Proceedings of the 4th International Conference on Document Analysis and Recognition (ICDAR'97), Ulm, Germany, August 1997, IEEE Computer Society Press, pp. 684–688.

Local Skew Angle Estimation from Background Space in Text Regions

Apostolos Antonacopoulos

Department of Computer Science, University of Liverpool,
Peach Street, Liverpool, L69 7ZF, United Kingdom (aa@csc.liv.ac.uk)

Abstract

*Almost all document analysis approaches need to perform a **global** analysis of the page orientation as a separate process at an early stage. It would be preferable to estimate the orientation locally after page segmentation and classification, when more knowledge about the different regions is available. In this paper, a novel local skew estimation method is presented that takes advantage of the information available after flexible and efficient page segmentation and classification methods have been applied to the document image. The proposed method accurately estimates the orientation of **individual** text regions by efficiently analysing the arrangement of background space contained in them. No assumption is made of the existence of a uniform or dominant orientation in the document. The whole process is very efficient, as only the regions of text are considered and the points used for the angle estimation are already available as by-products of previous document analysis stages.*

1. Introduction

The importance of the need for the estimation of skew in document images has been widely agreed upon in the Document Image Analysis literature. However, in the vast majority of cases skew detection is performed during a pre-processing step and is followed by an image rotation operation to restore the page image to its 'correct' orientation. The main problems associated with such an approach are as follows. Firstly, detection and correction of the skew in the whole of the image is computationally expensive. Secondly, a corrective rotation often leads to a degradation of image quality due to quantisation errors. Finally, there might not be a unique 'correct' orientation for the entire image, as separate regions may have been intentionally printed in different orientations.

There is a need for local skew detection, and for only a selected set of regions. For instance, in the majority of situations it is pointless to attempt to estimate the skew in non-textual regions. Ideally, a skew detection method should exploit the knowledge gathered about the printed regions during page segmentation and classification. If information produced during these document analysis stages is not used, skew detection becomes wasteful and remains computationally expensive. Therefore, if possible, skew should be estimated after page segmentation and classification. The orientation angle can then be passed as a parameter to subsequent recognition and understanding processes (it is assumed that it may not be desirable to change intentional skew). In reality, however, most document analysis approaches are not robust enough to proceed successfully in the presence of skew and measure it in the segmented and classified regions. An exception is [1]. Most approaches require a mandatory skew correction pre-processing step, while others require an intermediate estimate of skew (e.g. [2] [3]).

As far as the process of estimating the skew angle is concerned, several methods have been proposed. One approach is based on the analysis of projections of pixels or other representative points in one [4] or in a series of angles [5][6]. Another approach involves the analysis of the histogram of angles between nearest neighbour pairs [3][7]. Finally, a popular approach is to attempt to identify straight lines, each describing a set of representative points, using the Hough transform [8][9] or some other method of line fitting (e.g. [2][10][11]). These methods are applied on a set of points that have been extracted from the image as representative of the position of characters or parts of text lines. For a more detailed overview of some of the methods mentioned above the reader is referred to O'Gorman and Kasturi [12].

In general, methods that require computations on the pixel-based image data (e.g. [5]) are time-consuming. Extensive calculations or comparisons are another disadvantage of many methods [3][5][6][7][8]. The extraction of connected components [3][6][7][11] is also a time-consuming step, especially when these components are not required for any further document understanding purpose. A further area of concern is the handling of complex-shaped regions. Whilst efficient otherwise, the method of Pavlidis and Zhou [2] is not applicable to non-rectangular regions. Finally, as far as local skew estimation is concerned, the only previous approach that is capable of identifying skew independently for each

region is that of O'Gorman [3], whilst the approach of Yu *et al.* [10] must first identify and correct the global skew before page segmentation and then re-calculate local skew.

In this paper, a novel local skew estimation approach, based on the description of the background space inside printed regions, is presented. It exploits the fact that gaps in text regions are aligned in the reading direction. As the description of these gaps is available as a by-product of page segmentation [1], no effort is required to identify the set of candidate points to be used as input to the estimation method. The orientation of each printed region is identified as the dominant slope of lines fitted to collinear centroids of selected gaps. The selection of appropriate collinear points is fast, in contrast to the identification of nearest-neighbour pairs. As the page segmentation and classification [13] steps are both efficient and flexible, even when complex-shaped printed regions are present, accurate estimation of the skew of individual regions can be safely performed.

