting redundant data (Brodic et al. (2013)). Fig. 4 illustrates the exclusion of the redundant data.

3.3 Convex hull extraction

Furthermore, the proposed algorithm exploits the convex hulls around text objects. Convex hull creates a smaller region around the text compared to the bounding box proposed in (Brodic et al. (2012)). Hence, the probability of touching the neighbor text fragments (connected components) has been reduced. After the extraction, they are filled with the white pixels. Fig. 5 illustrates it.

Currently, the text image is given by the matrix $\mathbf{C}$.

3.4 Morphology based object grouping

To estimate the text skew, connected components (CCs) can be enlarged in the horizontal direction. Hence, the morphological erosion is applied to the complementary image $\mathbf{C}$. This way, the adjacent CCs are merged establishing parts of the text line. In this process structuring element $\mathbf{S}_v$ is given as the line. Its thickness is given by 1 pixel, while the length is variable. In order not to touch

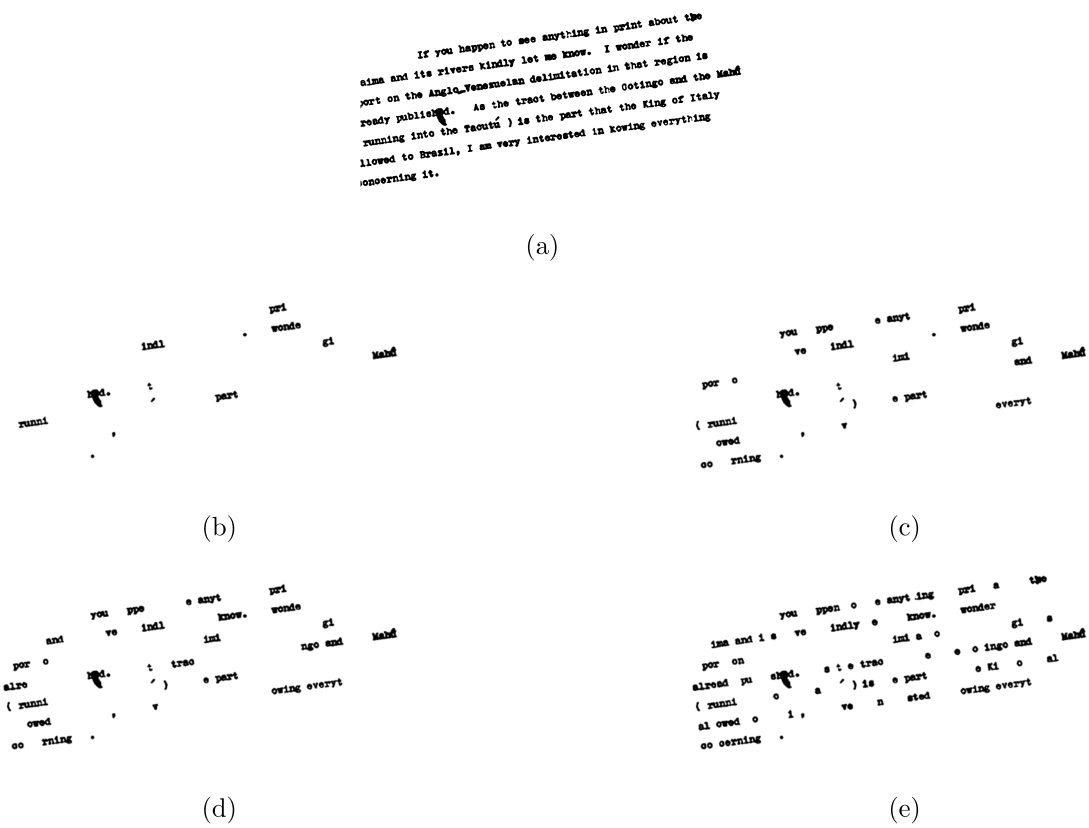


Fig. 4. Difference height values inclusion for different ranges of percentile distribution: (a) Original document, (b) Percentiles 5 – 95%, (c) Percentiles 10 – 90%, (d) Percentiles 15 – 85%, (e) Percentiles 25 – 75%.

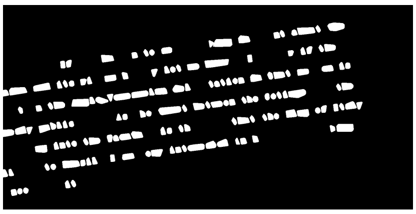


Fig. 5. Text fragment after convex hull extraction.

