INT J COMPUT COMMUN, ISSN 1841-9836
8(5):673-680, October, 2013.

Estimation of the Text Skew in the Old Printed Documents

D. Brodić, Č.A. Maluckov, L. Peng

Darko Brodić, Čedomir A. Maluckov
Technical Faculty in Bor,
University of Belgrade
V. J. 12, 19210 Bor, Serbia
dbrodic@tf.bor.ac.rs, cmaluckov@tf.bor.ac.rs

Liangrui Peng
Department of Electronic Engineering,
Tsinghua University
Beijing 100084, P.R. China
penglr@tsinghua.edu.cn

Abstract: Old printed documents represent the significant part of our heritage. In order to preserve them, the digitalization is indispensable. The paper proposed a robust skew estimation method for old printed document. It is based on the connected components made by filled convex hulls around text element. The connected components are enlarged by oriented morphological operation. Then, the longest connected component is extracted. The global orientation of the document is detected by its orientation. Accordingly, document image was globally de-skewed. The algorithm is tested on synthetic and real datasets. Obtained results proved the algorithms correctness.
Keywords: document image analysis, moment methods, optical character recognition, skew adjustment.

1 Introduction

Old documents represent the part of great cultural and scientific importance. Due to 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. In document automatic recognition systems, the quality of the input image is crucial to 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 document image can interfere in the layout analysis and consequently in the rest of the process. That's 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 so due to optical character recognition (OCR) system sensitivity to any skew appearance in the text.

In this paper, we deal with old printed documents like letters, technical notes, etc. They are characterized with the shape regularity as any other printed text, [1] which contain letters with similar sizes. The distance between text lines is adequate, which facilitates separation of text lines. The orientation of the text lines is similar. That considers pretty the same skew, which represents the global text skew.

A large amount of techniques has been developed in order to identify text skew. They are classified as: [1] projection profiles methods, k-nearest neighbor 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. [3] 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 those methods represent the incorporation of log-polar transformation. [7] However, sometimes it is unstable in application. 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. Such methods are proposed in [8]- [9]. Preprocessing of document image is made by complex decision making. It is performed with complex geometrical filtering. The 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.

Organization of the paper is as follows. Section 2 describes the algorithm. Section 3 defines text experiments. Section 4 gives the results and discusses them. Section 5 makes conclusions.

2 Proposed Algorithm

The proposed algorithm identifies the skew, which represents the dominant skew of the whole printed document. It consists of the steps that follows: 1) Uneven illumination reduction with binarization, 2) Convex hulls extraction, 3) Joining text objects with oriented binary morphology, 4) Extraction of the longest object, 5) Skew estimation of the longest object by the moments, and 6) Global de-skewing of the original document.

2.1 Uneven illumination reduction with binarization

The binarization method adopts both global and local adaptive thresholds. [10]- [11] First, multiple candidate thresholds are computed via histogram of the original gray image. Those pixels which are definitely background and foreground pixels are recorded, and the remained gray pixels will be binarized by the adaptive thresholding method. Second, we get the binarized results of the remained gray pixels by multiple candidate thresholds respectively, the statistical parameters such as run-length are calculated for each binarized images. Finally, the optimal threshold is selected when the statistical parameters are stable. If the statistical parameter analysis fails, a global threshold will be calculated by histogram analysis. After the binarization, document image is transformed into a binary matrix $\mathbf{B}$ featuring M rows and N columns. It has two intensity levels, i.e. $B(i,j) = \{0, 1\}$. Figure 1 shows the document image before and after binarization process.

2.2 Convex hull extraction

Instead of using bounding boxes, [12] the proposed algorithm exploits the convex hulls over text. Convex hull creates a smaller region around the text compared to bounding box. Hence, the probability of touching the neighbor text fragments has been reduced because of smaller contour. Upon the extraction of convex hulls, they are filled with white pixels due to complementary image. Such a text image is given with matrix $\mathbf{C}$. Figure 2(a) shows convex hulls extraction.

(a) (b)

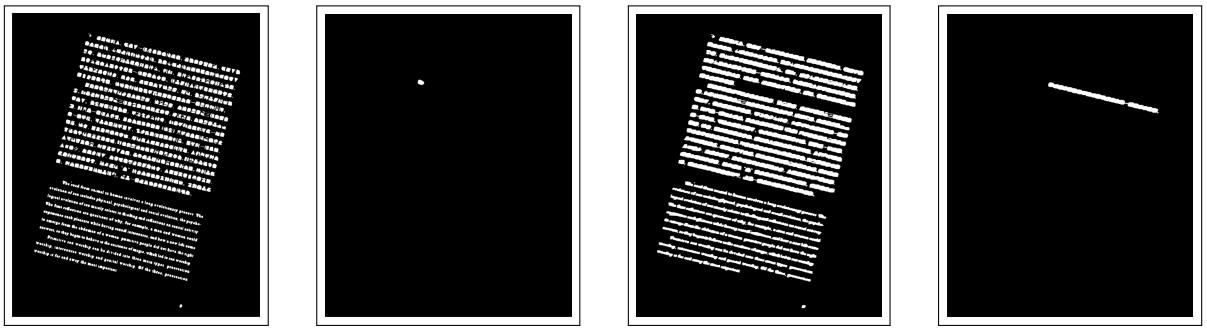
Figure 1: Document text image: (a) Before binarization, (b) After binarization