In the following section, a brief description is given of the rationale and of the methods for page segmentation and classification using the background space. This description provides the context for the proposed skew determination method and explains the derivation of the description of the background space. In Section 3, the main stages of the new method are detailed. Finally, in Section 4, the performance of the method and related issues are discussed and some representative results are presented.

2. The White Tiles Approach

The White Tiles document analysis approach is based on the analysis of the structure of the background space (usually white) in a page image. The White Tiles *page segmentation* method [1] exploits the fact that the printed regions on a document page are surrounded by background space. The arrangement of this space can be thought of as an irregular (because of the different shapes of the regions) net. The idea is that by reconstructing this net of white space one can identify and describe the holes i.e. the printed regions. The White Tiles *page classification* method [13] is based on the textural analysis of the background space inside the regions identified by page segmentation. A considerable advantage of this method is that it derives classification features from the description of background space (see below) already obtained during page segmentation. The whole White Tiles approach does not assume anything about the shapes of the regions or their orientation and is, therefore, quite flexible and efficient in the identification and description of complex-shaped and skewed regions.

A flexible way to describe the background space is by *white tiles* [1]. Each tile represents the widest area of white space that can be accurately represented by a rectangle. Hence, the net surrounding the printed regions is represented as a set of connected white tiles of different sizes. Meanwhile, the space inside regions is also represented by white tiles. The white tiles are identified in the page image after a vertical smearing process which reduces the amount of superfluous space. The smearing value is identified beforehand, in proportion to the dominant distance between baseline peaks in the projection profiles of a small number of vertical narrow strips in the image. It must be noted that, after the identification of white tiles, no other image access occurs. All subsequent document analysis stages perform operations on the white tile representation only. This fact is a significant efficiency-related advantage of the white tiles approach.

During page segmentation, appropriate edges of the region-surrounding white tiles are selected to create the contours that describe the printed regions. The method for obtaining the white tiles and for identifying the regions of interest is described in [1]. Furthermore, the textural characteristics of areas of different types (text, graphics and line-art) can also be expressed in terms of white tile information. Those tiles that are not used during the segmentation, i.e. the white tiles inside the identified regions, are used to derive the necessary features for the classification of the segmented regions [13].

The white tiles that describe the background space inside text regions can be further used to estimate the individual orientation of such regions. Since the characters inside a text region are aligned in parallel text lines, the space that separates characters inside and between text lines is also aligned in parallel lines of the same orientation. The proposed skew estimation method is based on this fact and its function is explained in the next section.

3. Skew Angle Estimation

The proposed method for orientation estimation is based on the following principles. Firstly, the estimation only takes place for text regions. The white tiles inside text regions describe some instances of space between characters inside words, space between words in the same textline and some fragments of space between textlines. It is observed that as the characters are printed along parallel straight lines, the separating space also occurs in a similar pattern. As for non-text regions, it should be noted that, unless there is prior knowledge of their structure which can be used in the estimation of their orientation, such an estimation process should be abandoned.

Secondly, it is assumed that text regions in the document are printed in such orientations that are easily readable by a human without rotating the paper. This means that, without any skew, the orientation differences between text regions should not be excessive. It is observed that, with the exception of mixed horizontally/vertically written documents (e.g. Japanese), text regions are not in general intentionally rotated more than 45° off the main reading direction of the text. Although the proposed method is equally applicable to documents where text is read from top to bottom, in this paper the description of the method is restricted to documents where text is expected to be read in a loosely horizontal direction. However, there is no restriction in individual text regions having different orientations. In addition, contrary to the overwhelming majority of skew detection methods (all except [3]), there is no assumption that the document should have a dominant orientation. An example image containing a complex-shaped text region as well as an intentionally rotated one is shown in Figure 1, while the segmented and classified text regions are shown (as contours) in Figure 2 together with the white tiles contained in them.

In each of the regions considered, the problem becomes to identify the text line orientation by estimating the orientation of rows of white tiles representing background gaps. The method proceeds from having the set of the coordinates of all white tiles inside a text region to identify suitable subsets of points to which straight lines can be fitted. The orientation of each of the fitted lines is then examined and the average of the most frequently occurring angles is taken as the orientation of the region. The following subsections describe these steps in more detail.

