2012 10th IAPR International Workshop on Document Analysis Systems

Skew Estimation of Sparsely Inscribed Document Fragments

Markus Diem, Florian Kleber and Robert Sablatnig
Computer Vision Lab
Vienna University of Technology
Vienna, Austria
{*diem, kleber, sab*}*@caa.tuwien.ac.at*

***Abstract*—Document analysis is done to analyze entire forms (e.g. intelligent form analysis, table detection) or to describe the layout/structure of a document for further processing. A pre-processing step of document analysis methods is a skew estimation of scanned or photographed documents. Current skew estimation methods require the existence of large text areas, are dependent on the text type and can be limited on a specific angle range. The proposed method is gradient based in combination with a Focused Nearest Neighbor Clustering of interest points and has no limitations regarding the detectable angle range. The upside/down decision is based on statistical analysis of ascenders and descenders. It can be applied to entire documents as well as to document fragments containing only a few words. Results show that the proposed skew estimation is comparable with state-of-the-art methods and outperforms them on a real dataset consisting of 658 snippets.**

***Keywords*-skew estimation; rotation; document fragments;**

I. Introduction

The mass digitalization of libraries, national archives or museums needs an automated processing of the acquired image data, comprising digital restoration or document analysis. Projects and institutions that are dealing with the digitalization of documents are amongst others the manuscript research center of Graz University (Vestigia[1]), Improving Access to Text (IMPACT[2]), or projects like Google Books of Google Inc.

Skew estimation is a pre-processing step of document layout analysis and OCR methods. Document analysis can be applied to index digitized images and cluster documents according to their content. Additionally "*for humans, rotated images are unpleasant for visualization and introduce extra difficulty in text reading*" [14]. A different example are OCR methods in mobile applications (Google Goggles, iBing Vision) that use images of mobile devices (e.g. smart phones) [9]. Ephstein [9] points out that current skew estimation algorithms have to be capable to deal with no restriction on the angle ("*full 360 degrees orientation detection*").

Within this paper a skew estimation suitable for sparse inscribed document fragments is presented. The data considered, consists of torn documents inscribed with German, English and Russian text. Thus, snippets have irregular shapes and their content varies from two words up to hundreds of words, either printed or handwritten. The appearance and properties of the dataset ranges from carbon copies, colored paper, lined or checked paper, up to old fashioned copies. The documents have been fragmented in 1989 when Stasi officers tried to destroy secret files [18]. In total, 600 million-odd fragments have been discovered after the fall of the Berlin Wall. Due to the mass of destructed documents and the complexity of matching [6], an automated reconstruction is needed. The Fraunhofer Institute for Production Systems and Design Technology has developed a system for the reconstruction of torn documents [17]. To support the matching algorithm the proposed skew estimation algorithm is applied to each document fragment. The evaluation of the algorithm has been done on a test set of 658 snippets of the Stasi-files and on synthetic images.

The proposed algorithm is based on the text's gradients [19], [16] in combination with a Focused Nearest Neighbor Clustering (FNNC) [11] of interest points. This allows the determination of the orientation up to 180 degree. The upside/down decision is based on statistical analysis of ascenders and descenders, which is dependent on the language and the script. The combination of both methods is able to handle also slanted handwritten text and snippets with at least 2 words. Thus, the detectable angle range is not restricted. Additionally there are no constraints on the layout and using integral images [10] allows for a fast implementation.

This paper is organized as follows: Section II reviews the state-of-the-art of skew estimation methods and gives a definition of skew estimation and the main problems. In Section III the proposed algorithm is described in detail, while Section IV presents the results. Finally, a conclusion is given in Section V.

II. Related Work

In Chen et al. [5] 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 occurring) text baseline direction*". Methods proposing algorithm for skew estimation include techniques based on projection profiles (e.g. [12]), the Hough transformation [1], [9], [8]

[1]www.vestigia.at/, accessed 14.10.2011
[2]www.impact-project.eu/, accessed 14.10.2011

978-0-7695-4661-2/12 $26.00 © 2012 IEEE
DOI 10.1109/DAS.2012.81

morphological based skew estimation [4] and methods based on properties of the Fourier transform.

