Journal of Universal Computer Science, vol. 20, no. 4 (2014), 488-506
submitted: 27/5/13, accepted: 27/2/14, appeared: 1/4/14 © J.UCS

An Approach to Skew Detection of Printed Documents

Darko Brodić
(Technical Faculty in Bor, University of Belgrade, Bor, Serbia
dbrodic@tf.bor.ac.rs)

Carlos A. B. Mello
(Centro de Informática, Universidade Federal de Pernambuco, Recife, Brazil
cabm@cin.ufpe.br)

Čedomir A. Maluckov
(Technical Faculty in Bor, University of Belgrade, Bor, Serbia
cmaluckov@tf.bor.ac.rs)

Zoran N. Milivojević
(College of Applied Technical Sciences Niš, Niš, Serbia
zoran.milivojevic@vtsnis.edu.rs)

Abstract: In this paper, we propose an approach to estimate the text skew for printed documents. This is an important step to prevent errors in further stages of an automatic document processing system (as text segmentation). Our approach is based on the statistical analysis of the height of the connected components. In a nutshell, our algorithm is comprised of four steps: (i) removal of redundant data; (ii) establishment of the connected components, which represent filled convex hulls around each text element; (iii) enlargement of these components using morphological erosion; (iv) removal of the largest connected component to identify the first estimation of text skew. According to it, the connected components are enlarged by oriented morphological erosion and the longest of them is extracted. Statistical moments are applied to this longest component to evaluate its orientation and the global text skew of the document is identified. At the end of this process, the original document is rotated back based on the calculated angle. The performance of the proposed algorithm is examined by testing on a custom dataset. The results support the robustness of our approach.

Keywords: Document image analysis, Connected component analysis, Statistical analysis, Moment based method, Skew estimation
Categories: I.4, I.4.5, I.4.7, I.4.10, I.5, I.5.3, I.7, I.7.2

1 Introduction

In spite of the increasing use of digital documents, humanity still has to deal with a large amount of paper documents (old and new). Among these, arguably the most common are printed documents, especially in the last 10 to 20 years, when printers were more widely used. Before that, typing machines were also common. Part of our archive of typed documents is composed by historical documents which represent an important part of our technical and cultural heritage.

In document recognition systems, the quality of the input image is essential to the output performance. During the scanning process, adverse effects, such as noise or skew, are inevitable, sometimes due to the large amount of digitization work which has to be manually carried out. This is especially noticeable when dealing with old documents, as the use of automatic feeding of the scanner is not allowed because it can damage the documents. As a result, the skew angle can be up to 15 degrees in any direction, although such large angles are fortunately not common. All these issues can decrease the performance of an automatic document recognition system. For example, a small inclination in a document image can interfere with layout analysis [Amin, 05], and optical character recognition tools (OCR) show a significant sensitivity to any skew, which generally leads the system to fail. Hence, skew detection is a key element in document image analysis [Amin, 05], [Manmatha, 99], [Rehman, 11].

Skew estimation and correction clearly play an important role in automatic image processing and they make the whole process very complex. Therefore, appropriate methods should be developed to remove redundant elements from printed documents during processing for further recognition. In this paper, we call these elements 'redundant' as they are not important for skew estimation. Some of redundant data correspond to the connection of different text lines that have close ascender or descender characters (as letters 'h' and 'g'). Some degrees of skew in the document image can also merge different text lines through the connection of these two types of characters. The other possible reason for this merging of different text lines is the small distance and positional perturbations of nearest-neighbour pairs from different text lines [Lu, 03]. The process of segmenting redundant data from a document image is mainly based on geometrical filters. In order to discard components which correspond to redundant elements, some text attributes are specified, and then classified [Saragiotis, 08], [Zagoris, 11]. The selection of geometrical filters is based on the analysis of some internal features, which are inherent to the connected components and their neighbourhood. Such internal features should recognize the connected components that have low probability of being text objects. Specifically, these features can be the area, density and width-to-height ratio [Saragiotis, 08], or some other geometrical attributes [Zagoris, 11], [Makridis, 10].

In this paper, we propose a statistical approach to estimate skew estimation in document images with the exclusion of redundant data. This approach is based on a size restriction of the connected components. According to this, all connected components that do not fall within pre-specified ranges, which are based on the percentile values of the distribution of the heights of the connected components, are excluded. Such an approach is suitable for text skew detection algorithms based on the principle of nearest neighbour clustering, as well as morphological transform. Both methodologies start with connected components which are enlarged by a boundary growing algorithm in order to distinguish skewed elements. The proposal deals with printed documents characterized by some shape regularity. In other words, the letters have similar sizes, the distance between text lines is adequate and the orientation of text lines is similar. All of the above constitutes dominant, i.e., global text skew. Hence, we propose an approach that groups two skew detection algorithms together to estimate the global text skew of printed documents. The document images are already binarized, so we do not deal with thresholding or noise removal.

In Section 2, we review some previous remarkable works on algorithms for text skew estimation. Section 3 describes the full procedure of our proposed algorithm.

Section 4 describes the experiment along with the proposed dataset. Section 5 evaluates the results of the experiment, while Section 6 presents the conclusions.

2 Related works

During the scanning process, text skew occurrence is unavoidable. Most document analysis systems require a prior skew detection before the image processing stages. Hence, an accurate method for determining document image skew is an essential need [Lu, 03]. To solve this problem, a large number of methods has been proposed, and they may be separated into the following classes [Amin, 05], [Rehman, 11]:

1. Projection profile methods,
2. K-nearest neighbour clustering methods,
3. Hough transforms methods,
4. Cross-correlation methods, and
5. Other methods.

Many of the above methods have some strengths as well as weaknesses. Some brief description and characteristics are given next.

