Annals of DAAAM for 2010 & Proceedings of the 21st International DAAAM Symposium, Volume 21, No. 1, ISSN 1726-9679
ISBN 978-3-901509-73-5, Editor B. Katalinic, Published by DAAAM International, Vienna, Austria, EU, 2010
Make Harmony Between Technology and Nature, and Your Mind will Fly Free as a Bird
Annals of DAAAM International

CONFIDENCE MEASURE FOR SKEW DETECTION IN PHOTOGRAPHED DOCUMENTS

BOIANGIU, C[ostin] - A[nton]; ROSNER, D[aniel]; OLTEANU, A[lexandra]; STEFANESCU, A[lexandru] V[ictor] & MOLDOVEANU, A[lin] D[ragos] B[ogdan]

Abstract: *Skew detection and correction poses particular challenges in deformed images, as well as images suffering from camera lens distortions or distortions caused by non-flat scanning surfaces. The current paper introduces an algorithm that yields good results on such images and proposes a confidence measurement to estimate the accuracy of the returned skew angle.*
Key words: *skew, neighbor clustering, confidence measurement*

1. INTRODUCTION

In recent years, more and more institutions, companies and libraries have made the transition from plain paper documents and books towards digital content. As many of them retain large paper based archives that need to be integrated in their new digital systems, faster and more robust automated content conversion systems are needed to speed up and cut down on costs for converting books and archives to digital form.

Given the size of such archives, which in many cases can be in the range of millions of pages, improvements in the automatic digitalization system can lead to savings in terms of time, man power and mainframe power consumption.

The first stage in an automated content conversion and one that can induce many errors is the scanning procedure. A typical fault, and the one this paper will address, is the scanning or photographing of the input image at a skew angle. There are many scenarios for this error to occur, but generally there are caused either by poor alignment on the scanning surface, by human error or by imperfections in the optic system of the scanning device. Skew errors can also be expected when scanning thick books that cannot be properly placed on the scanner surface, or place leveled in front of a camera.

The current article presents an approach for detecting the skew angle of an image, with a new method for estimating a confidence level for the detected skew angle in the presence camera lens distortions or similar faults.

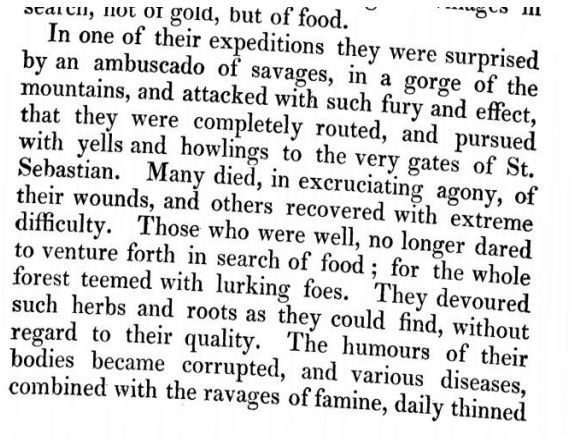
2. RELATED WORK

Skew detection and correction for scanned or photographed images is a field of great interest and thus many solutions have been proposed over the years.

There are four main classes of algorithms (Baird, H. S. 1995): Hough transform, projection profiling, cross-correlation and neighbor clustering.

Hough transform (Chandan Singh et al., 2008) and projection profiling based algorithms are constructed on the presumption that parallel lines of text will generate collinear agglomerations of foreground pixels, but use different mathematic representations of the idea. Cross correlation based algorithms use the general coherency between stripes from lines of text that can be expected in a skew free document.

Neighbor clustering techniques use clustering methods to rebuild lines of text. Then, using specific feature-points selected from inside a detected line of text, the skew angle is estimated by various line fitting techniques (R. Smith., 1995), (Lu, Y et al., 2003).


search, not of gold, but of food.

In one of their expeditions they were surprised by an ambuscado of savages, in a gorge of the mountains, and attacked with such fury and effect, that they were completely routed, and pursued with yells and howlings to the very gates of St. Sebastian. Many died, in excruciating agony, of their wounds, and others recovered with extreme difficulty. Those who were well, no longer dared to venture forth in search of food; for the whole forest teemed with lurking foes. They devoured such herbs and roots as they could find, without regard to their quality. The humours of their bodies became corrupted, and various diseases, combined with the ravages of famine, daily thinned

Fig. 1. Example input image

All this techniques will generally yield excellent results, but will lose precision in the presence of camera lens distortions. Moreover, these techniques cannot offer a measure of the accuracy of the returned skew angle.

3. ALGORITHM

The present algorithm is based on the observation that good skew features can be extracted from the borders of image entities. For most characters, especially Latin ones, bottom borderlines can provide excellent measurement of skew angles. The present skew detection algorithm identifies lines of text and extracts skew angle information based on these lines.

In order to overcome errors introduced by camera lens distortions in images, the current paper introduces a modified skew angle estimation method and a new approximation of the accuracy of results.

3.1 Preprocessing

If the image is not bitonal, the image will first be binarized, using an existing algorithm, like (J. Sauvola, 2000).Entities are then discovered and contour information is preserved.

3.2 Noise removal and entities filter

The main filter in the present work is a minimal size threshold used to remove noise. Considering the 6 point letter ‘e’ as a minimum size reference, it can be shown that most Latin characters are not readable below this size. Thus, entities below 6 pixels can be viewed as noise.