Algorithms based on projection profiles (e.g. [12]) determine the horizontal histogram which is obtained by summing pixel values along a horizontal projection line. The distribution of the horizontal histogram is analyzed to determine the correct skew. Due to the irregular shape of torn document fragments, lines differ in their length and will produce only small peaks in the histogram. Additionally the method requires "large" text areas (not a single line, or single words).

Calculating the skew by a FNNC method of feature points is presented by Jiang et al. [11]. Feature points within this work are based on connected components and pass codes.

Bar-Yosef et al. [2] use a distance transform of binarized images to determine the skew. The algorithm is evaluated on entire pages from scientific journals/articles (chinese, tables and figures) and Hebrew handwritten text with a skew angle range of $\pm 90\,^{\circ}$. It is stated that *"the dominant orientation of the gradient vectors of the distance transform accurately reflects the skew"* [2].

Ephstein [9] proposes a method based on the Hough transform of the white inter-line regions. It has been tested on 8 synthetic images of printed text and a database of 500 images (books, newspapers, ...) taken with an iPhone 3GS. There is no limitation of the angle range and the "smallest" page of the synthetic images contains 2 words.

III. Methodology

In this paper, a skew estimation is proposed that is based upon two methodologies with different characteristics. The gradient based method accurately detects the main orientation and is robust with respect to sparse document content. Methods based on gradients are presented by Sun and Si [19] and Omar et al. [16]. In contrast, the FNNC [11] is robust with respected to slanted handwritten text and is able to detect the angle up to $180\,^{\circ}$. Combining these two methods allows for a correct skew estimation of documents with various layouts, supporting material and sparse content.

A. Gradient Orientation Measure

The gradient orientation estimation is a pixel based method. Its key concept is that script comprises mostly vertical or horizontal strokes. This assumption can be verified if solely printed text is considered. However, for handwritten text with a slant, the modal angle corresponds to the slant and not to the text line angle. Nevertheless, this methodology has advantages in comparison with the second method presented:

- Fast computation.
- Good performance on small snippets with less than two words content.
- Considers additional information such as ruling.
- Accurate angle estimation (median error $< 0.5\,^{\circ}$).

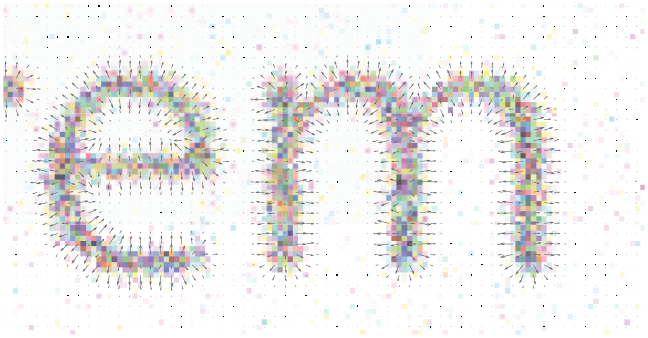

Figure 1. Gradient vectors of noisy text. For illustration a vector is assigned to every 2nd pixel.

Before computing the gradient vector of each pixel, the image is smoothed with a Gaussian kernel ($\sigma = 1.75$) so that noise or background clutter is suppressed. Then, the gradient vectors are computed:

$$f_x(x,y) = f(x+1,y) - f(x-1,y) \tag{1}$$

$$f_y(x,y) = f(x,y+1) - f(x,y-1) \tag{2}$$

$$m(x,y) = \sqrt{f_x(x,y)^2 + f_y(x,y)^2} \tag{3}$$

$$\theta(x,y) = \tan^{-1}\frac{f_y(x,y)}{f_x(x,y)} \tag{4}$$

where $f(x,y)$ represents the image, $m(x,y)$ denotes the gradient magnitude of a pixel (x,y) and $\theta(x,y)$ is the gradient vector's angle. Figure 1 shows the gradient vectors of two characters. Note that the gradients are robust with respect to noise because of the Gaussian smoothing.