2.3 Joining text objects with oriented binary morphology

Currently, some of filled convex hulls are joined. They create short connected components (CC). The longest of them CC_{ISR} accompanies the attribute of the orientation intention. It is called initial skew rate (ISR). CC_{ISR} is extracted by the application of the longest common subsequence (LCS). [13]

$$\mathrm{CC}_{ISR} = \underbrace{\max}_{i,j} \left(\bigcap_{m=1}^{K} \mathrm{CC}_m \right), \tag{1}$$

where $K = 1$ is the total number of CC. CC_{ISR} is shown in Figure 2(b). Its orientation is calculated by the moments (See eq. (7) for reference).



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

Figure 2: Document text image: (a) CC extraction (b) CC_{ISR} extraction, (c) Extended CC extraction (b) The longest CC extraction

In order to correctly estimate the text skew, connected components should be extended. Hence, morphological erosion is applied to **C**. This way, the adjacent CC's are merged establishing parts of the text line. Structuring element **S** representing a variable width line is used. In order not to touch or join separate neighbor text lines, the width of the line should be chosen with caution. It heavily depends on each CC's height. Empirically, it is used as 30% of the connected component's height, which means that width of structuring element **S** applied to each CC's is different. Furthermore, its variability is a function of ISR, because structuring element **S** is skewed according to ISR orientation. Morphological operation is given as:

$$\mathbf{Y} = \mathbf{C} \oplus \mathbf{S}(\angle ISR). \tag{2}$$

Figure 2(c) shows extended CC made by oriented morphology.

2.4 Extraction of the longest object

Currently, extended CCs are created. They represent partial text lines. It is clear that the longest of them CC_{LNG} incorporates the orientation which is similar to text skew. Hence, it is mandatory to extract CC_{LNG} from $\mathbf{Y}$. Again, it is performed with LCS method: [13]

$$\mathrm{CC}_{LNG} = \underbrace{\max}_{i,j}\left(\bigcap_{n=1}^{L}\mathrm{CC}_n\right), \tag{3}$$

where L is the total number of extended CC. CC_{LNG} is shown in Figure 2(d). Document text skew can be estimated by identifying the orientation of CC_{LNG}.

2.5 Skew estimation of the longest object by the moments

In order to estimate the skew orientation of CC_{LNG}, 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 $\mathbf{Y}$ featuring M rows and N columns are: [14]

$$m_{pq} = \sum_{i=1}^{M}\sum_{j=1}^{N} i^p j^q, \tag{4}$$

where p and q = 0, 1, 2, 3, ... , r, and r represents the order of the moment. The central moment's μ_{pq} of the binary image $\mathbf{Y}$ can be calculated as:

$$\mu_{pq} = \sum_{i=1}^{M}\sum_{j=1}^{N}(i-\overline{x})^p(j-\overline{y})^q. \tag{5}$$

The image feature which represents the object orientation θ is obtained from the moments. It illustrates the angle between the object and the horizontal axis. It is given as: [14]

$$\theta = \frac{1}{2}\arctan\left(\frac{2\mu_{11}}{\mu_{20}-\mu_{02}}\right), \tag{6}$$

Hence, the orientation θ of the longest object CC_{LNG} estimates the global text skew.

2.6 Global de-skewing of the original document

According to the orientation of the longest object θ, the initial document image is de-skewed. Figure 3 shows the document image before and after de-skewing process.

3 Experiments

Main goal of the experiment is the evaluation of the algorithm's ability to estimate text skew. It is performed on real and synthetic datasets. In this case, synthetic dataset consist of the samples that include single-line printed text. The samples are given in the resolution of 300 dpi. They are rotated for the angle θ, from 0° to 10° by 1° and from 10° to 40° by 5° steps around x-axis in the positive direction. It is shown in Figure 4(a).