3.1 Identification of points

The starting set of candidate points, from which suitable collinear subsets are chosen, is obtained in this step. It is a very simple operation that comprises storing the x and y coordinates of the centroids of white tiles. However, there are some white tiles whose height is greater than the majority of other tiles. These tall tiles describe instances of background space extending for more than the height of a textline. As the centroids of such tiles would distort the overall alignment of points, they are adjusted (the dominant height of gaps between textlines is known from page segmentation).

3.2 Identification of strings

In the set of all points from a text region, it is observed that most of the points are organised along parallel strings. A *string* is referred to here as a maximal subset of collinear neighbouring points. The aim of this step, which is the most crucial in the whole approach, is to identify such strings of points in an efficient way.

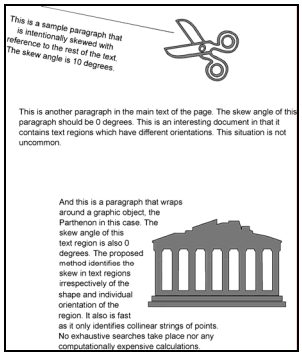

Figure 1. An example image.

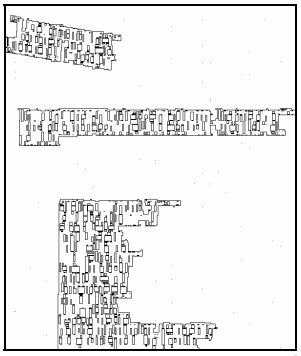

Figure 2. The text regions of Figure 1 with the white tiles contained in them.

The method proposed in this paper exploits the following information, which is available from the previous document analysis stages. Firstly, it is known whether the orientation of each text region is closer to the horizontal or to the vertical. Secondly, the dominant baseline distance in the document is known.

The idea at this stage is that if the points plane is divided into intervals spanning the reading direction (horizontal here), parts of the required strings will be contained in them. These substrings are more straightforward to identify because consecutive points in each interval are much more likely to belong to the same substring. In contrast, in the undivided set of points, the possible connections are very much larger in number and, therefore, it is more computationally expensive to identify the required ones. In the approach described in this paper, the height of the interval is chosen according to the dominant baseline separation.

The method proceeds as follows. In the beginning, the points are sorted according to their y-coordinate. Consecutive subsets of points that fall in different intervals are then sorted according to their x-coordinate. An example of the division of the points plane into

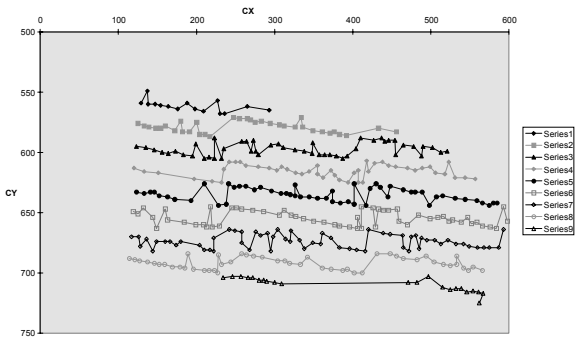

Figure 3. The points divided into intervals (each series comprises the points in an interval).

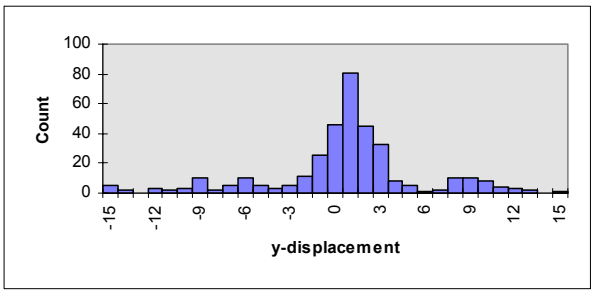

Figure 4. Example histogram of y-displacements.

intervals is shown in Figure 3.

The next task is to identify substrings in each interval. The main concern in doing so is to avoid outliers, while allowing minor variations in slope between consecutive points. As a measure of variation between pairs of points, the y-displacement is used. This has proved adequate in experiments due to the fact that the white tiles are arranged in more or less regular distances. In the histogram of y-displacements between consecutive points in every interval, there is a clearly dominant peak. An example of such a histogram, for the points in the region shown in Figure 3, can be seen in Figure 4.