These gradient vectors are then accumulated into an orientation histogram that consists of 180 bins representing $[-\frac{\pi}{2}\ \ \frac{\pi}{2})$. The other two quadrants are neglected since a gradient difference of π stands for a black-to-white instead of a white-to-black transition.

In order to build the orientation histogram, the gradient magnitude $m(x,y)$ of each pixel is accumulated to the bin that corresponds with the pixel's angle $\theta(x,y)$. Thus, pixels with a low gradient magnitude (weak edge) are weighted less than those having strong edges. In addition, numerical artifacts and noise is reduced if the gradients are linearly interpolated. The snippet's main orientation is then defined as the maximum of the orientation histogram. In order to improve the method's results, a spline is fit into the maximal bin and its neighbors which results in an accuracy $< 1\,^{\circ}$.

Since the proposed method is not capable to determine a snippet's quadrant and may fail if slanted handwriting is present, it is combined with a second method which can deal with these challenges.

B. Focused Nearest Neighbor Clustering

The second method presented was introduced by Jiang et al. [11]. This method focuses on the words' skew

which allows for a skew estimation up to $180\,^{\circ}$. Even though, it is not as accurate as the previously described method, it is more robust if handwritten snippets are observed.

The FNNC is based on local skew lines that are fit into small subsets of points. In contrast to Jiang et al. [11] we propose to use Difference-of-Gaussians (DoG) [15] interest points for the FNNC, since they are fast to compute and more robust than centroids of connected components. Interest points of a handwritten text are illustrated in Figure 2.

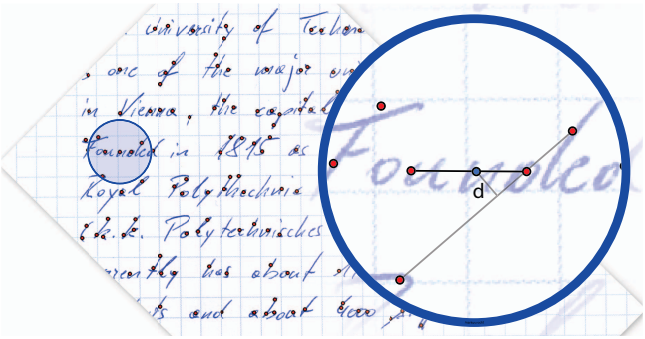

Figure 2. Nearest neighbors with $k = 7$, rejected local skew line having distance d and selected local skew line of the FNNC.

Since the interest points should represent roughly centers of characters rather than junctions or corners, a large scale is chosen. Experiments showed that the second scale in the third octave of the Gaussian scale space is a good choice for the feature point selection. Thus, the image is resized by $\frac{1}{4}$ and subsequently smoothed with a Gaussian kernel having $\sigma = 2$. Since the interest points represent characters rather than edges or junctions the word's orientation is observed rather than its slant angle.

The k nearest neighbors $n_1, n_2, ..., n_k$ of each interest point p are then regarded. For the local skew line computation, the point pair n_i, n_j needs to be found whose connecting line minimizes the distance to p:

$$d = \frac{|n * (p - n_i)|}{|n|} \tag{5}$$

with n being the normal vector of n_i, n_j and d is the distance of the connecting line to p. If the points n_i, n_j are close, the accuracy of the local skew line suffers from small deviations of the interest points. In order to find a robust local skew line, the longest line connecting p, n_i, n_j is chosen. Figure 2 shows a cluster with 7 interest points. The gray line shows a rejected local skew line with distance d to p. In addition, the final local skew line is illustrated.