Projection profile methods can be classified as a computationally unclaimed technique. They use a binary image as a basis, running through each of the image's lines, counting the total number of black pixels. This creates a black pixel density histogram. To detect the text skew, the highest peak of this histogram is determined [Baird, 87]. The strength of this method is its insensitivity to noise spikes intrinsic to text. However, it is suitable for uniform text skew only, which renders it unable to be used for text skew estimation in multi-column documents. As an extension of this method, an algorithm based on the Fourier transform was proposed [Postl, 86]. The basic idea is similar to projection profile methods; however, it is applied in the frequency domain. As such, the position where the density of Fourier space has the highest peak corresponds to the text skew angle. Unfortunately, this algorithm is complex and computationally intensive [Postl, 86].

K-nearest neighbour clustering methods are based on page layout analysis [Hashizume, 86]. In order to create connected components, the nearest neighbour cluster is used. Each connected component is created by connecting the nearest cluster, which is typically made of characters. Furthermore, the histogram is calculated, and it represents the direction vectors for all pairs of nearest neighbours. The peak of the histogram is then determined and used to estimate the text skew angle. The primary weakness of this method is its sensitivity to noise, as it cannot handle noisy subparts which usually are present in text [O'Gorman, 93].

Hough transform-based methods use a well-known technique for detecting lines in images [Ballard, 81]. These methods need a pre-processing stage, which defines candidate mapping points. An example is presented in [Le, 93]. These methods map the image points from Cartesian- into parametric Hough- space, where a majority voting process is carried out. The object candidates are acquired in accumulator space, where local maxima are identified. However, the way of finding an accumulator peak is very time consuming [Pal, 96].

Cross-correlation methods are based on the correlation between text- and reference-lines, comparing the deviation of interline cross-correlation. To determine text skew, a cross-correlation function is created in which peak value designates the text skew orientation. The method is limited to small skew angles, up to 10 degrees [Yan, 93]. A particular extension to this method is given in [Brodic, 13]. It presents an idea of cross-correlation, which is applied in log-polar domain, reducing some disadvantages of the original cross-correlation method.

Other methods are based mostly on the combination of different techniques. The algorithm proposed in [Najman, 04] incorporates morphological and geometrical transformations. Here morphological transformations include dilation and erosion, which regularly utilizes a structuring element with different length. In addition, geometrical moments can also be used to measure the pixel distribution in the image [Flusser, 09]. As a consequence, some of these are sensitive to rotation, which makes them suitable text skew estimation [Kapogiannopoulos, 02]. Another interesting method was proposed in [Chou, 07]. It detects a dominant skew in document images using parallelograms to cover the document in a piecewise fashion. The method proved to be very accurate (absolute error is up to 0.1 degree), but the skew range is limited to angles lower than 15 degrees. This method was further improved in [Mascaro, 10]. However, it is not robust to noise or to larger skew angles either. A much more complex method was proposed in [Makridis, 10]. It includes a complex pre-processing stage composed of multi-stage decision making and geometrical filtering. Global text skew is identified with the cross-correlation methods, which are applied to the remaining connected components. At the end, the local text skew is calculated using least squares. This technique performs local skew estimation, with reliable text localization without restriction on the skew angle value. However, the method is computationally intensive.

3 New Skew Detection Algorithm

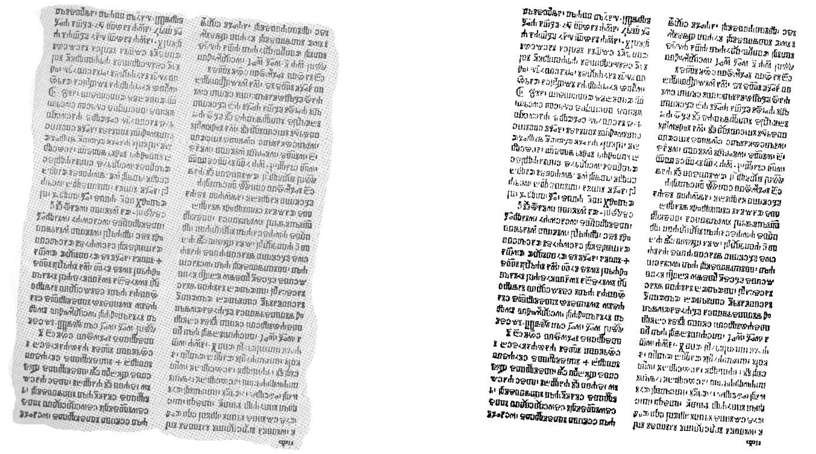
The skew detection algorithm proposed herein identifies global text skew of printed documents, which represents the dominant text skew of the whole document. It consists of the following major steps:

1. Binarization,
2. Statistical pre-processing based on connected component heights,
3. Extraction of the convex hulls,
4. Extraction of the connected components linked with the initial skew rate,
5. Application of an oriented binary morphology,
6. Extraction of the longest connected component,
7. Skew estimation of the longest connected component based on moments.

Figure 1(a) shows a sample printed document which was skewed during the scanning process.

3.1 Binarization

In order to binarize the document image, it was used a segmentation algorithm based on visual perception [Palmer, 99] as proposed in [Mello, 14]. The idea is to simulate a person going far from an object. As the distance increases our visual system loses visual information about details (in documents, details are the text) although the main colours are still perceived (the paper, for document images). Thus, if the perception of a document image at certain distance is simulated, the letters will not be perceived anymore but the colours of the background will still be visible 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 pre-processed image after equalization and the resized image, dark pixels (which means zero difference) are converted into white, non-white pixels are complemented (new colour = 255 - previous colour, for greyscale document images), bright pixels are converted into white and the image is equalized again.



(a) (b)

Figure 1: Printed document image in Glagolitic script (Missal dated from 1483): (a) initial state, (b) after uneven illumination reduction and binarization.

This process removes most part of the background of the image generating a greyscale image which is thresholded by a combination of Otsu thresholding method and k-means producing the final bi-level image. Figure 1(b) shows the document image of Figure 1(a) after the application of this binarization algorithm.

