Estimation of the Global and Local Text Skew in the Old Printed Documents

Darko Brodić[1], Ivo Draganov[2], Dragan R. Milivojević[3], Viša Tasić[3]

Abstract – **The paper proposed a robust skew estimation method for the old printed document that estimates global and local skew angles. The global skew angle characterizes the main skew angle of the document image. In order to identify it, the longest object was extracted. Then, its skew was estimated with the moments. Accordingly, document image was globally de-skewed. The local skew angle represents a fine variation of each text line after de-skewing document image. Creating vertical projection profiles of de-skewed document identified different text lines. Line through text objects joined them in each text line. At the end, text skew of such objects was estimated by the moments identifying local text skews.**

Keywords – **Document image analysis, Moment methods, Optical character recognition, Projection algorithms, Skew Adjustment.**

I. Introduction

Old historical documents represent the part of great cultural and scientific importance. Due to its age, it is quite common for such documents to suffer from degradation. Examples of degradations include shadows and variable background intensity, smudges, ink seeping, smear and strains. These degradations make image preprocessing particularly difficult and produce recognition errors. The printed text is mostly uniform [1]. That considers pretty the same skew, which represents the global text skew. However, old historical documents incorporate paper stretching. It contributes to non-uniformity of text orientation. Hence, each text line has a slightly different skew, which represents local text skew.

In document automatic recognition systems, the quality of the input image is crucial to the final performance. There are a variety of interfering effects such as noise and skewing that appear during the scanning process. These components disturb the proceeding and decrease the performance of the recognizer. Skew correction plays an important role in the image preprocessing. A small inclination in the document image can interfere in the layout analysis and consequently, in the rest of the process. That is why the identification of the object skew in the image is one of the most important tasks in digital image processing and document image analysis. It is due to optical character recognition (OCR) system sensitivity to any skew appearance in the text.

Because of the scanning process, the text skew occurrence is unavoidable. It is an implication of the digitization process. To solve a problem, a large amount of techniques has been developed. They can be classified as [1]: projection profiles method, k-nearest neighbour clustering methods, Hough transforms methods, Fourier transformation methods, cross-correlation methods, and other methods.

Many of these methods have strong points as well as weaknesses. Projection profile method is a straightforward method, which is suitable for text with uniform skew only [2]. K-nearest neighbor clustering method cannot handle incorporation of noisy subparts in text, which leads to reduced accuracy. The Hough transforms method needs preprocessing stage, which defines candidate mapping points [4]. The method is complex and computer time intensive. The Fourier transforms method is even more complex [5]. The cross-correlation method is limited only to small skew angles up to 10° [6]. Interesting extension of this group of methods is given in [7].

The techniques classified as other methods are based mostly on combination techniques. They have been reputed as the most efficient ones. However, they are multistage and computer time intensive. In [8]-[9], preprocessing of document image is made by complex decision making. It is performed with complex geometrical filtering. Global text skew is identified with the cross-correlation method applied to remain connected components. At the end, local text skew is calculated with the least square method. This technique performs local skew estimation and reliable text localization without restriction of the skew angle value.

The main contribution of this paper is the algorithm suitable for the recognition of the text skew in the old printed documents characterized with dominant skew and small variation of the local skew in each text line.

Organization of this paper is as follows. Section 2 presents previous works. Section 3 describes proposed algorithm. Section 4 defines text experiments. Section 5 shows and discusses test result. Section 6 makes conclusions.

II. Previous Works

The algorithm identifies global and local text skew of the printed documents. The global text skew represents the dominant text skew of the whole document, while the local text skew shows a small variation of the skew in each text line. It consists of the global and local skew estimation stage.

First, the stage that estimates the global text skew consists of the steps that follow:

- Uneven illumination reduction with binarization.
- Convex hulls extraction.
- Joining text objects with binary morphology.
- Extraction of the longest object.
- Skew estimation of the longest object by the moments.

[1] Darko Brodić is with the University of Belgrade, Technical Faculty in Bor, Vojske Jugoslavije 12, 19210 Bor, Serbia.

[2] Ivo Draganov is with the Technical University of Sofia, Faculty of Telecommunications, Bul. Kl. Ohridsky 8, Sofia 1797, Bulgaria.