The snippet's dominant angle is determined by accumulating the local skew lines' angles to an orientation histogram having 180 bins. In contrast to the proposed method [11] we linearly interpolate the skew lines to its two nearest neighboring bins which reduces artifacts. In addition a weight is introduced, that rewards close matches:

$$\omega_i = \frac{1}{1 + d_i} \tag{6}$$

where ω_i is the weight of the $i - th$ local skew line that is accumulated to the orientation histogram with d_i being the minimal distance of p to the local skew line. A spline is fitted into the maximal bin and its two neighbors in order to find the dominant angle. In contrast to the gradient based method, the FNNC determines the dominant angle up to $[0\ \pi)$. Hence, the snippet can still be rotated upside down, but not $\frac{\pi}{2}$ to the text line orientation.

C. Method Combination

As previously discussed, each orientation estimation method is designed for different kinds of snippets. Hence, the FNNC performs better if handwritten text is supplied while the gradient method is more accurate than FNNC and regards background information.

Since both orientation histograms have the same number of bins, one could simply accumulate both histograms. However, the results improve if it is detect which method failed, and then assign a lower weight to that method. Therefore, a weighting scheme is established that incorporates knowledge about the orientation histogram's shape. A perfect orientation histogram is present if all information regarded is accumulated into one angle bin. On the other hand, the method failed if most or all angle bins have the same height. Hence, the weight assigned to each of the two orientation histogram is based on their integral:

$$\omega_o = 1 - \frac{\sum h(x)}{n \cdot \max h(x)} \tag{7}$$

where ω_o is the weight of the orientation histogram $h(x)$ and n is the number of bins. If $\min h(x) = \max h(x)$ or if less than 5 interest points were detected in an image ω_o is set to 0 for that histogram. Figure 3 shows the histogram combination. It can be seen, that the gradient method (c) considers the background ruling but has a higher peak resulting from the writing's slant. The FNNC (d) completely disregards the slant and the background information. It solely detects the correct orientation. Since the correct orientation was detected by both methods, the wrong peaks are suppressed (b).

D. Up/Down Orientation

The page up/down orientation determination is based on the work of Caprari [3]. Thus the decision is based on the frequency of ascenders and descenders of roman letters and Arabic numerals. Statistics of German and English text show that the occurrence of ascenders is dominating [13]. Caprari analyzes the asymmetry of the line histogram based on the ascender and descender frequency. Since the algorithm is sensitive to the correct skew, the entire page is divided into

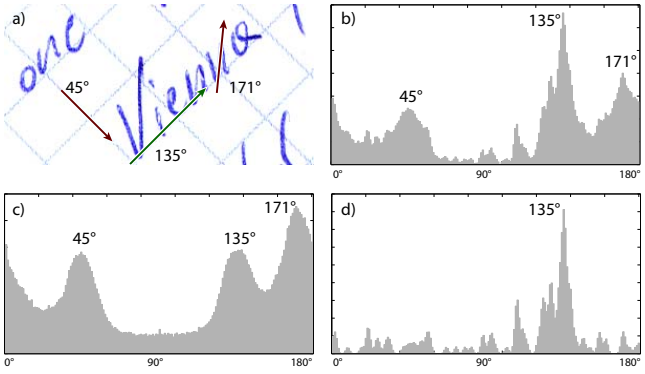

Figure 3. (a) Text region on checked paper (b) combined histogram (c) gradient histogram (d) FNNC histogram

stripes. It turned out that the best results are gained when a snippet is divided into 6 stripes.

IV. Evaluation

The proposed method has been evaluated on a synthetic test set and on 658 document fragments of the Stasi files. The following sections describe the test sets in more detail and present the results.