3.2 Statistical Pre-processing Based on Connected Component Heights

Text elements in a bi-level image represent distinct objects. These objects are referred as connected components (CC). They usually correspond to separated characters, fragment of characters or even some noise elements. These elements have different distributions of their sizes. The main goal of the preprocessing stage is to disregard redundant data according to their heights. Elements with small height typically represent noise, while elements with great height correspond to capital letters, merged neighbor elements from different text lines, etc. These elements contribute to possible errors in text skew estimation. Hence, they are considered as redundant data. The segmentation of these redundant data from document image will make easier and more accurate the skew estimation. Accordingly, the analysis of the objects height distribution in the image is performed. The process of data rejection is investigated jointly with the object heights that do not fall within ranges based on the percentile values of its distribution. The percentile of the distribution is the value of a variable below which a certain percentage of observations falls [Wackerly, 96].

Figure 2 shows the distribution function *F*(*x*) of the connected component heights from the image given in Figure 1(b).

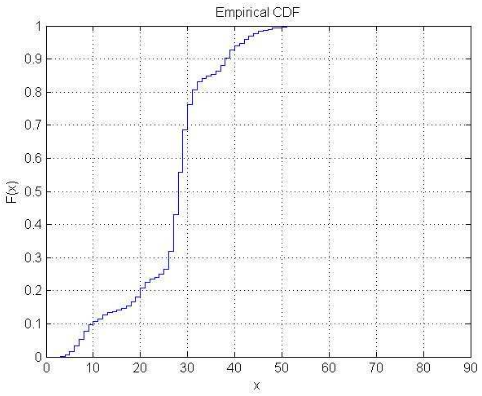


Figure 2: Distribution function F(x) of the connected component heights for the printed document image.

In order to optimize the segmentation of redundant data, the percentile distribution ranges from 0-100% to 25-75% is investigated [Brodic, 14]. To find the optimum percentile distribution range (for data reduction) the minimum absolute angle deviation is analyzed. The difference between maximal and minimal object heights is assumed as an input, while the difference between measured and reference text skew, i.e., the absolute angle deviation, represents output. The objective is to find the percentile distribution range which produces the minimum absolute angle deviation. The experiments are performed for three skew ranges: [0°-10°], [0°-15°], and [0°-40°]. Figure 3 shows the dependence of the absolute angle deviation in accordance to the heights percentile distribution.

Experiments show that the best results are obtained for 10-90% percentile, which uses 80% of measured data around the mean value. The proposed statistical approach proved to be successful in the process of detecting redundant data [Brodic, 14]. Figure 4 illustrates excluded data which consists of joined connected component inherent to different text lines.

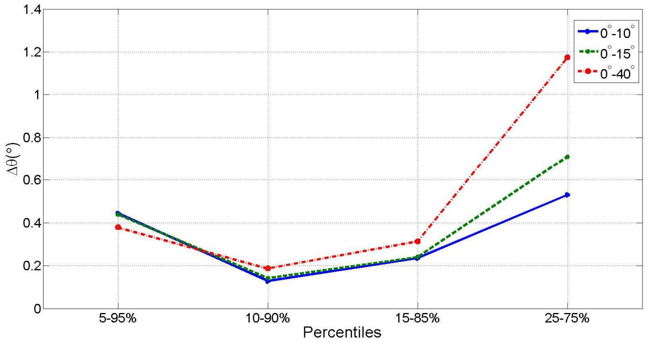


Figure 3: The absolute angle deviation (Δθ) vs. the heights percentile distribution.

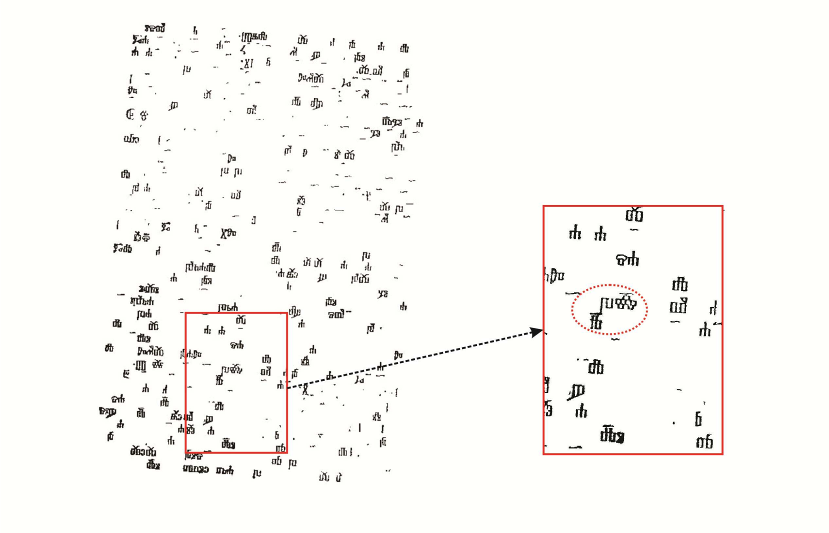
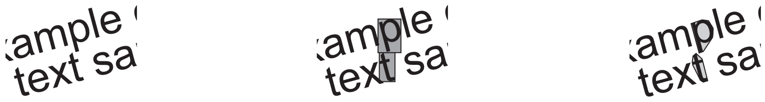


Figure 4: Examples of redundant data excluded with proposed statistical approach (connected component fragments that include joined text objects from different text lines).