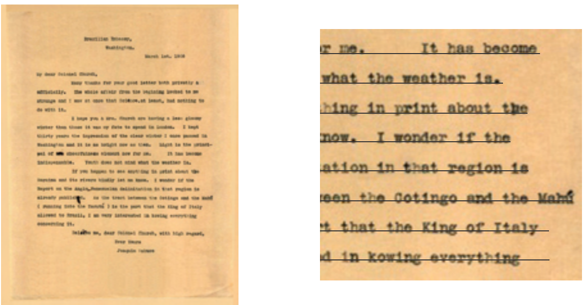
[3] Dragan R. Milivojević and Viša Tasić are with the Mining and Metallurgy Institute Bor, Zeleni Bulevar 35, 19210 Bor, Serbia.

- Global de-skewing of the original document.

The stage that estimates the local text skew includes following:

- Vertical projection profiles of de-skewed document.
- Joining objects in each text line striking them with line.
- Skew estimation of each text line.

Fig. 1.(a) shows the old historical printed document.



(a) (b)

Fig. 1. The old historical printed document: (a) Whole document, and (b) Fragment (lines are subsequently drawn).

It is obvious that the orientation of text lines is quite similar, which represents dominant i.e. global text skew. Fig. 1.(b) shows the fragment from the historical document. To illustrate a small difference between text lines orientation, the lines are subsequently drawn in text. These variation of the orientation is characterized by paper stretching, which forms a local text skew.

Global Text Skew Estimation

Uneven illumination reduction with binarization

The uneven illumination distribution in the background of historical documents is very frequent. It degrades the visual quality of the text image and makes it difficult to recognize the content. In order to efficiently solve this problem, ref. [10] proposed an efficient edge-based light balancing scheme (ELBS) for text images. After application of this algorithm, the document text image is a grayscale without uneven illumination. It is given by matrix **D**, which consists of M rows, N columns, and L intensity levels of gray. L is the integer from {0, …, 255}. $D(i, j) \in \{0, …, 255\}$, where $i = 1, …, M$ and $j = 1, …, N$.

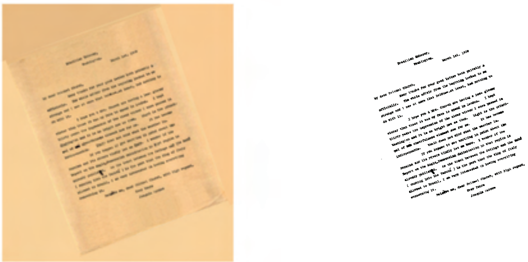
The binarization transforms grayscale into a binary image $B(i, j)$. Hence, if $D(i, j) \geq D_{\text{th}}$ then $B(i, j) = 1$, else if $D(i, j) < D_{\text{th}}$ then $B(i, j) = 0$, where D_{th} represents the global threshold sensitivity value [11]. Document text image is a binary image represented by matrix **B** featuring M rows and N columns, and two intensity levels. Fig. 2 shows skewed document: (a) original, and (b) binarized.

Convex hull extraction

Instead of using bounding boxes [12], the convex hulls over text objects have been exploited. Convex hull creates a smaller region around the text compared to the bounding box. Hence, the probability for touching the neighbor text fragments has been reduced. Upon the extraction of the convex hulls, they are filled with white pixels. Such a text image is given with matrix **C**.

Joining text objects with binary morphology

Each convex hull is growing in all directions creating connected components (CC). It is achieved with erosion applied to **C**. This way, the adjacent CC is merged establishing the text line.



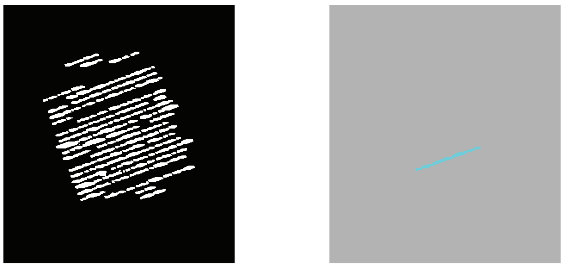
(a) (b)

Fig. 2. Skewed document: (a) Original, and (b) Binarized.

The structuring element **S** represents a line with variable width. It is given as:

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

In order not to touch or join separate neighbour text lines, the width of the line should be chosen carefully. It heavily depends on each CC's height. Empirically, it is calculated as approx. 30% of the CC's height. Fig. 3.(a) shows connected components established by erosion.