A. Results on Synthetic Images

The synthetic test set consists of 8 images, which has been reproduced from the test set presented in Ephstein [9]. First, Gaussian blur with $\sigma = 1.5$ and then Gaussian noise $\sigma = 0.05$ are added to all images. The images are rotated form 0 to π with a step of 0.05 radians as suggested in [9]. Figure 4 shows an example image of the test set. It can be seen that strong edges are not present in the synthetic test images and that the proposed method must be able to deal with noisy images even if solely two words are present. Table I shows the results of the proposed method compared to Ephstein's [9] and Bar-Yosef's et al. method [2]. The proposed method outperforms the skew estimation presented by Bar-Yosef et al. [2]. Table I additionally shows that the proposed method has no catastrophic errors (errors $> \pi/10$) and a variance $\sigma^2 < 1\,^\circ$. Although the synthetic test set has been reproduced from Ephsteins paper, the blur and Gaussian noise parameters are unknown, and thus leading to not directly comparable results.

Figure 4. Synthetic test image with Gaussian blur and noise added.

	median	mean	variance	catastrophic
proposed method	$0.35\,^\circ$	$0.64\,^\circ$	± 0.88	0
Bar-Yosef et al. [2]	$0.37\,^\circ$	$0.67\,^\circ$	± 2.05	2
Ephstein [9]	-	$0.497\,^\circ$	-	0

Table I
Synthetic images (504), see [9]

	median	mean	variance	catastrophic
proposed method	$0.56\,^\circ$	$1.75\,^\circ$	± 4.57	120
Bar-Yosef et al. [2]	$0.62\,^\circ$	$4.24\,^\circ$	± 9.10	106

Table II
Document fragments (658)

B. Results on Document Fragments

The second test set consists of 658 document fragments of Stasi files, which have been fragmented in 1989. The snippets have irregular shapes from a stamps size up to a full DIN A4 page (for a graph of the size distribution see [7]) and different paper types (checked, lined, void). The content ranges from two up to hundreds of words. The snippets have been manually tagged. In the case of handwritten text one global orientation is defined as ground-truth (direction with the most words aligned). Figure 5 shows a similarly

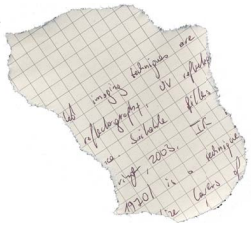

Figure 5. Exemplarily document fragment.

well preserved fragment as contained in the dataset, original fragments must not be shown due to privacy reasons.

Table II shows the results of the proposed method and Bar-Yosef's et al. method applied to 658 document fragments. Since Bar-Yosef's method needs a binarized image, all snippets have been thresholded using Otsu's method. Additionally a mask has been generated to ignore the border region of the fragments. Hence, the resulting binarization is improved.

The proposed method has a mean error of $1.75\,^\circ$ and 120 catastrophic errors ("*cases where the detected text orientation differs from the ground truth by more than* $\pi/10$" [9]), compared to a mean error of $4.24\,^\circ$ and 106 catastrophic errors. Catastrophic errors result from slanted text if solely a few words are present on a snippet. In these cases, the FNNC detects not enough interest points for a statistically significant orientation estimation.

V. Conclusion

A skew estimation for sparsely inscribed document fragments has been presented. The method is based on the gradient information and a FNNC of interest points. On the one hand the algorithm is suitable for document fragments with irregular shape and document pages inscribed with at least 2 words, on the other. Additionally there is no restriction on the detectable angle range. Due to the up/down orientation decision the method is limited to scripts where ascenders or descenders are dominating. The skew estimation has been evaluated on a synthetic test set (see [9]) and on 658 document fragments of the Stasi-files. It has been shown that the orientation can be reliably calculated without any restrictions on the detectable angle range.

Acknowledgment

The authors would like to thank the Fraunhofer-Institute for Production Systems and Design Technology (IPK), Berlin for supporting the work. We would also like to thank Itay Bar-Yosef, Nate Hagbi, Klara Kedem and Itshak Dinstein for providing the source code of the skew estimation presented in "Fast and Accurate Skew Estimation Based on Distance Transform" [2].

