Detecting a Document's Skew: A Simple Stochastic Approach

A. Hennig, G. Raza, N. Sherkat, R. J. Whitrow
Department of Computing, The Nottingham Trent University,
Burton Street, Nottingham NG1 4BU, UK
{amr,ghr,ns,rjw}@doc.ntu.ac.uk

Abstract

A simple stochastic approach to the detection of the skew angle of a scanned document is proposed in this paper. The method first estimates the undirected skew within the range of -90° to +90°. I-dots and full-stops are then used to determine whether the document has been scanned upside down, yielding the directed skew. Only a fraction of the information available in the document is exploited, without assumptions as to the general layout of the page. The method has been shown to be robust and accurate for a variety of documents containing both hand-written and printed text. The average error of the directed skew angle was observed to be 0.05° for a set of synthetic documents. For facsimile images, however, the upside-down detection failed in 9.6% of the documents examined.

1 Introduction

The detection of the overall skew of an electronic document is a vital early step in its analysis and recognition. In the majority of documents, lines of text are oriented horizontally. In a rotated document, skew angle is defined as the difference between the dominant orientation of the text lines and the horizontal. When a document is scanned, skew might be introduced in several ways. A relatively small skew might be added by imprecise feeding of the paper into the scanner or fax machine. A document in landscape orientation might have to be scanned in portrait orientation if the size of the scanner is insufficient. The resulting image is then rotated by +90° or -90°, depending on the preferences of the person that feeds the document. The document might even be inserted upside down, resulting in an additional skew of 180°. If a paper document is photocopied before the scanning, these effects might accumulate further.

In this paper, the angle of the '*directed skew*' (denoted by $\varphi, \varphi \in (-180°, +180°]$) describes the total rotation that has been applied to the original (right way up) document. The '*undirected skew*' (denoted with $\overline{\varphi}, \overline{\varphi} \in (-90°, +90°]$), however, assumes that the document has been scanned with a normal orientation. Most documents have a uniform skew. If they are moved during scanning or copying, or if the original contains curved or non-parallel lines of text, the skew angle is not constant throughout the document. This problem has been addressed in work described in [6][3] and is beyond the scope of this paper.

Various methods have been applied to the detection of a uniform skew. Projection profiles of connected components have been used by [2]. The Hough transform has been widely used, e.g. by [4], [1] and [9], even though it is computationally expensive and usually requires an assumption about the interval that contains the skew. A multi-layer perceptron is used for cursive handwriting in [5], whereas [8] applies a least square linear regression to reference points found on the page. Most of these methods, however, assume a right-way-up document, or even require that lines of text can be identified correctly.

The method presented in this paper uses a simple stochastic method to estimate the overall undirected skew of the document. A second step aims to detect whether the document has been scanned upside down, thus obtaining the directed skew. These steps are described in the following sections, followed by experimental results and concluding remarks.

2 Detection of the undirected skew angle

The undirected skew detection algorithm is based on the fact that elements of one straight line of text correlate with each other. The directions between elements of the same line of text distribute around the skew, whereas the connection to other lines can appear in almost any other direction. The histogram obtained from the angles of all possible connections hence shows a maximum at the desired skew angle. Elements might be black pixels or the centres of regions of connected black pixels. The example in Fig. 1a shows the connections from the centre of the region that forms the letter 'n' to the other regions of the page. Fig. 1b shows the resulting histogram with a clear maximum at the document's skew.