Fig. 6. The longest CC_{ISR}, which is extracted to detect the initial skew rate.

or join separate neighbor text lines, the length of the line should be chosen carefully. It heavily depends on each CC's height. Empirically, it is given as approximately 30% of the connected component's height (Brodic et al. (2013)). This morphological operation is given as:

$$\mathbf{Y}_v = \mathbf{C} \ominus \mathbf{S}_v. \tag{1}$$

3.5 Initial skew rate detection

Currently, some of the filled convex hulls are joined. They create short connected components (CCs). The longest of them approximates orientation of the text. It is called initial skew rate (ISR). From $\mathbf{C}$, the longest connected components CC_{ISR} is extracted according to the longest common subsequence (LCS) (Brodic et al. (2012)). After that, the initial text skew is calculated by the least square method (See eq. (7)). Fig. 6 shows CC_{ISR}, which is extracted to estimate the initial skew rate.

To better estimate the text skew, the connected components (CCs) can be enlarged in the ISR direction. Again, the structuring element represents the line with the thickness of 1 pixel. Furthermore, the variability of the line width is a function of ISR. Empirically, it is given as approximately 25% of the connected component's height. According to the ISR, the structuring element, $\mathbf{S}_v$ is skewed creating oriented structuring element.

$$\mathbf{Y}_v = \mathbf{C} \ominus \mathbf{S}_v(\angle isr). \tag{2}$$

Fig. 7 shows CCs, which are established with the erosion using structuring element $\mathbf{S}_v$.

3.6 Extraction of the longest object

From $\mathbf{Y}_v$, the longest connected components CC_L is extracted with the longest common subsequence (LCS) method (Brodic et al. (2012)). It is given as $\mathbf{Y}_L$ and shown in Fig. 8.

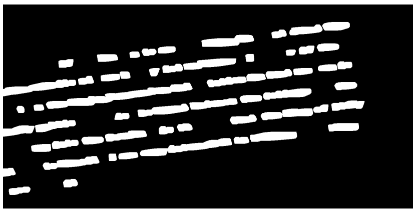


Fig. 7. The enlarged CC objects established by oriented morphological operation in ISR direction.

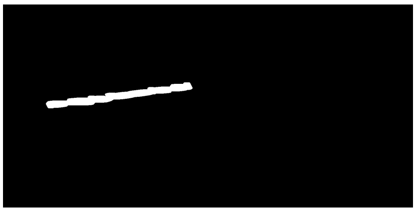
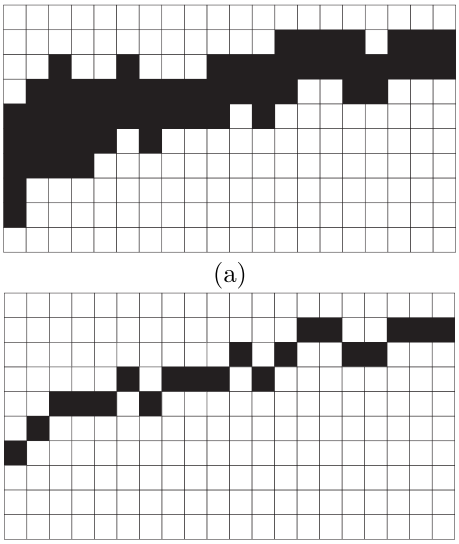


Fig. 8. The extraction of the longest CCs, i.e. CC_L.



(a)

(b)

Fig. 9. The fragment of the longest connected component from complementary image: (a) contour, (b) averaging pixels.

Document text skew can be estimated by identifying the orientation of CC_L (Dhandra et al. (2006)).

3.7 Global skew estimation based on the least square method

The text baseline of the printed text is linear. Hence, the orientation of the connected components CC_L is obtained by linear regression. Prior to that the connected components CC_L should be thinned. This step is illustrated in Fig. 9.

The calculation of the averaging pixels is given as:

$$ap(j) = \frac{\lceil \sum_{i=1}^{N} Y_L(i,j) \rceil}{N}, \tag{3}$$

where $\forall j,\ j = 1, \ldots, M,\ k = 1, \ldots, Q$ and $Q \leq M$. k represents the columns that include the position of the ap only (where exists).