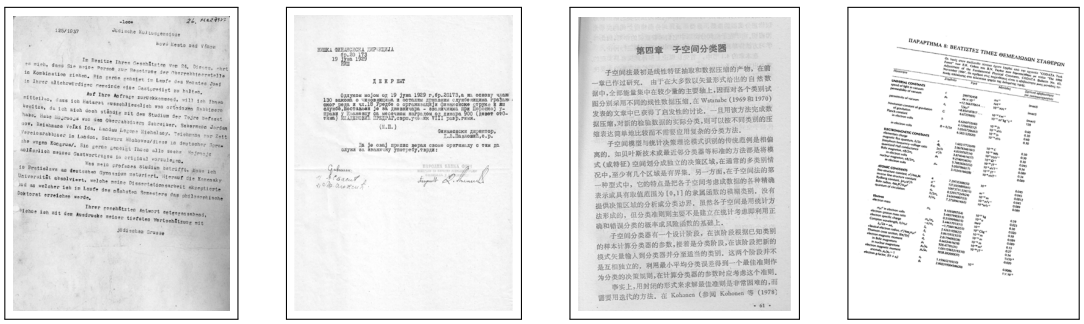
(a) (b)

Figure 3: Document text image: (a) With skew (b) After de-skew

(a) (b)

Figure 4: Dataset rotation: (a) Synthetic dataset, (b) Real dataset

Real dataset consists of document image samples given in the resolution of 150 and 300 dpi. They are rotated for the angle θ, from 0° to 10° by 1° and from 10° to 40° by 5° steps around x-axis. Figure 5 shows document samples of the real dataset in Latin, Serbian Cyrillic, Chinese and Greek Cyrillic (excerpt from ICDAR 2013 text skew test).



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

Figure 5: Samples from real dataset: (a) Latin document, (b) Serbian Cyrillic document, (c) Chinese document, (d) Greek Cyrillic document

After the algorithm's application to dataset obtained result represents the estimated text skew. This result is compared to the reference text skew of the document samples from dataset. The evaluation of the algorithm's result is made by the absolute deviation:

$$\Delta\theta_A = |\theta_{REF} - \theta_A|, \tag{7}$$

where θ_{REF} is the reference skew of the input text sample and θ_A is text skew estimated by the algorithm.

4 Results and Discussion

Table 1 shows the absolute deviation of global text skew for synthetic and real dataset. These result are given for the full range of rotation angles ($0° - 40°$).

Table 1: Absolute deviation for synthetic and real dataset

Resolution	300 dpi			150 dpi		
Dataset	Synthetic	Real		Synthetic	Real	
θ_{REF}	$\Delta\theta$	isr	$\Delta\theta$	$\Delta\theta$	isr	$\Delta\theta$
0	0.0734	1.0362	0.0376	0.0676	-0.6384	0.2490
1	0.0408	2.2297	0.0147	0.0481	0.8453	0.0329
2	0.0355	1.6194	0.1067	0.0141	1.0806	0.0005
3	0.0569	-0.0450	0.0020	0.0078	2.4496	0.2352
4	0.0001	0.7877	0.0209	0.0468	3.3863	0.0086
5	0.0188	2.2632	0.2458	0.0001	1.7279	0.2227
6	0.0118	2.9291	0.0221	0.0492	5.6370	0.1314
7	0.0265	3.9265	0.2889	0.0107	6.0385	0.0679
8	0.0740	4.9008	0.0504	0.0532	6.9414	0.0769
9	0.0604	7.0503	0.0987	0.0398	8.5425	0.0045
10	0.0470	5.5578	0.2775	0.0493	8.4272	0.9760
15	0.0586	11.9536	0.3697	0.2334	15.0123	0.0648
20	0.0680	15.9705	0.3497	0.2476	19.5475	0.1556
25	0.0793	21.9282	0.2223	0.2942	26.5648	0.1537
30	0.0797	27.0247	0.2245	0.3048	28.7789	0.2365
35	0.3266	31.9050	0.1856	0.3770	34.5256	0.2569
40	0.3553	36.8890	0.2028	0.4086	38.9246	4.6753
Average	0.0831	-	0.1600	0.1325	-	0.4440

From Table 1 results are as follows:

- for synthetic dataset given in the resolution of 300 dpi the absolute deviation is below $0.08°$ for the angles up to $30°$ and below $0.35°$ for the angles between $30°$ and $40°$ with the average value of $0.08°$,

- for real dataset given in the resolution of 300 dpi the absolute deviation is below $0.37°$ for the angles up to $40°$ with the average value of $0.16°$,

- for synthetic dataset given in the resolution of 150 dpi the absolute deviation is below $0.05°$ for the angles up to $10°$ and up to $0.415°$ for the angles between $10°$ and $40°$ with the average value of $0.13°$,

- for real dataset given in the resolution of 150 dpi the absolute deviation is below $0.97°$ for the angles up to $35°$ with the average value of $0.44°$.