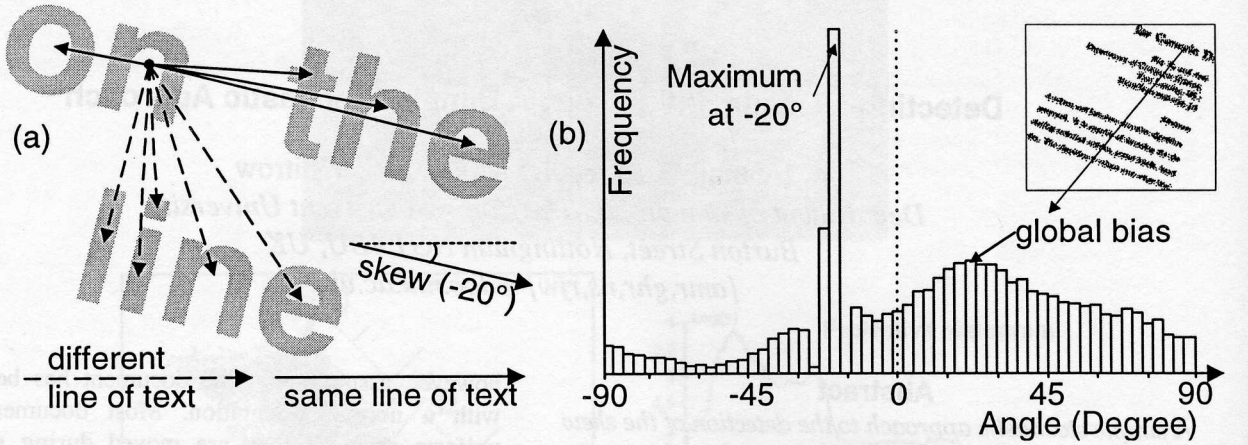

Fig. 1: Connections between black regions: a) Regions belonging to the same line of text appear under the same angle, the regions of other lines can appear at any other angle. b) The histogram shows a clear maximum around the undirected skew of the document.

This method, however, requires an excessive computational effort if all possible connections between regions or even pixels are considered. Furthermore, the histogram is biased by the overall arrangement of the elements. Elements in the corners of the document, for example, 'see' the majority of the remaining ones in direction of the opposite corner. This effect can be observed in the secondary maximum around 25°, derived from the document shown in the inset of Fig. 1b. The extraction of regions of connected pixels can be computationally expensive. If the document contains many lines such as tables, forms or text written on lined paper, only a few large regions can be detected. The centre point then becomes a poor representation of the region and the method becomes inappropriate.

To overcome these problems, the proposed method uses pixels instead of regions. The black pixels Q in a chosen distance r from a selected pixel P are considered (Fig. 2a). The angles the points Q appear at are computed and the histogram is updated accordingly. If this step is repeated for a number of different pixels P, the resulting histogram shows a clear global maximum (Fig. 2b) at the skew angle. Pixels P are selected randomly in order to avoid the exhaustive observation of all elements in the document without favouring a particular area. This allows the computational cost to be reduced dramatically without a substantial loss in accuracy. It also avoids the need of assumptions about the general layout of the document, e.g. the assumption that the lowest line on the page is a stable feature in [1].

Secondary maxima in the histogram are caused by the text lines above and below the line containing P, particularly if the lines are equally spaced. With increasing radii, these secondary maxima move closer to the global one. In order to emphasise the global maximum and to reduce the height of the secondary maxima, histograms of different radii are superimposed (Fig. 2c). The choice of the allowed radii is limited by the size of the letters in the document and the total size of the document. If r becomes too small, both P and Q might belong to the same letter (to or immediately adjacent ones) and no clear maximum will be observed in the histogram. If to large a radius is chosen, only pairs of that distance can be used, and observation is effectively reduced to the corners and margins of the document.

In order to simplify the computation further, the city-block distance is used instead of the Euclidean one. The shapes described by the points Q then degenerate from a circle around P to a square rotated through 45°. The cells of the histogram are denoted by $h_i \; ; i = 0(1)H-1$; H being the size of the histogram.

The complete algorithm for the detection of the skew angle is:

1 initialise $h_i=0$; $i=0(1)H-1$
2 for n randomly chosen black pixels P
3 choose a radius $r \in [r_{min}, r_{max}]$ randomly
4 for every pixel Q_i ; $i=0(1)4r-1$ in city-block distance r from P
5 if (Q_i is a black pixel)
6 increment the histogram:

$$h_j^* = h_j + 1 \;\; ; j = round\left(i * \frac{H-1}{4r-1} \right) \quad (1)$$

7 done
8 done
9 smooth the histogram by the weighted sum of the s neighbours

$$h_i^* = \sum_{j=i-s}^{i+s} \left(1 - \frac{|i-j|}{s} \right) * h_{(j \bmod H)} \quad ; i = 0(1)H-1 \quad (2)$$

10 use the position of the global maximum in the histogram to compute the undirected skew angle