The simplest linear estimator is the least square method. Accordingly, the linear function is given as:

$$y = a_{lsm}x + b_{lsm} \tag{4}$$

where a_{lsm} is the slope and b_{lsm} is y-intercept. The slope a_{lsm} and y-intercept b_{lsm} can be calculated as (Bolstad (2004)):

$$a_{lsm} = \frac{Q \cdot \sum_{k=1}^{Q} x_{ap,k} \cdot y_{ap,k} - c}{Q \cdot \sum_{k=1}^{Q} {x_{ap,k}}^2 - d}, \tag{5}$$

where

$$c = \left(\sum_{k=1}^{Q} x_{ap,k}\right) \cdot \left(\sum_{k=1}^{Q} y_{ap,k}\right),$$

$$d = \left(\sum_{k=1}^{Q} x_{ap,k}\right)^2,$$

$$b_{lsm} = \frac{\sum_{k=1}^{Q} y_{ap,k} - a_{lsm} \cdot \sum_{k=1}^{Q} x_{ap,k}}{Q}. \tag{6}$$

The slope and intercept determine the line that approximates the longest connected component CC_L. Furthermore, the orientation angle β_{lsm} of the line is given as:

$$\beta_{lsm} = \arctan(a_{lsm}) \tag{7}$$

3.8 Global de-skewing of document

The orientation of the longest connected component CC_L estimates the global text skew β_{lsm}. According to the orientation β_{lsm}, the bi-level document image $\mathbf{B}$ featuring M rows and N columns is de-skewed. It is accomplished by direct method (Yu and Jain (1996)). This method proposed that the black pixel p in the input image $\mathbf{B}$ is transformed into p' by multiplying the coordinates of p by a rotational matrix given as:

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos(\beta_{lsm}) & -\sin(\beta_{lsm}) \\ \sin(\beta_{lsm}) & \cos(\beta_{lsm}) \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}, \tag{8}$$

where $(x, y)^t$ are the coordinates of p in the original input image $\mathbf{B}$ and $(x', y')^t$ are the coordinates of p' in the output image $\mathbf{B}'$. Fig. 10 shows a de-skewed document.

3.9 Horizontal projection profiles of de-skewed document

To determine the local text skew, the horizontal projection profile method has been exploited. It extracts features from the projection profiles of text lines, which gives the sum of the black pixels perpendicular to y-axis. It is given by the vector P_v (Zramdini and Ingold (1993)):

$$P_v[i] = \sum_{j=1}^{N} B'(i,j), \tag{9}$$

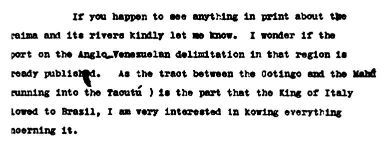

If you happen to see anything in print about the
raima and its rivers kindly let me know. I wonder if the
port on the Anglo_Venezuelan delimitation in that region is
ready published. As the tract between the Cotingo and the Mahú
running into the Tacutú) is the part that the King of Italy
lowed to Brazil, I am very interested in kowing everything
ncerning it.

Fig. 10. Document de-skew according to estimated global text skew.

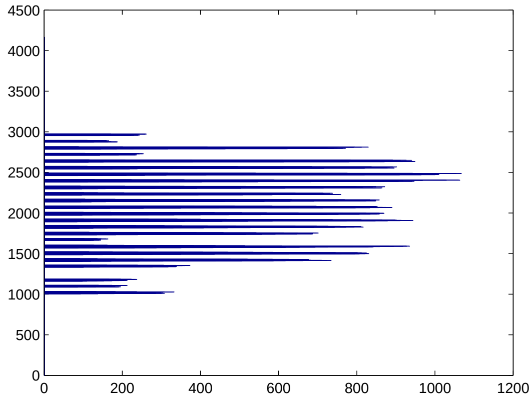


Fig. 11. The horizontal projection profiles of de-skewed document.

where $B'(i, j)$ is the instance of de-skewed binary image. The valleys of the horizontal projection correspond to the background of the image. However, due to the touching of the neighbor text lines, the projection profiles should be analyzed carefully. This way, the histogram is redrawn with the segment between 10% and 100% of the projected profile. The segment below 10% of the projected profile counts the text components that mutually interchange their positions in different neighbor text lines. According to such a histogram, each text line is separated and extracted. Fig. 11 illustrates the horizontal projection profiles of de-skewed document.

3.10 Grouping objects by striking them with the line

The maximum of each peak in projection profiles represents the local maxima. It defines the position, i.e. y-coordinate of each text line. Accordingly, the text objects in each text line are grouped together by striking them through with the line (1-pixel thick). Fig. 12 shows fragments of grouped text lines.

Fig. 13 shows a text line extracted after grouping procedure.