3.3 Extraction of the Convex Hulls

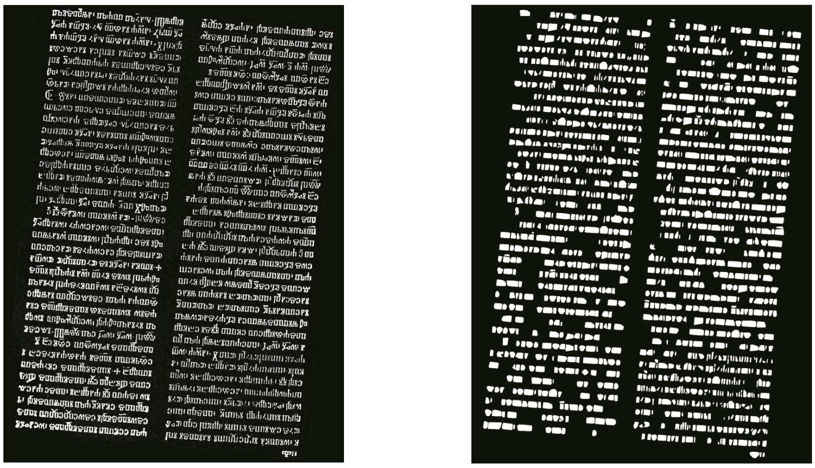
Many algorithms use bounding boxes to extract connected components needed for text skew estimation [Shivakumara, 05], [Brodić, 12]. However, convex hulls create smaller regions around the text if compared to bounding boxes. Figure 5 illustrates this comparison.



(a) (b) (c)

Figure 5: Comparison of the use of bounding boxes against convex hull: (a) original text, (b) sample bounding boxes of two letters (the boxes are connected although the letters are not), and (c)the letters are segmented with convex hulls that do not touch each other.

Thus, the fragments of touching neighbour (elements from different text lines) are reduced by using convex hulls. Due to the advantage of convex hulls utilization, this approach is employed in the proposed algorithm. First, the original bi-level image is complemented. Accordingly, the convex hulls over text elements are extracted in that image. Then, they are filled with white pixels (as it is applied to the complementary image). This creates the matrix **C** (see Figure 6(b) for reference).



(a) (b)

Figure 6: Convex hulls extraction: (a) Full complementary document image, (b) Convex hulls extraction after redundant data exclusion

3.4 Extraction of the Connected Components Linked with the Initial Skew Rate

At this point, some of the filled convex hulls can be grouped. Consequently, they create short connected components. The longest one incorporates the attribute of approximate text skew called Initial Skew Rate (ISR). From **C**, the longest connected components, CC_{ISR}, is extracted according to the longest common subsequence (LCS)

[Brodić, 11]. After that, initial text skew is calculated based on statistical moments.Figure 7 shows CC_{ISR} extracted from Figure 6(b) in order to estimate the initial skew rate.

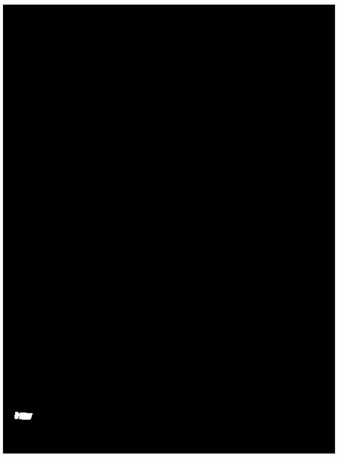
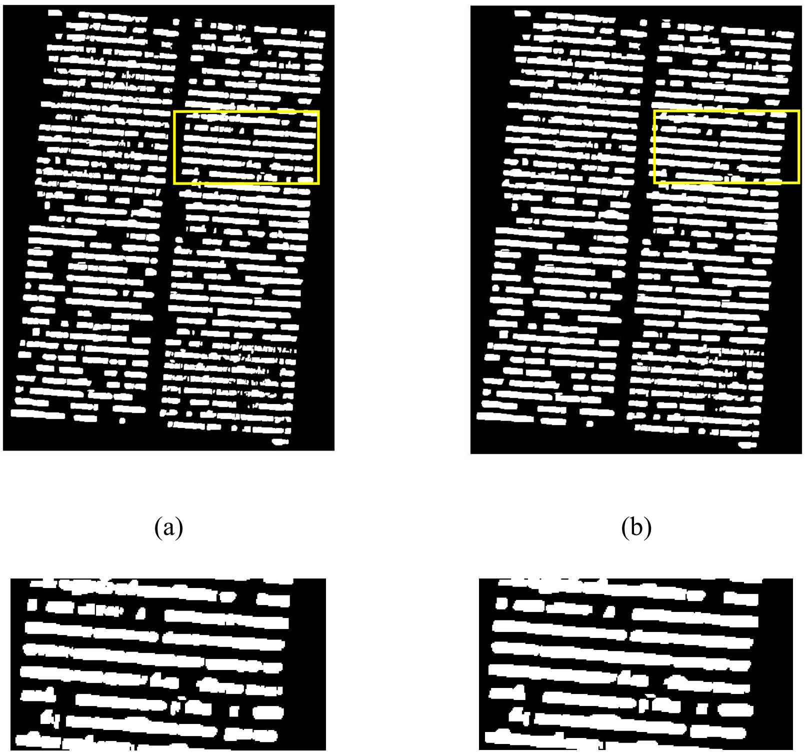


Figure 7: Extraction of the longest connected component CC_{ISR}

3.5 Application of an Oriented Binary Morphology

CC_{ISR} is too small to efficiently represent the skew of the text line. To correctly estimate the text skew all connected components should be extended. Hence, a morphological operation of dilation is applied to matrix **C**. Adjacent connected components are merged establishing parts of the text line. The first approach applies an algorithm which incorporates a non-oriented morphology. It uses a structuring element, $\mathbf{S}_{f,}$ which is an horizontal line with fixed width. It should be chosen carefully in order to not connect separate neighbour text lines. Its width heavily depends on each connected component height. Empirically, it is given as 30% of CC's height. The second approach includes an oriented morphology. It utilizes the structuring element $\mathbf{S}_v$, representing a fixed width line with variable orientation. This orientation variability is function of ISR. Figure 8 shows image $\boldsymbol{Y}_{\mathrm{f}}$ and $\boldsymbol{Y}_v$ established by dilating **C** with structuring elements $\mathbf{S}_f$ and $\mathbf{S}_v$, respectively.