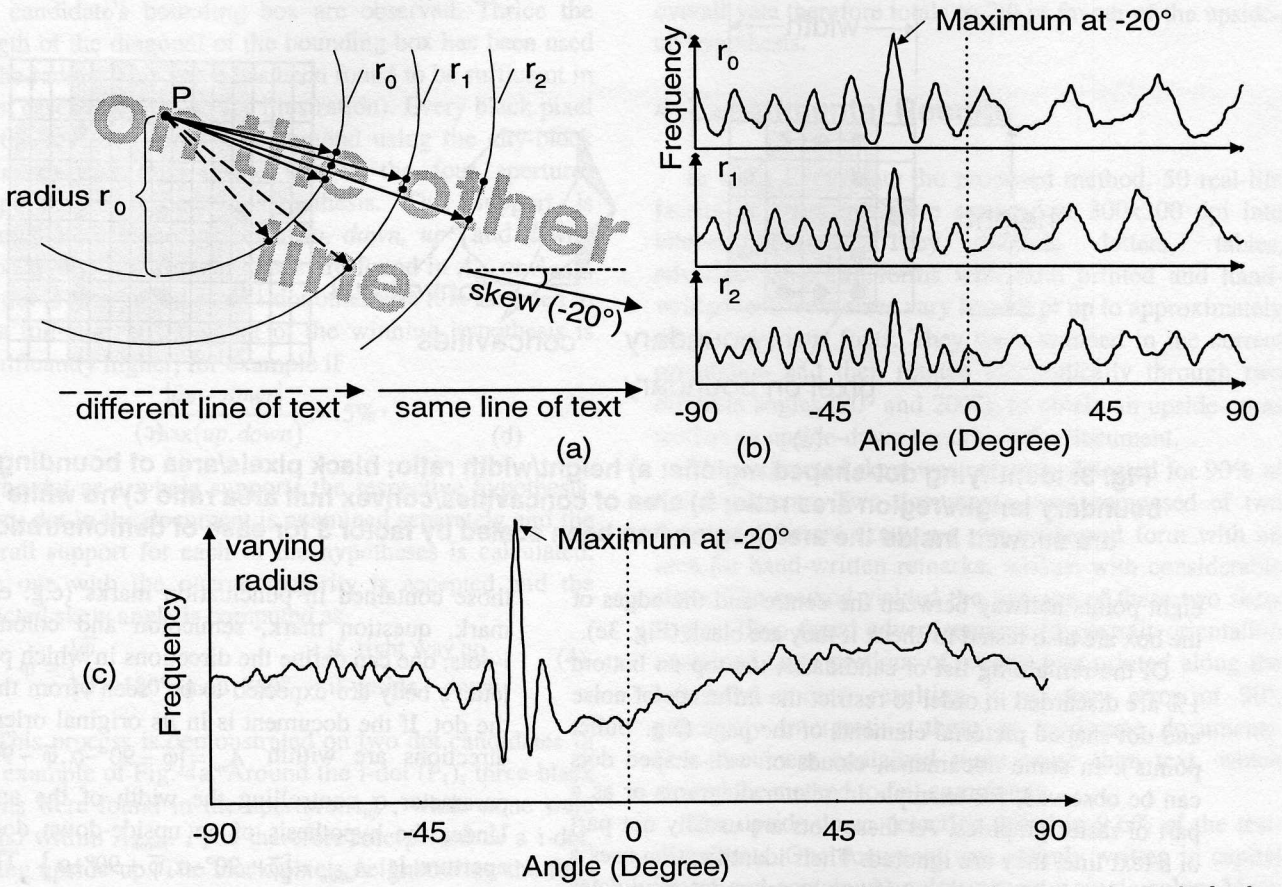

Fig. 2: Observing pixels correlation: a) the black pixels in a chosen distance r from a selected pixel *P* are investigated. b) The histogram of the angles of the lines they form shows a maximum at the overall skew angle. The positions of the secondary maxima vary with different radii. c) Superimposing the histograms of various radii further emphasises the global maximum.

3 Detection of upside-down documents

In order to detect whether a text document has been scanned upside-down or not, two simple properties of text are used. Firstly, the dots in the letters i and j are written above their body and are usually closer to their body than to the line of text above. Secondly, the full-stops terminating a sentence or abbreviation are closer to the preceding letter than to the following one. The first property depends on the way i's and j's are written, i.e. on the roman alphabet itself, while the second exploits the correlation between consecutive characters, i.e. language dependent conventions.