(a) (b)

Fig. 3. (a) Image after morphological operation, (b) Extraction of CC_L.

Extraction of the longest object

From **Y**, the longest connected components CC_L is extracted with the longest common subsequence (LCS) method [12]:

$$CC_L = \underbrace{\max}_{i,j}\left(\bigcap_{n=1}^{K} CC_{ncc}\right), \tag{2}$$

where n_{cc} is a number of CC. The longest CC, i.e. CC_L is shown in Fig 3.(b). Document text skew can be estimated by identifying the orientation of CC_L [12].

Skew estimation of the longest object by the moments

In order to estimate the skew orientation of CC_L, the moment based technique is used. Moment defines the measure of the pixel distribution in the image. It identifies global image information that depends on its contour. Moments of the binary image **B** are given as [13]:

$$m_{pq} = \sum_{i=1}^{N}\sum_{j=1}^{M} i^p j^q , \tag{3}$$

where p and q = 0, 1, 2, 3, ..., n, and n represent the order of the moment. The central moment μ_{pq} for binary image **B** is calculated as:

$$\mu_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} (i - \bar{x})^p (j - \bar{y})^q \ . \qquad (4)$$

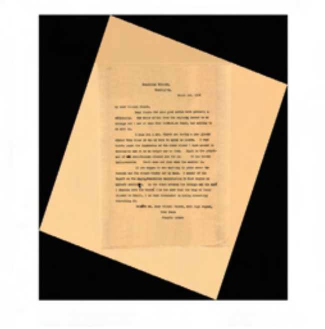
The image feature representing the object orientation θ is obtained from the moments. It illustrates the angle between the object and the horizontal axis. It is obtained as [13]:

$$\theta = 0.5 \cdot \arctan(2\mu_{11} / (\mu_{20} - \mu_{02})) \qquad (5)$$

Skew of the longest object represents the global text skew.

Global de-skewing of the original document

According to the global text skew, the binary document is de-skewed. It is shown in Fig. 4.(a) and 4.(b).



(a) (b)

Fig. 4. (a) Original document, and (b) De-skewed document.

After de-skewing, the local text skew should be estimated.

Local Text Skew Estimation

Vertical projection profiles of de-skewed document

First, vertical projection profile method is exploited. It extracts features from the projection profiles of text lines, which gives the sum of black pixels perpendicular to the *y*-axis. It is represented with the vector P_v defined as [14]:

$$P_v[i] = \sum_{j=1}^{N} B_d(i, j) \qquad (6)$$

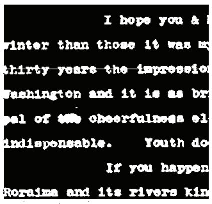
where B_d (i, j) is the instance of de-skewed binary image. The valleys of the vertical projection correspond to the background of the image. Finding local peaks in the vertical projection gives the center position of each text line [14].

Joining text objects in each text line

The maximum of each peak represents the coordinate of the line, which is drawn. This way, text objects in each text line are joined together. Fig. 5 shows joined text lines.

Skew estimation of each text line

Local text skew is calculated using (6) for each text line. That's way, for the example of text document skewed by 20°, the global text skew is 19.7918°, while the local text skews are (26 lines) between -0.2468° and 0.2692° (300 dpi).

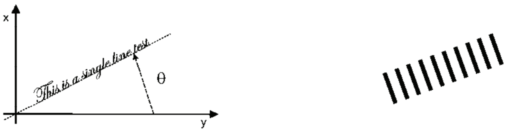


I hope you &
winter than those it was my
thirty years the impressio
Washington and it is as br
pal of [illegible] cheerfulness el
indispensable. Youth do
If you happen
Roraima and its rivers kin

Fig. 5. Zoom of the text lines after joining procedure.

III. Experiments

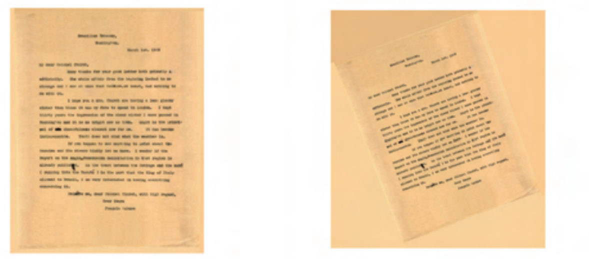
The task of the experiments is to evaluate algorithm's ability to estimate the text skew. The experiments were performed on real and synthetic datasets. Synthetic dataset represent the single-line samples of the printed text, which are rotated. The angle θ is rotated from 0° to 45° by the 5° steps around *x*-axis in the positive direction. It is illustrated in Fig. 6.



