A Skew Detection Technique Suitable for Degraded Ancient Manuscripts

Florian KLEBER – Robert SABLATNIG

Institute of Computer Aided Automation, Vienna University of Technology, Vienna, Austria
{kleber; sab}@prip.tuwien.ac.at

Abstract

In order to preserve our cultural heritage and for automated document processing libraries and National Archives have started to digitize historical documents. Automated document analyzing can comprise tasks like character segmentation, text extraction and layout analysis. In the case of degraded manuscripts (e.g. by mold, humidity, bad storage conditions) parts of the text disappear (e.g. be washed out) and therefore only fragments of an entire page are still visible. Due to the digitization process of ancient manuscripts/documents a rotation angle is introduced to the captured images in relation to the original image axes. The automated skew correction of documents is necessary because the algorithms for document processing are sensitive to the page skew and the amount of pages is too large to do it manually. This paper presents an algorithm for skew estimation that can be applied to handwritten degraded documents, and also for other archaeological relevant documentations like type written cards or reports.

Keywords

Document Analysis, skew estimation

1. Introduction

Digitalization permits a detailed storage and organisation of the contents of a document and a worldwide exchange or access of documents (e.g. via the internet). If entire libraries or parts of libraries of National Archives or museums are digitized manual analysis or digital restoration of documents by a human will not be feasible due to costs and time.

Projects and institutions that are dealing with the digitalization of documents are amongst others the digitalization center of the university library Graz[1], Vienna and *DIAMM: Digital Image Archive of Medieval Music, University of Oxford and Royal Holloway University of London*[2]. Methods for e.g. the layout analysis of handwritten historical documents (Bulacu *et al.* 2007), text line segmentation (Louloudis *et al.* 2006; Likforman-Sulem *et al.* 2007) in handwritten documents and automated layout segmentation and classification of printed documents (Cinque *et al.* 2003) have been published. For correct solutions these algorithm estimate that the input image has no skew. Additional "*for humans, rotated images are unpleasant for visualisation and introduce extra difficulty in text reading*" (Lins and Avila 2004). As a result algorithms for skew correction are necessary.

In (Chen *et al.* 1995) skew is defined as follows: "*The text skew angle of a document image is denoted by φ and is defined as its dominant (most frequently occuring) text baseline direction*". They (Chen *et al.* 1995) define also a formalization of the skew estimation problem: "*Given a document image I with text lines at unknown skews of* $\varphi_1, \ldots, \varphi_n$. *Find , an estimate of the true document text skew angle φ, to maximize the probability* $P(\varphi|I)$". Since in skewed printed and handwritten documents the baselines of the text (as well as text lines) are supposed to be parallel, the skews $\varphi_1, \varphi_2, \ldots, \varphi_n$ are equal (see Section 2.). The algorithm proposed in this paper assumes that it is possible that different lines differ in skew and also lines are not written straight according to a baseline as it can be seen in *Fig. 1*. Therefore connected components (related to words, part of words) are analysed for each line and the skew is calculated such that the distribution of skew angels is minimized. The algorithm was tested on synthetic data and on images of the folios of the Missale Sinaiticum (Cod. Sin.slav.5/N, contain Glagolitic text).

This paper is organized as follows. Section 2 reviews the state of the art of skew detection methods

[1] http://www.kfunigraz.ac.at/ub/sosa/digitalisierung/, accessed feb. 2008

[2] http://www.diamm.ac.uk/, accessed feb. 2008

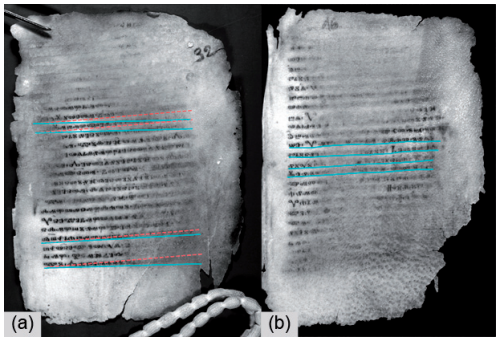

Fig. 1. (a) different line skews within a single folio (b) different skews within a single line.

and explains why the different algorithms are not suitable for degraded text. In Section 3 the proposed algorithm is described in detail while in Section 4 the experimental results are discussed.