The histogram of the y-displacements is computed and the left and the right cut-off values are chosen around the peak. The identification of these cut-off values takes into account the ratio R_y of the value of each candidate (adjacent to the peak) bin to the value of the peak. When a ratio of less than R_y is reached, the candidate bin is chosen as one of the two cut-off values (left or right). The value of $R_y = 0.5$ was shown to be sufficient in the experiments.

The points in each interval are considered one at a time and, if their y-displacement is acceptable, they are appended to the substring containing the previous point. If there is no substring (at the beginning of each interval) or the y-displacement of the current point falls outside the permitted values, a new substring is created. When no more points can be allocated to the current substring, its length is checked and if it is above a certain threshold T_l the substring is kept. All other substrings shorter than this threshold are marked to be discarded as experimental evidence has shown that orientation angles can not be reliably calculated from them. In the experiments the value used for T_l is 5 (pixel units).

Smaller substrings, which otherwise would be discarded, can be eligible to be joined with other larger substrings in the same interval, if they were separated by an abrupt change that immediately cancels out. This situation will occur if there is a single outlier encountered whilst identifying the substring (see Figure 3). If after excluding the outlier the second substring forms a continuation of the first, the two strings are joined. An additional criterion is that the x-displacement between the end-points to be joined should also be acceptable. The threshold for this displacement is calculated from the histogram of x-displacements between consecutive points in every interval of the region in question. Substrings from adjoining intervals can also be joined in the same manner to produce longer strings. By joining smaller substrings, greater overall accuracy is obtained, as more substrings will contribute to the estimation of the orientation of the region.

An illustration of the selected substrings identified among points in some intervals of the example of Figure 3 can be seen in Figure 5 indicated with a darker tone.

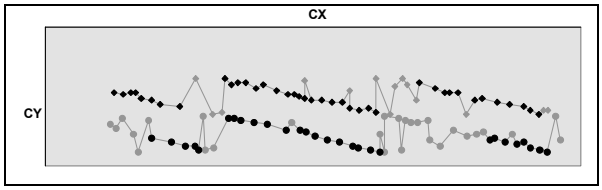

Figure 5. Example of identified substrings (darker points).

3.3 Estimation of orientation angle

At this final step, the orientation angle of the region is estimated from the identified strings of points. The slope of the straight line, fitted using the least squares method, is recorded for each string. A histogram of the identified angles is then computed, and the dominant neighbourhood is identified. The required angle of orientation of the region is computed as the average of the angles in that dominant neighbourhood.

4. Results and Discussion

In preliminary tests (using a set of documents with traditional and complex layouts) with individual text regions being skewed (by software) at angles ranging from -10° to 10°, the method has shown very high accuracy. More precisely, the errors observed range from 0° to 0.1°, with error values at the lower end also observed for the larger angles. The regions were exactly as output from the page segmentation and classification stages of the White Tiles approach. As the White Tiles page segmentation and classification methods can successfully proceed with skew angles larger than ±10° (at least up to ±15°) work is being carried out to adapt the method to accurately identify such large skew angles. However, it should be pointed out that it has been observed that in practice skew angles greater than ±10° are relatively uncommon.

Ideally, document image analysis and understanding approaches should work well irrespective of skew. Unfortunately, such robust methods are not available for all the stages involved (e.g. OCR). Therefore, the skew

has to be estimated and either passed on as a parameter to subsequent stages or corrected by a rotation in the opposite direction. Almost all document analysis approaches perform an analysis of the skew at the early stages. However, it is preferable to perform skew estimation after page segmentation and classification, when more knowledge about the different regions is available. In this paper, a novel skew estimation method has been presented that takes advantage of the information available after flexible and efficient page segmentation and classification methods have been applied to the document image.

The proposed method estimates the orientation of individual text regions by analysing the arrangement of background space contained in them. The whole process is very efficient for the following reasons. Firstly, only the appropriate regions are analysed in contrast to the other methods which are applied to the whole of the image. Secondly, the points used for the angle estimation are already available as by-products of previous document analysis stages. Finally, no extensive calculations or comparisons are performed. For instance, the division of points in intervals makes the identification of strings of collinear points faster as there is no need to consider a large number points as in the identification of k-nearest neighbours.