3.11 Local skew estimation of each text line

As a result of grouping procedure, each text line represents a distinct object. The distinct object consists of the letters (grouped by strike through line) that are part of the exact text line. Due to paper stretching, these letters are not quite in line even after the global text skew correction.

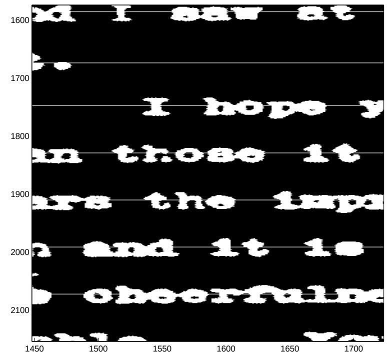


Fig. 12. Fragment of text lines after grouping procedure.

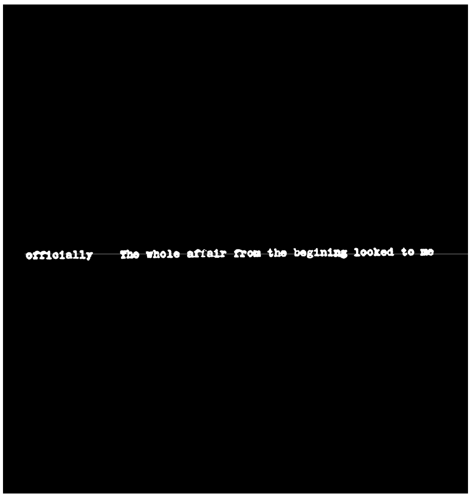


Fig. 13. Fragment of the extended text line after grouping procedure.

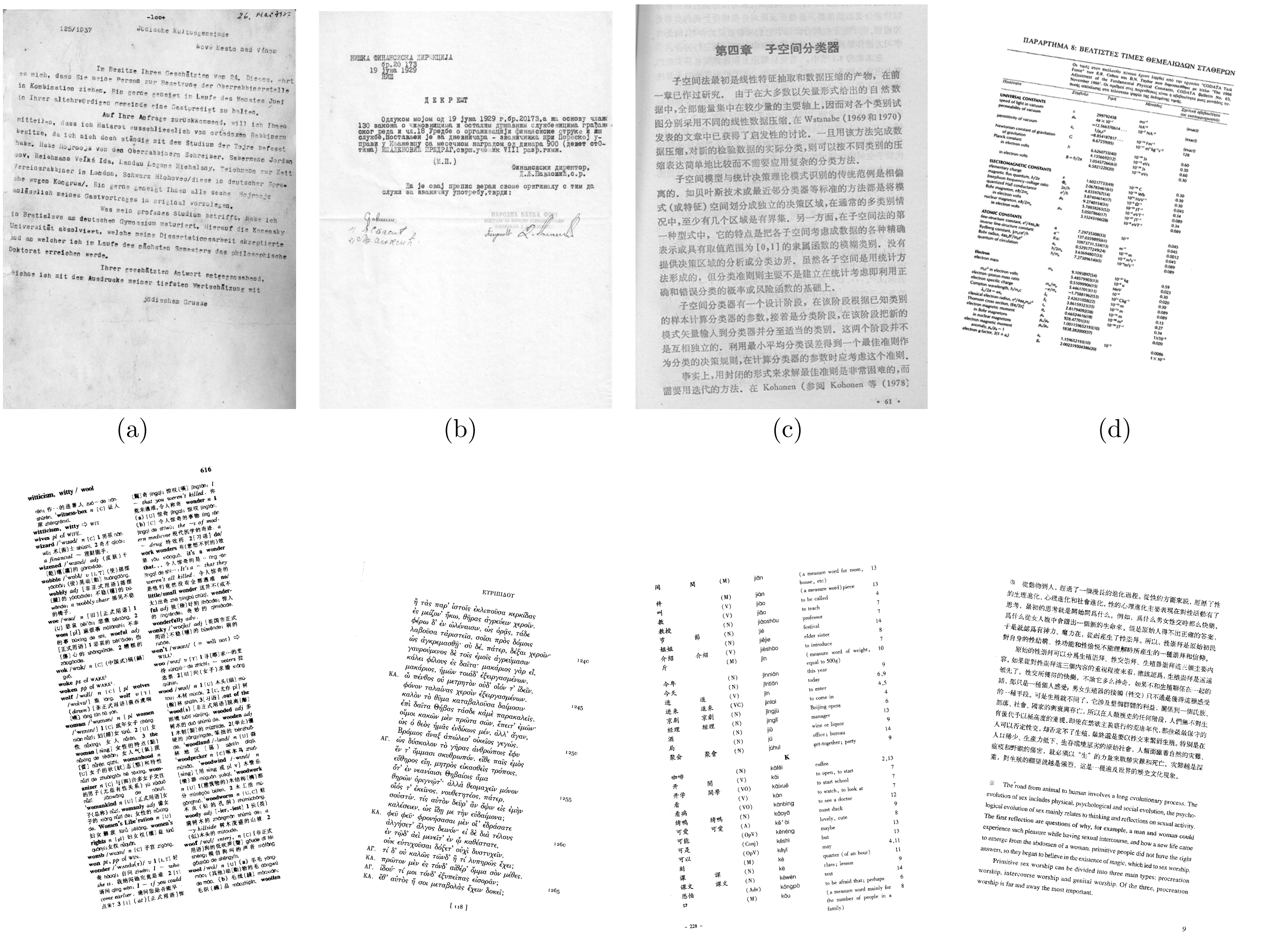
Hence, each text line contains a small inclination. In effect, it represents the local text skew, which can be calculated using eq. (7).