2. Related work

Methods proposing algorithm for skew estimation include techniques based on projection profiles (Bagdanov and Kanai 1997; Su *et al.* 2007), the Hough transformation (Hinds *et al.* 1995; Amin and Fischer 2000) morphological based skew estimation (Chen and Haralick 1994) and methods based on properties of the Fourier transform. A summary and a classification of skew estimation algorithm is shown in (Lins and Avila 2004) and Hull (Hull 1998). Algorithms based on projection profiles determine the horizontal histogram which is obtained by summing pixel values along a horizontal projection line. The distribution of the horizontal histogram is analysed to determine the correct skew. In the case of degraded documents, where it is assumed that parts of the document are faded/washed out, lines or parts of lines disappear and will produce only small peaks. A second drawback of projection profile based techniques is that narrow lines do not produce a significant peak (Likforman-Sulem *et al.* 2007).

A different solution to estimate the skew angle is to calculate feature points and apply the Hough transform on the image to find straight lines. In (Egozi *et al.* 2007) a statistical model containing straight lines that are distorted by Gaussian noise is used to determine the skew. In (Safabakhsh and Khadivi 2000) the document image is smoothed and then a bounding rectangle of the Connected Components (CC) with a minimal area is calculated. The orientation of the bounding rectangle determines the skew of the document.

Other methods include solutions based on clustering (Avila and Lins 2005; Okun *et al.* 1999), Principle Component Analyzes (PCA) (Sarfraz *et al.* 2007; Steinherz *et al.* 1999) and moment based analysis. According to (Kavallieratou *et al.* 2002) the problems of skew estimation algorithms are the following:

1. "Restriction of detectable angle range
2. Restriction on type or size of fonts
3. Dependence on page layout
4. A specific document resolution is required
5. High computational cost
6. Limitation to specific application
7. Large text areas are required
8. [...] Furthermore, the proposed algorithms can estimate the dominant skew angle and cannot deal with the cases of handwritten pages where the text lines may not be parallel to each other."

All cited skew estimation methods are always solving only parts of the listed problems since they are designed for specific styles of documents. A general solution, which solves all of the listed problems is, to our knowledge, not published. The proposed skew estimation method is able to handle handwritten pages where the text lines are not parallel or straight (8) and is restricted to a defined angle range (1) of ±85°.

Fig. 1 (a) shows a folio with different line skews (slashed red line indicates the skew of the upper part of the folio whereas the full turquoise line indicates the skew of the bottom lines of the folio) and *Fig. 1 (b)* visualizes differences in declination within single lines. A restriction that is not limited according to the skew has to consider single characters or words and therefore perform an OCR. Additional the size and type of the font (within an page) is not limited to a specific type.

3. Proposed algorithm

Working with degraded and poor quality documents need a pre-processing stage that eliminates distortions. The noise removal is followed by a initial skew estimation. This is done to align the page within a skew range of ±15°, because within this range the calculation of the Adaptive Local Connectivity Map (ALCM, see Section 3.2) leads to blobs that belong to words or even sentences and are not fragmented. Therefore a calculation of the alignment of the

different blobs can be done: for the pixels of every CC a corresponding orientation is calculated by principle component analysis. The skew of every line is weighted according to the area size and length of the corresponding pixel area. As a result the pre-dominant words/lines determine the correct skew. The single steps are described in detail in the following subsections.