The proposed text skew algorithm has the average absolute deviation of $0.16°$ for the text skew angle θ up to $40°$. Compared obtained result with the result of the algorithm without using oriented morphology, [16] the average value of absolute deviation is lower approx. $0.2°$. It has quite acceptable values of the absolute deviation in the wide range of angles. Furthermore, the algorithm has been applied to different types of documents (including few examples from Document Image Skew Estimation Contest - ICDAR 2013) and different types of letters. It can be

used for documents like letters, technical articles, journals, dictionary, etc. Furthermore, above results are quite acceptable because geometrical filtering in preprocessing stage was excluded. However, proposed algorithm doesn't have the versatility of the multi-stage method proposed in [8]- [9]. This complex methods include complicated steps of geometrical filtering in preprocessing stages in order to exclude some redundant elements. However, such methods are much more computer time intensive. In further development, proposed method should be expanded with the inclusion of some additional geometrical filtering steps. This step will contribute to lower dispersion of absolute error value (up to 0.1°).

5 Conclusions

The paper proposed robust method for the estimation of global text skew. The method shows good results of global skew estimation for different resolution of test images. It is a merit of the moments exploration. Furthermore, the algorithm is suitable for text skew identification of document types like letters, technical articles, journals, dictionary, etc. Due to the exclusion of the preprocessing elements, some of redundant data were included in the process of text skew identification. Hence, further development of the algorithm should include geometrical filtering, which will lead to lower dispersion of estimated skew value.

Acknowledgment

This work was partially supported by the Grant of the Ministry of Science from Republic of Serbia, as a part of the project TR33037 and III43011 within the framework of Technological development program and the National Natural Science Foundation of China under Grant No. 61261130590.

Bibliography

[1] Amin, A.; Wu, S. (2005); Robust Skew Detection in Mixed Text/Graphics Documents, *Proc.of 8th ICDAR*, Seoul, Korea, 247-251.

[2] Manmatha, R.; Srimal, N. (1999); Scale Space Technique for Word Segmentation in Handwritten Manuscripts, *Proc. of 2nd ICSSTCV*, LNCS 1682, London, Great Britain, 22-33.

[3] O'Gorman, L. (1993); The Document Spectrum for Page Layout Analysis, *IEEE Trans Pattern Anal Mach Intell*, ISSN 0162-8828, 15(11): 1162-1173.

[4] Louloudis, G.; Gatos, B.; Pratikakis, I.; Halatsis, C. (2008); Text Line Detection in Handwritten Documents, *Pattern Recognition*, ISSN 0031-3203, 41(12): 3758-3772.

[5] Postl,W. (1986); Detection of Linear Oblique Structures and Skew Scan in Digitized Documents, *Proc. of 8th ICPR*, Paris, France, 687-689.

[6] Yan, H. (1993); Skew Correction of Document Images Using Interline Cross-Correlation, *CVGIP: Graphical Models and Image Processing*, ISSN 1049-9652, 55(6): 538-543.

[7] Brodić, D.; Milivojević, Z.N. (2013); Log-polar Transformation as a Tool for Text Skew Estimation, *Elektronika Ir Elektrotechnika* ISSN 1392-1215, 19(2): 61-64.

[8] Saragiotis, P.; Papamarkos, N. (2008); Local Skew Correction in Documents, *Int J Pattern Recognit Artif Intell* ISSN 0218-0014, 22(4): 691-710.

[9] Makridis, M.; Nikolau, N.; Papamarkos, N. (2010); An Adaptive Technique for Global and Local Skew Correction in Color Documents, *Expert Syst Appl*, ISSN 0957-4174, 37(10): 6832-6843.

[10] Otsu, N. (1979); A Threshold Selection Method from Gray-level Histograms, *IEEE Trans Sys, Man, Cyber,* ISSN 0018-9472, 9(1): 62-66.

[11] Chen, Kuo-Nan; Chen, Chin-Hao; Chang, Chin-Chen (2012); Efficient Illumination Compensation Techniques for Text Images, *Digit Signal Prog* ISSN 0165-1684, 22(5): 726-733.

[12] Brodić, D.; Milivojević, D.R. (2012); An Algorithm for the Estimation of the Initial Text Skew, *Inf Technol Control*, ISSN 1392-124X, 41(3): 211-219.

[13] Brodić, D. (2011); The Evaluation of the Initial Skew Rate for Printed Text, *J Electr Eng*, ISSN 1335-3632, 62(3): 142-148.

[14] Kapogiannopoulos, G.; Kalouptsidis, N. (2002); A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments, *Proc. of SPPRA*, Crete, Greece, 275-279.

[15] Zramdini, A.; Ingold, R. (1993); Optical Font Recognition from Projection Profiles, *Electronic Publishing*, ISSN 0194-4851, 6(3): 249-260.

[16] Brodić, D.; Milivojević, D.; Tasić, V.; Milivojević, Z. (2013); Identification of the Global Text Skew Based on the Convex Hulls, *Proc. of MIPRO*, Opatija, Croatia, 1282-1286.