The nature of the background space in text regions gives rise to some issues which are particular to methods that analyse the structure of this space. It can be observed (see Figure 2) that white tiles are not arranged along straight lines with the same regularity as characters are. Therefore, a histogram of angles between nearest neighbour pairs [3] will not give as sharp a peak as it perhaps would for angles between characters. However, the results of the proposed method show that not only is it feasible to estimate the orientation from white tiles but it is very accurate too.

Another issue is the lack of enough background space in a region to warrant accurate skew estimation. This can be the situation where a paragraph consists of only one or two textlines, in which case the result will not be as reliable. In this case, as the number of characters is very small, an alternative solution would be to estimate the skew using one of the existing character-based approaches. In a different situation, there might be the case where a small isolated word fragment contains enough space to be correctly classified but the tiles describing this space may be so few that the orientation of the line passing through their centroids is not the same as the actual one. Such a case, however, arises relatively rarely in the White Tiles approach and, it should be noted, that the issue of reduced accuracy in situations where so little information is present, is common to all approaches.

Furthermore, one of the most significant advantages of the proposed method is its flexibility. The orientation of each text region is estimated *independently* of the other regions without any assumption about the existence of a dominant orientation in the document page. Such flexibility is a key characteristic of the whole White Tiles approach as the preceding document analysis stages successfully segment and classify printed regions (possibly complex-shaped) irrespective of skew. Finally, it should be pointed out that if it is known that there are no individually skewed regions in a document, the global skew can be very quickly estimated by applying the proposed approach to one text region only.

References

[1] A. Antonacopoulos and R.T. Ritchings, "Flexible Page Segmentation Using the Background", *Proc. 12th ICPR*, Vol. II, Jerusalem, Israel, Oct. 1994, pp. 339–344.

[2] T. Pavlidis and J. Zhou, "Page Segmentation and Classification", *CVGIP: Graphical Models and Image Processing*, **54**, no. 6, November 1992, pp. 484-496.

[3] L. O'Gorman, "The Document Spectrum for Page Layout Analysis", *IEEE Trans. on PAMI*, **15**, no. 11, November 1993, pp. 1162–1173.

[4] T. Akiyama and N. Hagita, "Automated Entry System for Printed Documents", *Pattern Recognition,* **23**, no. 11, 1990, pp. 1141–1154.

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

[6] H.S. Baird, "The Skew Angle of Printed Documents", *Proc. SPSE 40th Conf. & Symp. on Hybrid Imaging Systems,* Rochester, N.Y., May 1987, pp. 21–24.

[7] A. Hashizume, P.-S. Yeh and A. Rosenfeld, "A Method of Detecting the Orientation of Aligned Components", *Pattern Recognition Letters,* **4**, April 1986, pp. 125–132.

[8] S.C. Hinds, J.L. Fisher and D.P. D'Amato, "A Document Skew Detection Method Using Run-Length Encoding and the Hough Transform", *Proc. 10th ICPR*, **1**, 16–21 June 1990, Atlantic City, NJ, U.S.A., pp. 464–468.

[9] Y. Nakano, Y. Shima, H. Fujisawa, J. Higashino and M. Fujinawa, "An Algorithm for the Skew Normalization of Document Image", *Proc. 10th ICPR*, **2**, 16–21 June 1990, Atlantic City, NJ, U.S.A., pp. 8–13.

[10] C.L. Yu, Y.Y. Tang and C.Y. Suen, "Document Skew Detection Based on the Fractal and Least Squares Method", *Proc. 3rd Int. Conf. on Doc. Anal. and Rec. (ICDAR'95)*, **2**, Montréal, Canada, August 14–16, 1995, pp. 1149–1152.

[11] R. Smith, "A Simple but Efficient Skew Detection Algorithm via Text Row Accumulation", *Proc. 3rd Int. Conf. on Doc. Anal. and Rec. (ICDAR'95)*, **2**, Montréal, Canada, August 14–16, 1995, pp. 1145–1148.

[12] L. O'Gorman and R. Kasturi, *Document Image Analysis*, IEEE Computer Society Press, 1995.

[13] A. Antonacopoulos and R.T. Ritchings, "Representation and Classification of Complex-Shaped printed Regions Using White Tiles", *Proc. 3rd Int. Conf. on Doc. Anal. and Rec. (ICDAR'95)*, **2**, Montréal, Canada, August 14–16, 1995, pp. 1132–1135.