3.1. Pre-processing

In the pre-processing stage first the entire image is binarized using an adaptive algorithm suitable for degraded document image binarization (Gatos *et al.* 2006). This algorithm extracts an estimated background surface of the image, from which the original image is subtracted and afterwards thresholded. The parameters for the estimation of the foreground regions are chosen as suggested in (Gatos *et al.* 2006; Sauvola and Pietikainen 2000). The size of the interpolation window dx × dy for the estimation of the background has to cover at least two image characters (Gatos *et al.* 2006); therefore a size of 131 × 131 pixels (images of the Missale Sinaiticum) is chosen. Integral images are used for an efficient implementation of the binarization algorithms (Shafait *et al.* 2008). A comparison of different thresholding techniques is also done in (Gatos *et al.* 2006). After the binarization the page is segmented and an additional noise removal is performed by applying a 9 × 9 median filter. *Fig. 2 (a)* shows the original view of a folio of the Missale Sinaiticum and *Fig. 2 (b)* shows the result of the binarization. The selected and median filtered folio is shown in *Fig. 2 (c)*. It is estimated that after the binarization and the noise removal every information in the image belongs to text or parts of the text. To calculate an initial estimation of the skew, the page is rotated in 15° steps and the image is horizontally smoothed using a convolution to fill the white gaps between consecutive symbols/characters (words). If the text direction is not parallel to the direction of the smoothing, connected components, which are related to words are subdivided in different blobs. The initial skew is determined by maximizing the area property of the single blobs. *Fig. 3* shows the size of the single blobs dependend on the skew. Note that this approach is independent from the type of the characters. The size of the filter kernel used for horizontal smoothing must be choosen as large as the gaps between single characters/words. In this application the filter width for smoothing is 60 pixels. *Fig. 3 (a)* shows the same text written in different orientations. *Fig. 3 (b)–(c)* shows the smoothed image with 3 different orientations of the smoothing filter (0°, 10° and 20°). It can be seen that filters with the same orientation as the text produce blobs with the largest area and length

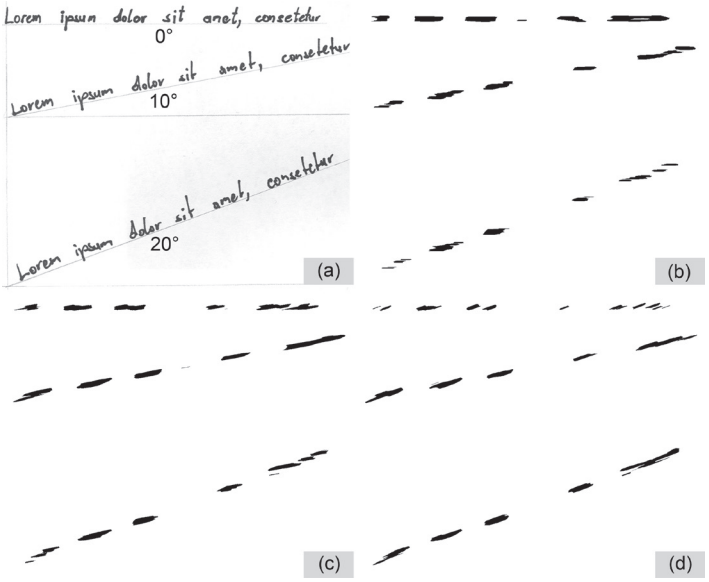

Fig. 3. (a) Original Image (b) 0◦ filtered image (c) 10◦ filtered image (d) 20◦ filtered image.

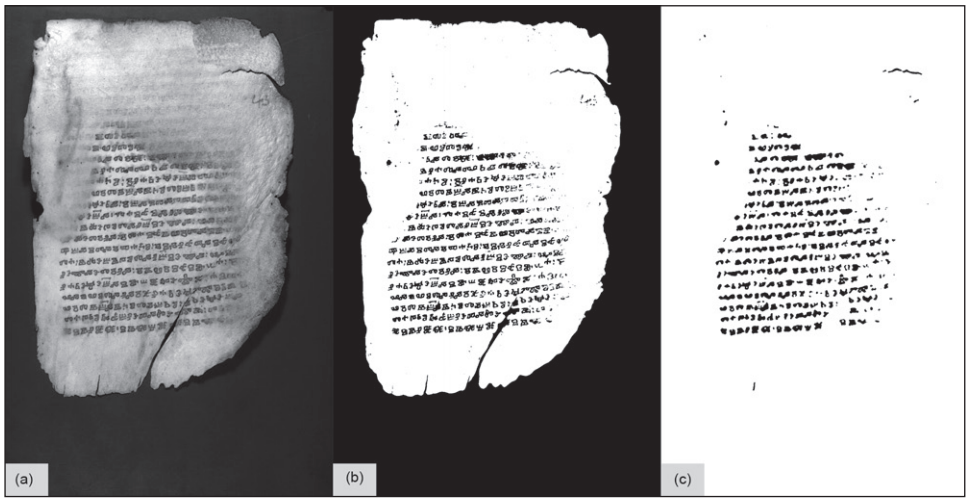

Fig. 2. (a) Original Image (b) Binarized Image using (Gatos et al. *2006) (c) Binarized and Median filtered Image.*

while blobs with other orientations are fragmented. The result of the initial estimation has an accuracy of ±15°.