References

[1] A. Amin and S. Fischer. A Document Skew Detection Method Using the Hough Transform. *Pattern Analysis and Applications*, 3(3 2000):243–253, 2000.

[2] I. Bar-Yosef, N. Hagbi, K. Kedem, and I. Dinstein. Fast and Accurate Skew Estimation Based on Distance Transform. In *The Eighth IAPR International Workshop on Document Analysis Systems, 2008. DAS '08.*, pages 402 –407, sep. 2008.

[3] Robert S. Caprari. Algorithm for text page up/down orientation determination. *Pattern Recogn. Lett.*, 21(4):311–317, 2000.

[4] S. Chen and R.M. Haralick. An automatic algorithm for text skew estimation in document images using recursive morphological transforms. In *ICIP94*, pages 139–143, 1994.

[5] Su Chen, R.M. Haralick, and I.T. Phillips. Automatic text skew estimation in document images. In *Proceedings of the Third International Conference on Document Analysis and Recognition, 1995.*, volume 2, pages 1153 –1156 vol.2, aug. 1995.

[6] Erik D. Demaine and Martin L. Demaine. Jigsaw Puzzles, Edge Matching, and Polyomino Packing: Connections and Complexity. *Graphs and Combinatorics*, 23(1):195–208, 2007.

[7] Markus Diem, Florian Kleber, and Robert Sablatnig. Document Analysis Applied to Fragments: Feature Set for the Reconstruction of Torn Documents. In *DAS*, pages 393–400, 2010.

[8] A. Egozi and I. Dinstein. An EM Based Algorithm for Skew Detection. In *Ninth International Conference on Document Analysis and Recognition, 2007. ICDAR 2007.*, volume 1, pages 277 –281, sep. 2007.

[9] Boris Epshtein. Determining document skew using interline spaces. In *11th International Conference on Document Analysis and Recognition, 2011. ICDAR '11.*, pages 27–31, 2011.

[10] Mohamed E. Hussein, Fatih Porikli, and Larry S. Davis. Kernel integral images: A framework for fast non-uniform filtering. In *International Conference on Computer Vision and Pattern Recognition, CVPR*, pages 1–8, 2008.

[11] Xiaoyi Jiang, H. Bunke, and D. Widmer-Kljajo. Skew detection of document images by focused nearest-neighbor clustering. In *Proceedings of the Fifth International Conference on Document Analysis and Recognition, 1999. ICDAR '99.*, pages 629 –632, sep. 1999.

[12] Junichi Kanai and Andrew D. Bagdanov. Projection profile based skew estimation algorithm for JBIG compressed images. *IJDAR*, 1(1):43–51, 1998.

[13] Robert Edward Lewand. *Cryptological Mathematics*. The Mathematical Association of America, 2005.

[14] Rafael Dueire Lins and Bruno Tenrio vila. A New Algorithm for Skew Detection in Images of Documents. In Aurlio C. Campilho and Mohamed S. Kamel, editors, *ICIAR (2)*, volume 3212 of *Lecture Notes in Computer Science*, pages 234–240. Springer, 2004.

[15] David G. Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[16] K. Omar, A. Ramli, R. Mahmod, and M. Sulaiman. Skew Detection and Correction of Jawi Images using Gradient Direction. *J. Techn.*, 37:117–126, 2002.

[17] Jan Schneider and Bertram Nickolay. Automatische virtuelle Rekonstruktion vernichteter Dokumente. *Fraunhofer FUTUR*, 2:6–7, 2006.

[18] Jan Schneider and Bertram Nickolay. The Stasi puzzle. *Fraunhofer Magazine, Special Issue*, 1:32–33, 2008.

[19] Changming Sun and Deyi Si. Skew and Slant Correction for Document Images Using Gradient Direction. In *ICDAR '97: Proceedings of the 4th International Conference on Document Analysis and Recognition*, pages 142–146, Washington, DC, USA, 1997. IEEE Computer Society.