(a) (b)

Fig. 6. Synthetic dataset: (a) Definition, (b) Example.

Fig. 7 shows the original document which is rotated creating dataset. The angle θ is rotated from 0° to 45° by the 5° steps around *x*-axis in the positive direction.



(a) (b)

Fig. 7. Real dataset: (a) Original document (b) Skewed document.

All text samples are given in the resolution of 300, 150, 100, 75 and 50 dpi. The results are evaluated by the absolute deviation. It is given as:

$$\Delta\theta_A = \left|\theta_{REF} - \theta_A\right| \ , \qquad (7)$$

where θ_{REF} is the referent skew of the input text sample and θ_A is the text skew obtained by the algorithm. The relative error (*RE*) is given as:

$$RE(\theta) = \frac{\Delta\theta_A}{\theta_{REF}} . \qquad (8)$$

IV. Results and Discussion

Tables I-II show absolute deviation of the global text skew for synthetic and real dataset, respectively.

TABLE I. ABSOLUTE DEVIATION FOR SYNTHETIC DATASET.

Synthetic	**300 dpi**	**150 dpi**	**100 dpi**	**75 dpi**	**50 dpi**
θ_{REF} (°)	Δθ(°)	Δθ(°)	Δθ(°)	Δθ(°)	Δθ(°)
0	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.0372	0.0517	0.0544	0.1157	0.0017
2	0.2582	0.0925	0.1351	0.0102	0.1088
3	0.3355	0.1292	0.0147	0.0134	0.2550
4	0.0576	0.0613	0.0718	0.1798	0.1198
5	0.0839	0.1000	0.0323	0.1010	0.0921
6	0.0614	0.0134	0.0148	0.2135	0.8336
7	0.0814	0.0492	0.1348	0.3068	0.4113
8	0.1055	0.1307	0.0278	0.0056	0.5080
9	0.0448	0.1036	0.2335	0.1013	0.3859
10	0.1340	0.0893	0.2268	0.2138	0.6438
15	0.2786	0.3033	0.3809	0.3628	1.0774
20	0.4534	0.5437	0.5745	0.9677	0.8270
25	0.6338	0.7805	0.9036	1.3814	2.2697
30	0.9245	1.1040	1.0183	1.3528	2.0654
35	1.1288	1.2527	1.4378	1.7182	2.3277
40	1.4909	1.4214	1.9928	2.3231	2.5742
45	1.5888	1.8475	1.9872	1.7570	2.2730

TABLE II. ABSOLUTE DEVIATION FOR REAL DATASET.

Real	**300 dpi**	**150 dpi**	**100 dpi**	**75 dpi**	**50 dpi**
θ_{REF} (°)	Δθ(°)	Δθ(°)	Δθ(°)	Δθ(°)	Δθ(°)
0	0.4764	0.4644	0.4167	0.4903	0.5773
1	0.4438	0.4449	0.4280	0.4059	0.5378
2	0.4385	0.4109	0.4062	0.4384	0.4943
3	0.4599	0.4046	0.4764	0.4224	0.4725
4	0.4030	0.4436	0.4472	0.5166	0.5029
5	0.4218	0.3968	0.4536	0.4674	0.5577
6	0.4148	0.4460	0.4597	0.4575	0.4262
7	0.4295	0.4075	0.4498	0.4628	0.6191
8	0.4770	0.4500	0.4529	0.4250	0.5965
9	0.4634	0.4366	0.4073	0.4218	0.6414
10	0.4500	0.4461	0.4934	0.4927	0.7326
15	0.4616	0.6302	0.6061	0.9292	1.0112
20	0.4710	0.6444	0.6452	1.1104	1.1068
25	0.4823	0.6910	0.6710	1.3061	1.5084
30	0.4827	0.7016	0.9542	1.4755	1.5822
35	0.7296	0.7738	1.3487	1.6119	1.4585
40	0.7583	0.8054	1.5281	1.6468	1.8245
45	0.7686	0.7925	1.4999	1.8097	1.9028