3.2. Skew estimation

After the pre-processing step the image is rotated with the previously calculated skew. It is estimated that the orientations of the single words lie within the accuracy of the initial skew and is therefore ±15°. To calculate the final skew, an ALCM (Shi *et al.* 2005) as connectivity measure is calculated (see *Eq. 1* and *Eq. 2*, (Shi *et al.* 2005)).

$$I(x,y)_{ALCM} = \int_{x,y} I(x,y)G_c(t-x,y)dt \quad \textit{Eq. 1}$$

where

$$G_c(x,y) = \begin{cases} 1, & if|x| < c \\ 2, & otherwise \end{cases} \quad \textit{Eq. 2}$$

According to (Shi *et al.* 2005) c is a fixed value and determines the size of the sliding window and is "*approximately equal to three time the average height of text*". In this application c is set to 140. Afterwords the image is labelled with a standard label algorithm. The blobs with the highest connectivity (and therefore length) are considered more significant (small blobs are dedicated as outliers). The determine the skew of every single blob a line is fitted into the blob (see *Fig. 4*, *Eq. 3*) using Principle Component Analysis (PCA) (Forsyth and Ponce 2002). The CC are considered as a scatter plot and PCA is used to define a new axes (orientation), which goes through the maximum variation in the data (or equivalent minimizes the square of distance of each point to that axes).

$$\forall CC_{xy} : PCA\{(x,y) \in CC_{xy}\} \quad \textit{Eq. 3}$$

arctan(dx/dy) and lenxy = $dx^2 + dy^2$. The obtaining values are weighted according to the length of each object and the mean of the angles are calculated, which defines the skew of the page. Therefore Equation 4 defines the skew of the input image.

$$skew = \frac{\sum_{(x,y)} skew_{xy} * len_{xy}}{\sum_{(x,y)} len_{xy}} \quad \textit{Eq. 4}$$

4. Results

The algorithm was tested on chosen images of the Missale Sinaiticum with different dimensions of degradations. As shown in *Fig. 1* the orientation of the lines of folios of the Missale Sinaiticum can be different. To get results of a page where all lines have the same skew, the algorithm was also tested on a parallel ruled and handwritten page. *Fig. 5 (a)* shows the input image with a skew of −45° (all lines have the same orientation) and *Fig. 5 (b)* shows the orientations (red full line) of all connected components. The binarized and smoothed image is taken as input image for the initial skew estimation. The calculated skew in the preprocessing step is −12° (accuracy of the initial skew estimation is within the range of ±15°). The dominant orientation of the calculated orientation is chosen as correct skew estimation and is shown in *Fig. 5 (c)* after applying the proposed algorithm. To test the algorithm, the page was rotated from −85° to +85° in 5° steps. Additional the algorithm was tested on chosen folios of the Missale Sinaiticum with different dimensions of dedgradations which