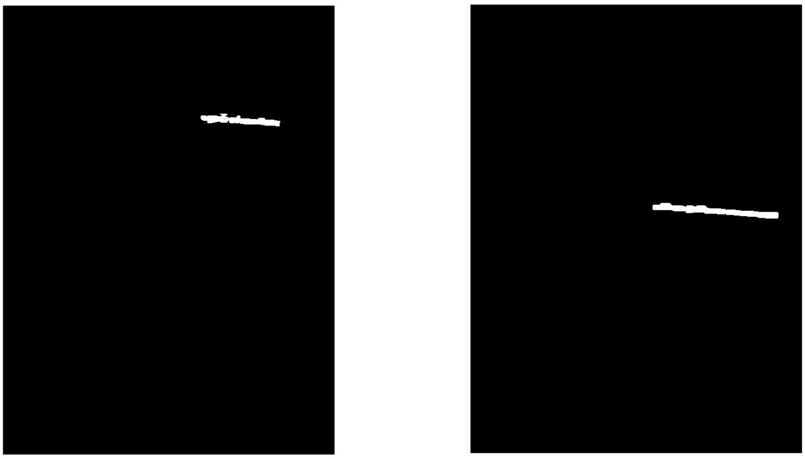
(a) (b)

(c) (d)

Figure 8: Document after morphological operation: (a) with element $\mathbf{S}_f$, *(b) with element* $\mathbf{S}_v$, *(c) zoomed fragment from (a), and (d) zoomed fragment from (b).*

3.6 Extraction of the Longest Connected Component

From Figure 8, the longest connected component commonly represents the text line. Accordingly, it incorporates the skew characteristics of the text line. Hence, the extraction of the longest connected component can detect the text skew. Figure 9 illustrates this circumstance.



(a) (b)

Figure 9: Longest connected component extraction: (a) from image $\mathbf{Y}_f$*(*CC_f*), (b) from image* $\mathbf{Y}_v$ *(*CC_v*).*

From image $\mathbf{Y}_f$ (or $\mathbf{Y}_v$), the longest connected component CC_f (or CC_v) is extracted with LCS [Brodić, 11]. The determination of its orientation estimates the global text skew in the document image.

3.7 Skew Estimation of the Longest Connected Component based on Moments

There exist many techniques to estimate the orientation of a set of pixels within an image. Due to its robustness, the last stage of the proposed algorithm incorporates the use of moments for this task.

Moments define the measure of the pixel distribution in an image. Furthermore, this information depends on the object contour in the image. Moments of the binary image **Y** featuring M rows and N columns are evaluated as [Kapogiannopoulos, 02]:

$$m_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} i^p j^q \ , \qquad (1)$$

where p and $q = 0, 1, 2, 3, ..., n$, and n represents the order of the moment. The central moment μ_{pq} for a binary image **Y** can be calculated as:

$$\mu_{pq} = \sum_{i=1}^{N} \sum_{j=1}^{M} (i - \bar{x})^p (j - \bar{y})^q \ . \qquad (2)$$

where ($\bar{x}$, $\bar{y}$) is the centroid.

The orientation θ of the object in the image can be obtained using moments. The angle θ determines the angle between the object and the horizontal axis. Its estimation is given by [Kapogiannopoulos, 02]:

$$\theta = \frac{1}{2}\arctan\left(\frac{2\mu_{11}}{\mu_{20} - \mu_{02}}\right). \tag{3}$$

with μ_{pq} evaluated using eq.(2).

Using eq.(3), the global skew is identified. Accordingly, the original image is subjected to skew correction. The result is given in Figure 10.

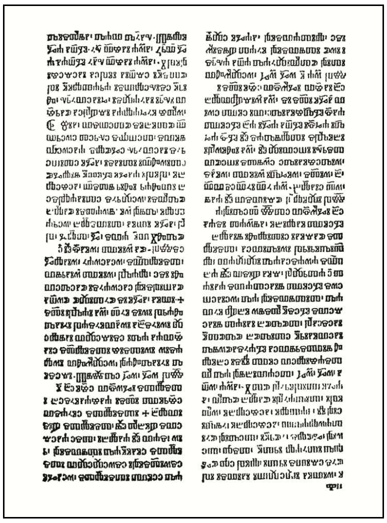
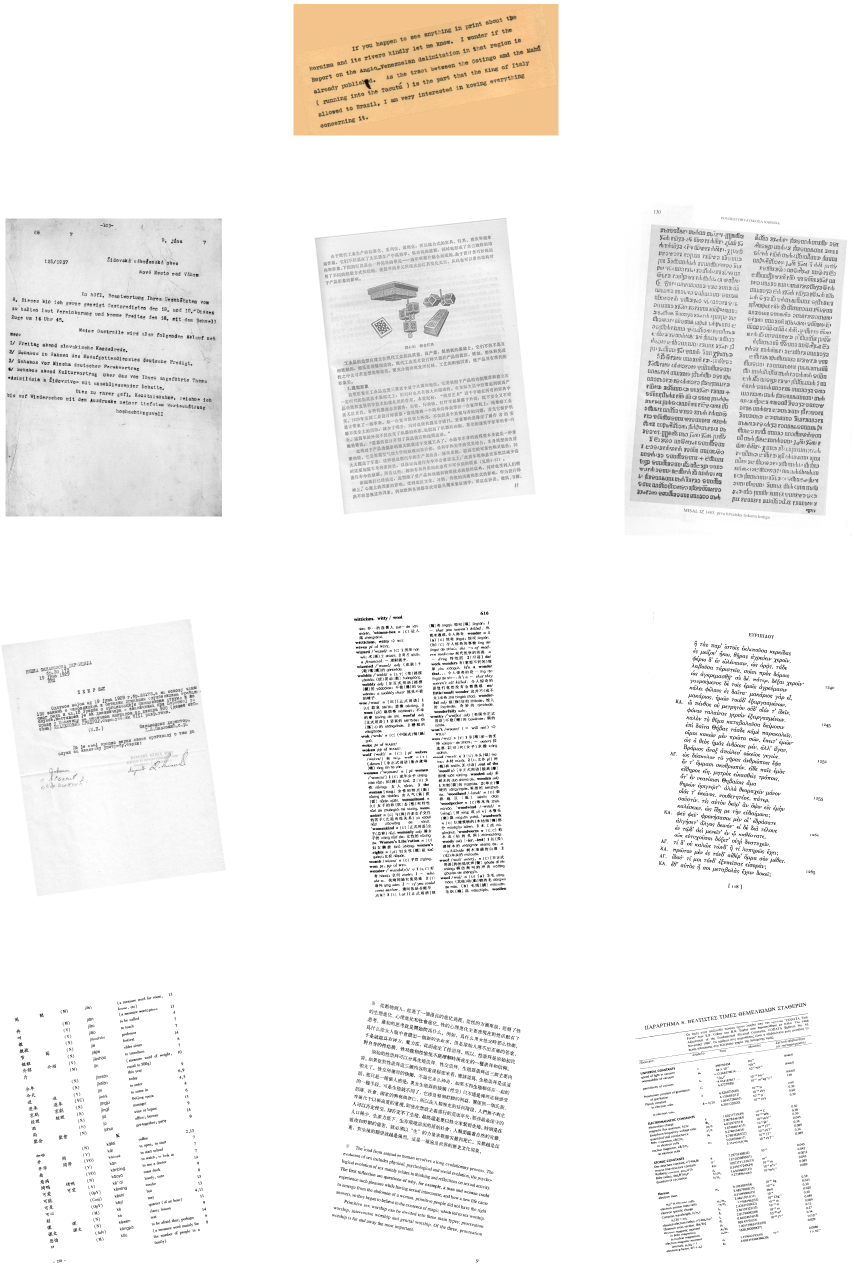