Global skew results are as follows: up to 10°, maximum absolute deviation is from 0.13° to 0.21°, except for images in 50 dpi, where it is 0.64°; up to 30°, maximum absolute deviation is from 0.92° to 1.35°, except for images in 50 dpi, where it is 2.06°; up to 45°, maximum absolute deviation is from 1.58° to 2.32°, except for images in 50 dpi, where it is 2.57° (for synthetic dataset), and up to 10°, maximum absolute deviation is from 0.45° to 0.49°, except for images in 50 dpi, where it is 0.73°; up to 30°, maximum absolute deviation is from 0.48° to 1.47°, except for images in 50 dpi, where it is 1.58°; up to 45°, maximum absolute deviation is from 0.76° to 1.80°, except for images in 50 dpi, where it is 1.90° (for real dataset). Maximum absolute deviation of the local skew are between -0.3718° and 0.2647°, except for images in 50 dpi, where it is between -1.0853° and 1.2994°. These results are quite acceptable, because geometrical filtering was excluded in preprocessing stage.

V. Conclusion

The paper proposed robust method for the estimation of global and local text skew. Method shows good results of global skew estimation for different resolution of test images from 300 down to 50 dpi. It is a merit of the moment based method exploration. Method shows an acceptable result for the local skew estimation. Accordingly, a fine-tuning of each text line orientation is feasible. Further research should be directed toward incorporation of preprocessing geometrical filtering, which will exclude elements like smudges, ink seeping, smear and strains.

Acknowledgement

This paper is supported by the Grant of the Ministry of Science of Republic of Serbia, as a part of the project TR33037 within the framework of Technological development program.

References

[1] A. Amin, S. Wu, "Robust Skew Detection in Mixed Text/Graphics Documents," in Proc. of 8th ICDAR, Seoul, Korea, 2005, pp. 247–251.

[2] R. Manmatha, N. Srimal, "Scale Space Technique for Word Segmentation in Handwritten Manuscripts", in Proc. of 2nd ICSSTCV, LNCS 1682, London, Great Britain, 1999, pp. 22–33.

[3] L. O'Gorman, "The Document Spectrum for Page Layout Analysis", IEEE Trans. Pattern Anal. Mach. Intell., 1993, Vol.15, No.11, pp. 1162–1173.

[4] G. Louloudis, B. Gatos, I. Pratikakis, C. Halatsis, "Text Line Detection in Handwritten Documents", Pattern Recognition, 2008, Vol.41, No.12, pp. 3758–3772.

[5] W. Postl, "Detection of Linear Oblique Structures and Skew Scan in Digitized Documents," Proc. of 8th ICPR, Paris, France, 1986, pp. 687–689.

[6] H. Yan, "Skew Correction of Document Images Using Interline Cross-Correlation", CVGIP: Graphical Models and Image Processing, 1993, Vol.55, No.6, pp. 538–543.

[7] D. Brodić, Z. N. Milivojević, "Log-polar Transformation as a Tool for Text Skew Estimation," Electronics and Electrical Engineering, 2013, Vol.19, No.2, pp. 61–64.

[8] P. Saragiotis, N. Papamarkos, "Local Skew Correction in Documents," Int. J. Pattern Recognit. Artif. Intell., 2008, Vol.22, No.4, pp. 691–710.

[9] M. Makridis, N. Nikolau, N. Papamarkos, "An Adaptive Technique for Global and Local Skew Correction in Color Documents," Expert Syst. Appl., 2010, Vol.37, No.10, pp. 6832–6843.

[10] Kuo-Nan Chen, Chin-Hao Chen, Chin-Chen Chang, "Efficient Illumination Compensation Techniques for Text Images", Digit. Signal Proc., 2012, Vol.22, No.5, pp. 726–733.

[11] N. Otsu, "A Threshold Selection Method from Gray-level Histograms," IEEE Trans. Sys., Man., Cyber., 1979, Vol.9, No.1, pp. 62–66.

[12] D. Brodić, D. R. Milivojević, "An Algorithm for the Estimation of the Initial Text Skew," Inf. Technol. Control, 2012, Vol.41, No.3, pp. 211–219.

[13] G. Kapogiannopoulos, N. Kalouptsidis, "A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments," Proc. of SPPRA, Crete, Greece, 2002, pp. 275–279.

[14] A. Zramdini, R. Ingold, "Optical Font Recognition from Projection Profiles", Electronic Publishing, 1993, Vol.6, No.3 pp 249–260.