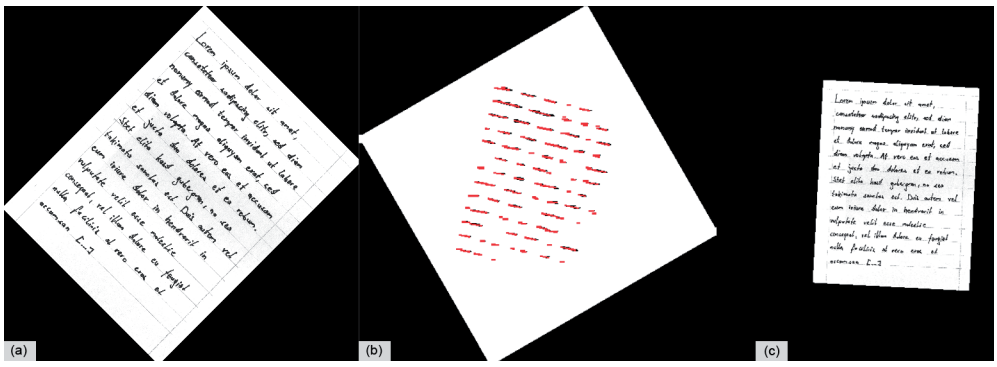

Fig. 5. (a) Input Image (b) initial skew estimation and calculated PCA (red line) for every blob (c) corrected Image.

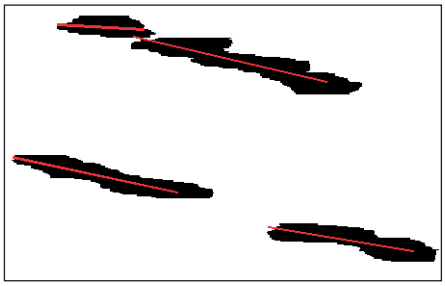

Fig. 4. Calculated orientation of single blobs (red full line).

Images	Proposed Method		
	Mean	Standard Deviation	Max Deviation
Handwritten page	0,64	0,54	1,98
27 verso	0,67	0,34	1,29
29 recto	3,49	1,24	5,41
36 recto	2,26	0,73	3,89
43 recto	0,73	0,47	1,92
44 recto	1,05	0,63	3,21
45 recto	1,41	0,96	3,03

Table 1. Mean, Standard and Max. Deviation of the estimated skew from the correct orientation.

are rotated by 20 random angular values in the range from −85° to +85° . The mean deviation, the standard deviation and the maximal deviation of the calculated skew to the correct skew is summarized in *Table 1*.

Representative folios of the Missale Sinaiticum (regarding appearence and possible dimension of degradation) are shown in *Fig. 1 (a)*, *(b)* and *Fig. 2 (a)*. The size of the images of the Missale Sinaiticum are 4000 × 2672 pixel and mean time needed for the binarization using (Gatos *et al.* 2006) is 8.79 seconds. The mean time for the inital skew estimation is 16.18 seconds and the mean time for the final skew estimation is 4.42 seconds. The explanation for the long time needed for skew estimation is on the one hand the high resolution of the images and on the other hand the search for the initial skew value, which is needed to get the best result regarding the length of the connected components. The best results are achieved on the handwritten page and on folios (27 verso, 43 recto) with a degradation less than 50% (50% of the text is disappeared). If the degradation is higher and additional words/sentences got fragmented the accuracy is less than 2° (see *Table 1*).

5. Conclusion

In this paper a skew estimation algorithm suitable for degraded documents was proposed. For that purpose for every word/line a weight according to the length of each blob is calculated such that after the correction most of the text is horizontally aligned. The algorithm is designed for pages that contain handwritten text without pictures. Degradation of parts of the text are ineffectual as long as determinative parts of the text for the skew are still available. Improvements of the algorithm can be done by introducing layout independency (also images etc. are treated correct) and by developing an OCR (optical character recognition) system to get a non restricted angular value for skew estimation.

Acknowledgement

Authors would like to thank the Austrian Science Fund for funding the project under grant P19608G12.

References

Amin, Adnan and Stephen Fischer (2000). A document skew detection method using the hough transform. *Pattern Analysis and Applications*, *3*(3), 243–253.

Avila, Bruno Tenorio and Rafael Dueire Lins (2005). A fast orientation and skew detection algorithm for monochromatic document images. In: *DocEng '05: Proceedings of the 2005 ACM symposium on Document engineering*. New York, NY, USA: ACM, 118–126.

Bagdanov, Andrew D. and Junichi Kanai (1997). Projection profile based skew estimation algorithm for jbig compressed images. In: *ICDAR '97: Proceedings of the 4th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 401–406.

Bulacu, Marius, Rutger van Koert, Lambert Schomaker and Tijn van der Zant (2007). Layout analysis of handwritten historical documents for searching the archive of the cabinet of the dutch queen. In: *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 1*. Washington, DC, USA: IEEE Computer Society, 357–361.