Figure 10: Sample document of Figure 1(a) subjected to global text skew correction.

4 Experiments

The experiments evaluate the algorithm's ability to estimate text skew. It is performed on real documents that come from a custom dataset. The dataset consists of ten printed documents (some of them represent vintage documents). They originate from different sources (50% from DISEC'13) [ICDAR, 13] (Figure 11).

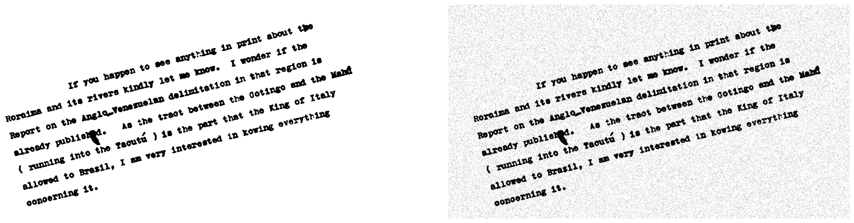


If you happen to see anything in print about the Roraima and its rivers kindly let me know. I wonder if the Report on the Anglo_Venezuelan delimitation in that region is already published. As the tract between the Cotingo and the Mahú (running into the Tacutú) is the part that the King of Italy allowed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

Figure 11: Dataset samples (Latin, Cyrillic, Glagolitic, Greek Cyrillic, Chinese, and Mixed)

For each document sample, ten instances are created by randomly rotating the document image. The rotation is made according to the angle θ_{REF}, which is changed from -15° to 15° by random step in the positive or negative direction. All document images from the dataset are digitized with 300 dpi resolution.

A second experiment is also performed with the adding of salt and pepper noise with densities from 0.01 to 0.05 with step 0.01. Figure 12 illustrates a sample document with and without noise.



If you happen to see anything in print about the
Roraima and its rivers kindly let me know. I wonder if the
Report on the Anglo_Venezuelan delimitation in that region is
already publish[illegible]d. As the tract between the Cotingo and the Mahú
(running into the Tacutú) is the part that the King of Italy
allowed to Brazil, I am very interested in kowing everything
concerning it.

If you happen to see anything in print about the
Roraima and its rivers kindly let me know. I wonder if the
Report on the Anglo_Venezuelan delimitation in that region is
already publish[illegible]d. As the tract between the Cotingo and the Mahú
(running into the Tacutú) is the part that the King of Italy
allowed to Brazil, I am very interested in kowing everything
concerning it.

(a) (b)

Figure 12. Document image: (a) without noise, (b) with 0.05 salt & pepper noise

Salt and pepper noise is a very common issue sometimes generated from an inappropriate thresholding stage (especially in old documents where the paper can have high levels of degradation).

All results are evaluated by the absolute deviation:

$$\Delta\theta_A = \left|\theta_{REF} - \theta_A\right|, \tag{4}$$

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

Furthermore, the measure average absolute deviation $\Delta\overline{\theta}$ is introduced. It represents the average value of all the absolute deviation values, which are obtained for document samples rotated in different angles.

$$\Delta\overline{\theta} = \frac{1}{K}\sum_{l=1}^{K}\Delta\theta_l\ , \tag{5}$$

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

5 Experimental Results and Analysis

All results are given for both versions of the algorithm, i.e., with and without ISR, in order to choose the best one. Furthermore, the results are compared to a sophisticated algorithm for text skew estimation [Mascaro, 10].

5.1 Results of Experiment 1

The first experiment tested all algorithms applied to the document images given in Figure 11, skewed at up to 15 degrees in any direction. Table 1 shows the absolute deviation $\Delta\theta$ for ten instances of ten different document samples.