4. EXPERIMENTS

The basic goal of the experiment is to properly estimate the algorithm's ability to correctly determine the text skew. It is performed on the real custom dataset, which consists of eight historical printed documents given in the resolution of 300 dpi. Five of them are excerpts from DISEC'13 database (DISEC13 (2013)). Fig. 14 shows a few document image samples from the dataset.

In the first experiment, ten instances of each document sample from the dataset are randomly rotated in the angle range β from $-15°$ to $+15°$. Second experiment includes the text samples from the dataset that are rotated according to the angle β. It is changed from $-10°$ to $+10°$ by $0.5°$ step and from $\pm10°$ to $\pm45°$ by $1°$ step around x-axis in the positive direction. In total, it represents 410 document instances.

Apart from ref.(Brodic et al. (2013)), the third experiment includes the images from the first experiment which are



(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 14. Document samples from test dataset: (a) Latin, (b) Serbian Cyrillic, (c) Chinese, (d) Greek, (e) English-Chinese combination, (f) Greek, (g)-(h) Chinese-English combination ((d)-(h) excerpt from Document Image Skew Estimation Contest 2013 (DISEC'13)).

exposed to the noise. Accordingly, they are subjected to the salt and pepper noise. This type of noise is generated from an inappropriate thresholding stage, which is a very common in old degraded documents. Hence, the clear images are contaminated with the salt and pepper noise from 0 (without any noise) to 0.05 by 0.01 step (in Matlab). Hence, it includes 400 document instances with artificially added noise. Fig. 15 shows the example of the clear and noise contaminated image.

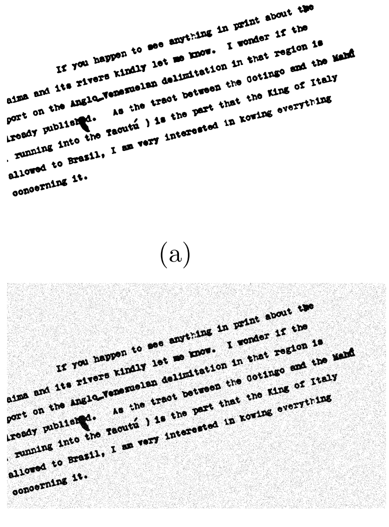
In the last experiment, the global de-skewed document is estimated at the local level. This way, the local skew of each text line in the document sample has been estimated. To compare the obtained results, documents are digitized in 150 and 300 dpi resolutions.

The results are evaluated by the absolute deviation:

$$\Delta\beta_A = |\beta_{REF} - \beta_A|, \tag{10}$$

where β_{REF} is the reference skew of the input text sample and β_A is the text skew obtained by the algorithm.

Furthermore, the average absolute deviation $\Delta\bar{\beta}$ is introduced. It represents the average value of all absolute deviation values, which are obtained from document samples rotated at different angles, i.e.:



(a)

(b)

Fig. 15. The binary image: (a) without noise, (b) with added noise.