Considering that the human eye cannot distinguish, unaided, details smaller than 0.1 mm, entities below 0.6 mm of height can be removed from the computation. Thus, the minimum accepted height, measured in pixels, will be (considering that an inch has 25.4 mm):

$$noise\_threshold = round\left(max\left(6, \frac{0.6 * Image\_DPI}{25.4}\right)\right). \quad (1)$$

A second heuristic threshold is used for large entities removal. If too much of the input entities are removed, the second threshold can be reduced in order to retain more input data. Further filtering is done within the clustering algorithm.

3.3 Cluster construction

Clusters are initially constructed based on a vicinity clustering algorithm that is optimized for horizontal Latin based text.

A character's vicinity is defined as rectangular area: of 120% height, and of a length of 200% the character's height, starting from its right bottom pixel. A maximum height difference of 200% is also used. This character clustering algorithm yields good results in the presence of skew angle in the range ±15° and normal camera lens distortions.

3.4 Cluster selection

It is important that the clusters offering better skew information are retained, while the clusters that can introduce errors are discarded.

Clusters of small length will generally not retain enough input information for an accurate skew detection. However, depending on the page layout, there are pictures that do not contain sufficiently long lines. Thus, cluster filtering will be tried in an iterative fashion, from highest precision, until a sufficient number of clusters are found:

$$cluster\_threshold = 1/\tan(i \times angular\_precison) \quad (2)$$

where i is the iteration number and $angular_precision$is the desired skew angle detection accuracy.

3.5 Skew angle estimation

For each cluster, two approximate skew angle computations are made by means of a least square fitting method performed on the bottom pixels of each entity in the cluster. The first is used to find the direction of the skew fault, and the second is used to find the angle, while removing descenders in the computation.

Next, the input points are rotated towards the OX axes with the determined skew angle.

Finally, the data points are fitted to a parabolic function in order to determine the straightness of the cluster. There are two indicators of the quality of a cluster for skew detection: cluster length and cluster "flatness" – determined by the distance between the focal point and the extreme of the parabolic function.

Thus, considering that the cluster length in pixels is *cluster_length* and that the fitting function was determined as:

$$a\,x^2 + b\,x + c \quad (3)$$

a confidence degree can be defined as:

$$confidence\_degree = cluster\_length \times a\,. \quad (4)$$

3.6 Final skew angle computation

A histogram is used to eliminate values that are too far off the main peaks. The final skew angle computation takes into account the confidence of each cluster, and is thus computed as a weighted mean between the skew angle and the confidence degree for each cluster.

$$final\_angle = \frac{\sum(angle\ [i] * confidence\ \_degree[i])}{\sum confidence\ \_degree\ [i]} \quad (5)$$

3.7 Cluster recomputation

After the final skew angle was computed, if the input data quality is considered of low quality, a new computation can be made, by recomputing clusters using a modified version of the vicinity clustering that account for the direction and the approximate angle of the skew fault.

Finally, steps *3.4* to *3.7* are reiterated and the final skew angle is returned.

4. DEGREE OF CONFIDENCE OF FINAL RESULTS ESTIMATION

As the next steps in an automated digitalization process depend on the results obtained by the skew detection algorithm, a degree of confidence is introduced as a second returned value for the presented algorithm.

The expected error is inverse proportional to the length of the clusters and with the straightness of the detected parabolas.

Thus, the algorithm will return two values:

- the skew angle, in degrees
- a confidence measure

The confidence measure is given by the formula:

$$\sum confidence\_degree[i] \,/\, number\_of\_clusters \quad (6)$$

In the present form, the confidence measure is not normalized, but represents a way of evaluating the quality of the returned value by comparing the confidence measure returned for various image.

Moreover, in systems that use feedback loops, an image can be processed several times, each time with different pre-processing filters, and the confidence measure can be used to determine which filters yield the best results.

5. FURTHER WORK

For now, the algorithm achieves great accuracy and provides a measure of the degree of accuracy for the returned skew angle, depending on the quality of the input data.

Further work will be done in order to obtain a normalized degree of confidence that will likewise be returned together with the skew angle, but providing a better measure of the skew detection success. Two values shall be returned.

First, the algorithm will return the angle accuracy, measured in degrees, which will represent the maximum possible error in the returned angle, and thus an angle space in which the true skew angle will surely reside.

The second value will be a percentage indicating the probability that the returned skew angle is the actual true skew angle.

6. ACKNOWLEDGMENTS

The research presented in this paper is supported by the national project "Excelenţă în cercetare prin programe post doctorale în domenii prioritare ale societăţii bazate pe cunoaştere (EXCEL)", Project POSDRU/89/1.5/S/62557.

7. REFERENCES

Baird, H. S. (1995). The skew angle of printed documents. *Document Image Analysis*, Eds. IEEE Computer Society Press, Los Alamitos, CA, 204-208

Bhatia, C.S.N. & Kaur, A. (2008). Hough transform based fast skew detection and accurate skew correction methods. *Pattern Recognition*, Volume 41, Issue 12, pp. 3528-3546

Sauvola, J. & Pietikainen, M. (2000).Adaptive document image binarization. *Pattern Recognition*, 33(2):225–236

Lu, Y. & Tan, C. L. (2003). Improved Nearest Neighbor Based Approach to Accurate Document Skew Estimation. *Proceedings of the Seventh international Conference on Document Analysis and Recognition* - Volume I. CDAR. IEEE Computer Society, Washington, DC, pp. 503

Smith, R. (1995). A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation. *Proc. of the 3th International Conference on Document Analysis and Recognition*, Montreal, Canada, pp. 1145 - 1148