Chen, Su and Robert Haralick (1994). An automatic algorithm for text skew estimation in document images using recursive morphological transforms. In: *ICIP94*. 139–143.

Chen, Su, Robert Haralick and Ihsin T. Phillips (1995). Automatic text skew estimation in document images. In: *ICDAR '95: Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 2)*. Washington, DC, USA: IEEE Computer Society, 1153–1156.

Cinque, Luigi, Stefano Levialdi and Alessio Malizia (2003). A system for the automatic layout segmentation and classification of digital documents. In: *Image Analysis and Processing,*

2003. Proceedings. 12th International Conference, 201–206.

Egozi, Amir, Itshak Dinstein, Joulia Chapran and Michael Fairhurst (2007). An em based algorithm for skew detection. In: *ICDAR '07: Proc. of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 1*. Washington, DC, USA: IEEE Computer Society, 277–281.

Forsyth, David and Jean Ponce (2002). *Computer Vision: A Modern Approach*. Prentice Hall.

Gatos, Basilis, Ioannis Pratikakis and Stavros J. Perantonis (2006). Adaptive degraded document image binarization. *Pattern Recogn* 39(3), 317–327.

Hinds, Stuart C., James L. Fisher and Donald P. D'Amato (1995). A document skew detection method using run-length encoding and the hough transform. In: *Document Image Analysis*. 209–213.

Hull, John (1998). Document image skew detection: Survey and annotated bibiliography. In: John Hull and Suzanne Taylor (eds.) *Document Analysis System II, World Scientific*. 40–64.

Kavallieratou, Ergina, Nikos Fakotakis and George Kokkinakis (2002). Skew angle estimation for printed and handwritten documents using the wigner-ville distribution. *Image and Vision Computing* 20, 813–824.

Likforman-Sulem, Laurence, Abderrazak Zahour and Bruno Taconet (2007). Text line segmentation of historical documents: a survey. *Int. Journal Document Analysis and Recognition* 9(2), 123–138.

Lins, Rafael Dueire and Bruno Avila (2004). A new algorithm for skew detection in images of documents. In: Aurelio Campilho and Mohammed Kamel (eds.) *ICIAR (2)*. Springer, vol. 3212 of *Lecture Notes in Computer Science*, 234–240.

Louloudis, Georgios, Basilis Gatos, Ioannis Pratikakis and Constantin Halatsis (2006). A block-based Hough transform mapping for text line detection in handwritten documents. In: *Proceedings of the Tenth Int. Workshop on Frontiers in Handwriting Recognition*. 515–520.

Okun, Oleg, Matti Pietikainen and Jaakko Sauvola (1999). Robust skew estimation on low-resolution document images. In: *ICDAR '99: Proceedings of the Fifth International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 621.

Safabakhsh, Reza and Shahram Khadivi (2000). Document skew detection using minimum-area bounding rectangle. In: *ITCC '00: Proceedings of the The International Conference on Information Technology: Coding and Computing (ITCC'00)*. Washington, DC, USA: IEEE Computer Society, 253.

Sauvola, Jakko and Matti Pietikainen (2000). Adaptive document image binarization. *Pattern Recogn* 33(2), 225–236.

Shafait, Faisal, Daniel Keysers and Thomas M. Breuel (2008). Efficient implementation of local adaptive thresholding techniques using integral images. SPIE, vol. 6815, 681510.

Shi, Zhixin, Srirangaraj Setlur and Venu Govindaraju (2005). Text extraction from gray scale historical document images using adaptive local connectivity map. In: *Proceedings of Eighth International Conference on Document Analysis and Recognition, 2005*, *2*, 794–798.

Steinherz, Tal, Nathan Intrator and Ehud Rivlin (1999). Skew detection via principal components analysis. In: *ICDAR '99: Proceedings of the Fifth International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 153.

Su, Tung-Hsin, Tian-Wen Zhang, Hung-Ju Huang and Yun Zhou (2007). Skew detection for chinese handwriting by horizontal stroke histogram. In: *ICDAR '07: Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007) Vol 2*. Washington, DC, USA: IEEE Computer Society, 899–903.