Skew angle	Absolute deviation		
Dataset sample	Chou alg.	Mascaro alg.	Propos. alg.
german latin 1	0.1875	0.1625	**0.0677**
serbian cyrillic 1	**0.0812**	0.1000	0.0968
chinese 1	0.0312	**0.0250**	0.0497
greek cyrillic 1	0.2540	0.4260	**0.1193**
english-chinese 1	0.1700	**0.1400**	0.1997
greek cyrillic 2	0.1730	0.1530	**0.1342**
chinese-english 1	**0.1140**	0.1540	0.1945
chinese-english 2	0.1000	**0.0940**	0.1288
Minimum	0.0312	**0.0250**	0.0497
Maximum	0.2540	0.4260	**0.1997**
Max-Min	0.2228	0.4010	**0.1500**
Standard dev.	0.0705	0.1179	**0.0538**
Average	0.1389	0.1563	**0.1238**

Table 1. The absolute deviation of the global text skew obtained for the angle range from 0° to 15° (bold represents the best results).

Skew angle	Absolute deviation		
Dataset sample	Chou alg.	Mascaro alg.	Proposed alg.
german latin 1	0.3750	3.5524	**0.1014**
serbian cyrillic 1	3.1571	3.5048	**0.1073**
chinese 1	3.1190	3.4476	**0.0502**
Minimum	0.3750	3.4476	0.0502
Maximum	3.1571	3.5524	**0.1073**
Max-Min	2.7821	0.1048	**0.0571**
Standard dev.	1.5954	0.0525	**0.0314**
Average	2.2170	3.5016	**0.0863**

Table 2. The absolute deviation of the global text skew obtained for the extended angle range from 0° to 40° (bold represents the best results).

$$\Delta\bar{\beta} = \frac{1}{K}\sum_{l=1}^{K}\Delta\beta_l, \qquad (11)$$

where $\Delta\beta_l$ is the absolute deviation of text skew estimation for each instance of β, $l = 1, ..., K$ is the ordinal of instance and K is the total number of instances.

5. RESULTS AND DISCUSSION

For convention, the best result is presented in bold. The results of the first experiment are shown in Table 1.

The proposed algorithm is compared to a well-known Chou algorithm (Chou et al. (2007)) as well as with its optimized version given by the Mascaro algorithm (Mascaro et al. (2013)). Our average absolute deviation is 0.1238° compared to 0.1389° and 0.1563° obtained by Chou and Mascaro algorithm, respectively. This means that the proposed algorithm shows at least 10% better estimation of the global text skew. Furthermore, the maximum deviation is up to 0.1997° compared to 0.2540° and 0.4260° obtained by Chou and Mascaro algorithm, respectively. As a consequence, the standard deviation of the proposed algorithm is the smallest one, i.e. 0.0538 compared to 0.0705 and 0.1179 given by Chou and Mascaro algorithm, respectively.

The results of the second experiment are shown in Table 2.

Our average absolute deviation is 0.0863° compared to 2.2170° and 3.5016° obtained by Chou and Mascaro algorithm, respectively. It should be pointed out that Chou

Skew angle	Absolute deviation		
Dataset sample	Chou alg.	Mascaro alg.	Propos. alg.
german latin 1	26.3000	5.2833	**0.1631**
serbian cyrillic 1	17.0000	2.2667	**0.0970**
chinese 1	16.0000	1.0167	**0.2281**
greek cyrillic 1	11.9300	0.1033	**0.0695**
english-chinese 1	9.1333	1.0833	**1.0408**
greek cyrillic 2	15.3033	0.4400	**0.2152**
chinese-english 1	6.7633	0.1300	**0.1398**
chinese-english 2	11.5567	0.1800	**0.1727**
Minimum	6.7633	0.1300	**0.0695**
Maximum	26.3000	5.2833	**1.0408**
Max-Min	19.5367	5.1800	**0.9713**
Standard dev.	5.9977	1.7616	**0.3177**
Average	14.2483	1.4997	**0.2658**

Table 3. The absolute deviation of the global text skew obtained from the noise images rotating from 0° to 15° (bold represents the best results).

and Mascaro algorithms are not predetermined for text skew angles larger than 15°. However, our intention was only to show the versatility of the proposed approach. For the extended angle range (up to 40°), the proposed algorithm has the average absolute deviation which is slightly increased, i.e. from 0.0714 to 0.0863 (See the first three rows in Table 1). Hence, it has acceptable values of the absolute deviation in a case of a wide range of angles.

The results of the third experiment are shown in Table 3. Accordingly, the noise images are obtained when the clear images are treated by the salt and pepper noise.

The average absolute deviation of the proposed algorithm is 0.2658° compared to 14.2483° and 1.4997° obtained by Chou and Mascaro algorithm, respectively. Both algorithms that are compared with our approach are not suitable for the estimation of the text skew in noisy images. On the contrary, our algorithm shows a substantial level of robustness to noise present in images.