Dataset sample	**Algorithm [Mascaro, 10]**	**Algorithm w/o ISR**	**Algorithm with ISR**
Latin fragment	0.1625	0.0510	0.0677
Latin 2	0.1000	0.1469	0.0968
Chinese 1	0.0250	1.6454	0.0497
Glagolithic 1	0.1930	0.3370	0.0920
Cyrillic 1	0.1150	1.4860	0.3230
IMG(004)*	0.4260	0.2722	0.1193
IMG(009)*	0.1400	0.8389	0.1997
IMG(014)*	0.1530	1.7154	0.1342
IMG(015)*	0.1540	0.8929	0.1945
IMG(016)*	0.0940	0.1769	0.1288
Average deviation	0.1563	0.7563	**0.1406**

Table 1: Absolute deviation $\Delta\theta$ for the skew angle range from -15° to +15°
(represents samples from DISEC'13)*

To correctly evaluate the given results, two criteria are important: absolute deviation range value of document sample instances and average absolute deviation. From Table 1, the algorithm [Mascaro, 10] and the algorithm with ISR have a far better average deviation than the algorithm w/o ISR. Mascaro et al.'s algorithm has absolute deviation from 0.0250 to 0.4260 with average deviation of 0.1563, while the algorithm with ISR has absolute deviation from 0.0497 to 0.3230 with average deviation of 0.1406. Accordingly, the average absolute deviation of the proposed algorithm compared to Mascaro et al.'s algorithm is about 10% better. Furthermore, the worst absolute deviation result is 0.3230 compared to 0.4260.

5.2 Results of Experiment 2

To prove the versatility of the proposed algorithm, it is tested on document images that have skew angles between 0° and 40°. Mascaro et al.'s algorithm is not expected to work with such wide range of text skew angles. However, it is used as a reference to show the strong points of our approach. Table 2 shows the results from the experiment with document images rotated between 0º and 40º.

Dataset sample	**Algorithm** [Mascaro, 10]	**Algorithm w/o ISR**	**Algorithm with ISR**
Latin fragment	3.5524	0.0780	0.1014
Latin 2	3.5048	0.1843	0.1073
Chinese 1	3.4476	2.3311	0.0502
Average deviation	3.5016	0.8644	**0.0863**

Table 2: Absolute deviation Δθ for the skew angle range 0º-40º

From Table 2, Mascaro et al.'s algorithm is unsuitable for the given skew angle range. By the other hand, both proposed algorithms are suitable for the wide range of text skew with the algorithm with ISR as a clear winner.

5.3 Results of Experiment 3

In the third experiment, the document image incorporates different levels of salt and pepper noise. These levels are in the range between 0.01 and 0.05. The results of the experiment with noisy document images are shown in Table 3.

Dataset sample	**Algorithm [Mascaro, 10]**	**Algorithm w/o ISR**	**Algorithm with ISR**
Latin fragment	5.2833	0.2065	0.1631
Latin 2	2.2667	0.2854	0.0970
Chinese 1	1.0167	1.0959	0.2281
Glagolithic 1	0.8500	0.5967	0.0283
Cyrillic 1	3.6433	0.3317	0.3333
IMG(004)*	0.1033	0.2295	0.0695
IMG(009)*	1.0833	1.0859	1.0408
IMG(014)*	0.4400	1.8804	0.2152
IMG(015)*	0.1300	0.1042	0.1398
IMG(016)*	0.1800	0.2436	0.1727
Average deviation	1.4997	0.6060	**0.2488**

Table 3: Absolute deviation Δθ for the skew angle range from -15º to +15º with salt & pepper noise from 0 to 0.05 by step 0.01 (represents samples from DISEC'13)*

Table 3 shows that both proposed algorithms are suitable for text skew estimation in noisy document images. Compared to the algorithm in [Mascaro, 10], the algorithm w/o ISR and the algorithm with ISR have 2.5 times and 6 times better estimation of the text skew. Such a good result shows another advantage of the proposed approach. At the end, the versatility of our approach is proven by testing a dataset which incorporates documents given in Latin, Cyrillic, Glagolitic and Chinese languages.

6 Conclusions

This paper presents a robust method for estimation of global text skew of printed documents. It is based on the convex hull's extraction over the text in the document image. Furthermore, convex hulls are extended with non-oriented as well as with oriented binary morphology determined by initial skew rate. After that, the longest object is extracted. A moment based approach estimates its orientation, which represents the global text skew. We have proposed two methods: one that includes the influence of an initial skew rate (ISR) and the other that does not. Both methods are examined on a real dataset which includes document images in Latin, Cyrillic, Glagolitic and Chinese languages given at 300 dpi resolution. The method with oriented morphology shows advantages over the other method. Comparison with the state of the art Mascaro et al.'s method shows clear advantage of the proposed approach with ISR, especially in the case of noisy documents. In near future, the proposed method should be extended by incorporating more sophisticated geometrical filtering in pre-processing stage in order to further reduce the error rate of text skew estimation for variety types of documents.

Acknowledgements

This work was partially supported by the Grant of the Ministry of Education, Science and Technological Development from Republic of Serbia, as a part of the project TR33037 and III43011 within the framework of Technological development program.

References

[Amin, 05] Amin, A., Wu, S.: Robust Skew Detection in Mixed Text/Graphics Documents, In Proc. of 8th International Conference on Document Analysis and Recognition, vol.1, pp.247–251, 2005.

[Baird, 87] Baird, H. S.: The Skew Angle of Printed Documents, In Proc. of SPSE 40th Symposium on Hybrid Imaging Systems, pp 739-743, 1987.

[Ballard, 81] Ballard, D. H.: Generalizing the Hough Transform to Detect Arbitrary Shapes, Pattern Recognition, vol.13, No.2, pp.111-122, 1981. doi: 10.1016/0031-3203(81)90009-1.

[Brodić, 11] Brodić, D.: The Evaluation of the Initial Skew Rate for Printed Text, Journal of Electrical Engineering - Elektrotechnický časopis, vol.62, no.3, pp.134-140, 2011.

[Brodić, 12] Brodić, D., Milivojević, D. R.: An algorithm for the Estimation of the Initial Text Skew, Information Technology and Control, 2012, Vol. 41, No. 3, pp. 211–219.

[Brodic, 13] Brodic, D. Milivojevic, Z. N.: Log-polar Transformation as a Tool for Text Skew Estimation, Elektronika ir Elektrotechnika, vol.19, no.2, pp.61-64, February, 2013. doi:10.5755/j01.eee.19.2.3471.

[Brodic, 14] Brodić, D., Maluckov, Č. A., Peng, L.: Statistics Oriented Preprocessing of Document Image, Computing and Informatics, *In press*.

[Chou, 07] Chou, C., Chu, S., Chang, F.: Estimation of Skew Angles for Scanned Documents Based on Piecewise Covering by Parallelograms, Pattern Recognition, vol.40, no.2, pp.443-455, 2007. doi:10.1016/j.patcog.2005.10.030.

[Flusser, 09] Flusser, J., Zitova, B., Suk T.: Moments and Moment Invariants in Pattern Recognition, John Wiley & Sons, 2009.

[Hashizume, 86] Hashizume, A., Yeh, P. S., Rosenfeld, A.: A method of Detecting the Orientation of Aligned Components, Pattern Recognition Letters, vol.4, no.4, pp. 125-132, April, 1986. doi: 10.1016/0167-8655(86)90034-6.

[ICDAR, 13] ICDAR 2013 - Document Image Skew Estimation Contest (DISEC'13), http://users.iit.demokritos.gr/~alexpap/DISEC13/.

[Johannsen, 82] Johannsen, G., Bille, J.: A Threshold Selection Method using Information Measures, In Proc. of 6th International Conference on Pattern Recognition, pp. 140-143, 1982.

[Kapogiannopoulos, 02] Kapogiannopoulos, G., Kalouptsidis, N.: A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments, In Proc. of SPPRA, pp. 275–279, 2002.

[Le, 93] Le, D. X., Thoma, G. R.: Document Skew-angle Detection Algorithm, In Proc. SPIE 1961 - Visual Information Processing II, 251, 1993, doi:10.1117/12.150957.

[Lu, 03] Lu, Y., Tan, C. L.: Improved Nearest Neighbor Based Approach to Accurate Document Skew Estimation, In Proc. of the Seventh International Conference on Document Analysis and Recognition, pp.503-507, 2003.

[Makridis, 10] Makridis, M., Nikolau, N., Papamarkos, N.: An Adaptive Technique for Global and Local Skew Correction in Color Documents, Expert Systems with Applications, vol.37, no.10, pp. 6832-6843, 2010. doi:10.1016/j.eswa.2010.03.041.

[Manmatha, 99] Manmatha, R. Srimal, N.: Scale Space Technique for Word Segmentation in Handwritten Manuscripts', M. Nielsen et al. (editors), Scale-Space Theories in Computer Vision, Lecture Notes in Computer Science 1682, pp.22-33, 1999. doi:10.1007/3-540-48236-9_3.

[Mascaro, 10] Mascaro, A. A., Cavalcanti, G. D. C., Mello, C. A. B.: Fast and Robust Skew Estimation of Scanned Documents Through Background Area Information, Pattern Recognition Letters, vol.31, no.11, pp.1403-1411, 2010. doi:10.1016/j.patrec.2010.03.016.

[Mello, 14] Mesquita, R. G., Mello, C. A. B., Almeida, L. H. E. V.: A New Thresholding Algorithm for Document Images based on the Perception of Objects by Distance, Integrated Computer-Aided Engineering, vol.21, no.2, pp.133-146, 2014.

[Najman, 04] Najman, L.: Using Mathematical Morphology for Document Skew Estimation, SPIE Document Recognition and retrievals XI, vol. 5296, pp 182-191, 2004. doi:10.1117/12.526615.

[O'Gorman, 93] O'Gorman, L.: The Document Spectrum for Page Layout Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.15, no.11, pp. 1162-1173, November, 1993. doi: 10.1109/34.244677.

[Pal, 96] Pal, U., Choudhuri, B. B.: An Improved Document Skew Angle Estimation Technique, Pattern Recognition Letters, vol.17, no. 8, pp.899-904, July, 1996. doi: 10.1016/0167-8655(96)00042-6.

[Palmer, 99] Palmer, S. E.: Vision Science Photons to Phenomenology, MIT Press, 1999.

[Postl, 86] Postl, W.: Detection of Linear Oblique Structures and Skew Scan in Digitized Documents, In Proc. 8th International Conference on Pattern Recognition, pp.687-689, 1986.

[Rehman, 11] Rehman, A., Saba, T.: Document skew Estimation and Correction: Analysis of Techniques, Common Problems and Possible solutions, Applied Artificial Intelligence, vol. 25, no.9, pp.769-787, 2011.

[Saragiotis, 08] Saragiotis, P., Papamarkos, N.: Local Skew Correction in Documents, International Journal of Pattern Recognition and Artificial Intelligence, vol.22, no.4, pp.691-710, 2008. doi:10.1142/S0218001408006417.

[Shivakumara, 05] Shivakumara, P., Kumar, G. H., Guru, D. S., Nagabhushan, A Novel Technique for Estimation of Skew in Binary Text Document Images based on Linear Regression Analysis, Sadhana, vol.30, no.1, pp.69-85, 2005.

[Wackerly, 96] Wackerly, D. D., Mendenhall, W. III., Scheaffer R. L.: Mathematical Statistics with Applications, Duxbury Press, Belmont, U.S.A., 1996.

[Yan, 93] Yan, H.: Skew Correction of Document Images Using Interline Cross-correlation, Computer Vision, Graphics, and Image Processing, vol.55, no.6, pp.538-543, November, 1993. doi:10.1006/cgip.1993.1041.

[Yu, 96] Yu, B., Jain, A. K.: A Robust and Fast Skew Detection Algorithm for Generic Documents, Pattern Recognition, vol.29, no.10, pp.1599-1629, 1996.

[Zagoris, 11] Zagoris, K., Chatzichristofis, S. A., Papamarkos, N.: Text Localization using Standard Deviation Analysis of Structure Elements and Support Vector Machines, EURASIP Journal on Advances in Signal Processing, vol.2011, no.47, pp.1-12, 2011. doi: 10.1186/1687-6180-2011-47.