From all above, our algorithm shows more correct skew estimation results compared to Chou or Mascaro algorithm. Its correctness improvement is around 10-15% in the text skew range from −15° to +15°. Furthermore, the proposed algorithm is suitable for wider skew range as well. Accordingly, the absolute deviation just slight increases in the text skew range from −40° to +40°, i.e. from 0.0714 to 0.0863. Finally, our algorithm is almost insensitive to noise present in images. This contributes to low levels of the absolute deviation obtained from noisy images. The maximum deviation of the proposed algorithm is considerably smaller compared to those obtained by the other two algorithms. This leads to smaller values of the standard deviation.

At the end, we propose an idea how to estimate the small text line variation represented as a local text skew. The results for different image resolutions are given in Table 4. They are similar for different image resolutions. It proves the robustness of the proposed algorithm in the applications where documents are given in different resolutions.

Real (20°)	300 dpi	150 dpi	$\Delta\beta_G(°)$
Global skew	19.8918	19.9147	0.0229
Text line	300 dpi	150 dpi	$\Delta\beta_L(°)$
#1	0.0330	0.0178	0.0152
#2	-0.0408	-0.0322	0.0086
#3	0.0669	0.0134	0.0535
#4	0.0878	0.0393	0.0485
#5	0.3004	0.1582	0.1422
#6	0.2801	0.1457	0.1344
#7	0.2936	0.1642	0.1294
#8	-0.0655	-0.0854	0.0199
#9	0.2232	0.0904	0.1328
#10	0.2692	0.1379	0.1313
#11	0.2548	0.1269	0.1279
#12	0.2647	0.1305	0.1342
#13	0.1700	0.0457	0.1243
#14	0.1419	0.0261	0.1158
#15	0.1156	-0.0132	0.1288
#16	0.0937	-0.0317	0.1254
#17	0.0262	-0.0898	0.1160
#18	0.1191	-0.0032	0.1223
#19	-0.0022	-0.1168	0.1146
#20	-0.1106	-0.2265	0.1159
#21	-0.0197	0.0173	0.0370
#22	-0.2468	-0.3188	0.0720
#23	-0.0406	-0.0196	0.0210
#24	-0.0044	0.0149	0.0193
#25	0.0330	0.0178	0.0152
#26	-0.0408	-0.0322	0.0086

Table 4. The absolute deviation of the local text skew.

6. CONCLUSIONS

This paper proposed the robust method for the estimation of global and local text skew. It is based on the convex hull's extraction over the text in the document image. Prior to that, geometrical filtering in preprocessing stage is performed. Furthermore, convex hulls are extended with oriented binary morphology determined by the initial skew rate. After that, the longest object is extracted. On the basis of its contour, the moment based method estimates its orientation, which determines the global text skew. According to obtained result, the image is globally deskewed. At the end, the method is enhanced in order to estimate small variations in each text line, which represent the local text skew. It is based on horizontal projection profiles. Finding local peaks in the horizontal projection gives the center position of each text line. Grouping them by striking with the line from this center position establishes distinct objects, which represent each text line. Its orientation is obtained by moments.

The proposed method is examined by different experiments applied to the real dataset. It shows good results of global skew estimation from document images given in the resolution of 300 dpi. Furthermore, the algorithm has a low level of sensitivity to noise in the images. Hence, the results are quite promising. Further improvement of the method will be made by including more sophisticated geometrical filtering in preprocessing stage, which will dynamically exclude redundant elements.

Acknowledgement

We thank Dr. Carlos A. B. Mello from Centro de Informtica, Universidade Federal de Pernambuco, Recife, Brazil, and Čedomir A. Maluckov from University of Belgrade, Technical Faculty in Bor, Bor, Serbia for the comments that greatly improved the manuscript.

REFERENCES

A. Rehman, T. Saba. Document Skew Estimation and Correction: Analysis of Techniques, Common Problems and Possible Solutions. *Applied Artificial Intelligence*, 25:(9), 769–787, 2011.

A. Amin, S. Wu. Robust Skew Detection in Mixed Text/Graphics Documents. In *Proceedings of 8^{th} International Conference on Document Analysis and Recognition*, 247–251, 2005.

W. Postl. Detection of Linear Oblique Structures and Skew Scan in Digitized Documents. In *Proceedings of the 8th International Conference on Pattern Recognition*, 687–689, 1986.

L. O'Gorman. The Document Spectrum for Page Layout Analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence*, 15:(11), 1162–1173, 1993.

S. N. Srihari and V. Govindaraju. Analysis of textual images using the Hough transform. *Machine Vision and Application*, 2:(3), 141–153, 1989.

A. Amin and S. Fischer. A Document Skew Detection Method using the Hough Transform. *Pattern Analysis and Application*, 3:(3), 243–253, 2000.

C. Singh, N. Bhatia, A. Kaur. Hough Transform based Fast Skew Detection and Accurate Skew Correction Methods. *Pattern Analysis*, 41:(12), 3528–3546, 2008.

S. Lowther, V. Chandran, S. Shidharan. An Accurate Method for Skew Determination in Document Images. In *Proceedings of Digital Image Computing Techniques and Applications (DICTA)*, 1–5, 2002.

H. Yan. Skew Correction of Document Images Using Interline Cross-Correlation. *CVGIP: Graphical Models and Image Processing*, 55:(6), 538–543, 1993.

D. Brodic, Z. N. Milivojevic. Log-polar Transformation as a Tool for Text Skew Estimation. *Elektronika Ir Elektrotechnika*, 19:(2), 61–64, 2013.

A. K. Das, Z. N. Milivojevic. A Fast Algorithm for Skew Detection of Document Images. *International Journal on Document Analysis and Recognition*, 2001:(4), 109–114, 2001.

B. V. Dhandra, V. S. Malemath, H. Mallikarjun and R. Hegadi. Skew Detection in Binary Image Documents based on Image Dilation and Region Labeling Approach. In *Proceedings of $18^{textrmth}$ International Conference on Pattern Recognition*, 954–957, 2006.

P. Shivakumara, G. H. Kumar, D. S. Guru and P. Nagabhushan. A Novel Technique for Estimation of Skew in Binary Text Document Images based on Linear Regression Analysis. *Sadhana*, 30:(1), 69–84, 2005.

J. Flusser, B. Zitova and T. Suk. *Moments and Moment Invariants in Pattern Recognition.* John Wiley & Sons, New York, 2009.

D. Brodic and Z. N. Milivojevic. Estimation of the Handwritten Text Skew Based on Binary Moments. *Radioengineering*, 21:(1), 162–169, 2012.

C. H. Chou, S. Y. Chu and F. Chang. Estimation of Skew Angles for Scanned Documents based on Piecewise Covering by Parallelograms. *Pattern Recognition*, 40:(2), 443–445, 2007.

M. Makridis, N. Nikolau and N. Papamarkos. An Adaptive Technique for Global and Local Skew Correction in Color Documents. *Expert Systems with Applications*, 37:(10), 6832–6843, 2010.

D. Brodic, C. A. Maluckov, L. Peng. Estimation of the Text Skew in the Old Printed Documents. *International Journal of Computers Communications & Control*, 8:(5), 673–680, 2013.

S. E. Palmer. *Vision Science Photons to Phenomenology.* MIT Press, Cambridge, 1999.

C. A. B. Mello. Segmentation of Images of Stained Papers Based on Distance Perception. In *IEEE International Conference on Systems, Man and Cybernetics*, 1636–1642, 2010.

G. Johannsen, J. Bille. A Threshold Selection Method using Information Measures. In 6^{th} *International Conference on Pattern Recognition*, 140–143, 1982.

D. D. Wakerly, W. III Mendenhall and R. L. Scheaffer. *Mathematical Statistics with Applications.* Duxbury Press, Belmont, 1996.

D. Brodic, D. R. Milivojevic. An algorithm for the Estimation of the Initial Text Skew. *Information Technology and Control*, 41:(3), 211–219, 2012.

W. M. Bolstad. *Introduction to Bayesian Statistics.* John Wiley & Sons, New Jersey, 2004.

B. Yu, A. K. Jain. A Robust and Fast skew Detection Algorithm for Generic Documents. *Pattern Recognition*, 29:(10), 1599–1629, 1996.

A. Zramdini, R. Ingold. Optical Font Recognition from Projection Profiles. *Electronic Publishing*, 6:(3), 249–260, 1993.

DISEC'13. http://users.iit.demokritos.gr/ alexpap/DISEC13/

A. A. Mascaro, G. D. C. Calvacanti, C. A. B. Mello. Fast and Robust Skew Estimation of Scanned Documents Through Background Area Information. *Pattern Recognition Letters*, 31:(11), 1403–1411, 1993.