To detect the dots in the document, regions are detected. A set of simple constraints is then used to identify the regions that are most likely to be dot-shaped, namely: aspect ratio, area to bounding box ratio, circumference to area ratio, concavity to convex hull area ratio (see Fig. 3 and Table 1).

A dot is ideally a filled circle. The aspect ratio of the bounding rectangle is therefore expected to be close to 1. The ratio of the area of a circle and the area of its bounding square is approximated by the number of black pixels within the detected region and the area of the bounding box. The ratio between the square root of the region's area and the length of the region's boundary (i.e. the circumference of an ideal circle) has to be within a given interval. The length of the region's boundary line is obtained by a simple edge following algorithm (Fig. 3a). The region's area is expressed as the number of black pixels therein. As an ideal circle is convex, the total area of the concavities found in the region should be zero. The concavities' area is expressed as the total of all areas that have to be added to the polygonal region in order to form a convex hull (Fig. 3b).

The above conditions do not exclude regions that contain a white area, such as simple circles or the letters 'e' or 'a' in poor quality printing. Rather than testing every pixel within the region boundary, only the following ones are verified: the centre of the bounding box should be the centre of the dot and must therefore be a black pixel. The

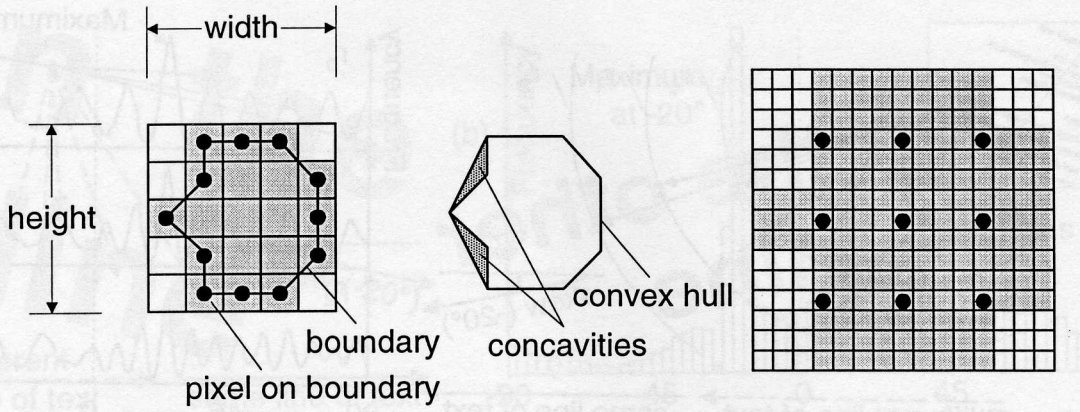

Fig. 3: Identifying dot-shaped regions: a) height/width ratio; black pixels/area of bounding box; boundary length/region area ratio; b) area of concavities/convex hull area ratio c) no white pixels are allowed inside the area (region has been scaled by factor 3 for ease of demonstration)

eight points halfway between the centre and the edges of the box are also tested to check if they are black (Fig. 3c).

Of the remaining list of candidates, the top an bottom 1% are discarded in order to restrict the influence of noise and dot-shaped pictorial elements of the page (e.g. 'bullet points'). In some documents, clouds of well-shaped dots can be observed, for example in scatter diagrams or as a part of raster graphics. As these dots are usually not part of a text line, they are ignored. Their identification uses a simple proximity criterion: If two or more dot candidates are located close to each other, all of them are discarded. The proximity threshold has been experimentally determined to be twice the average of the dot's diagonal extent.

In printed documents, the remaining candidates are mostly i- and j-dots and full-stops, with the exception of those contained in punctuation marks (e.g. exclamation mark, question mark, semicolon and colon). For the i-dots, one can define the directions in which pixels of the letter's body are expected to be "seen" from the centre of the dot. If the document is in its original orientation, the directions are within $A_{up} = [\overline{\varphi} - 90° - \alpha, \overline{\varphi} - 90° + \alpha]$, the parameter α controlling the width of the aperture A_{up}. Under the hypothesis of an upside-down document the aperture is $A_{down} = [\overline{\varphi} + 90° - \alpha, \overline{\varphi} + 90° + \alpha]$. The body of the letter immediately preceding the full-stop is expected to appear within $A^*_{up} = [\overline{\varphi} + 135° - \alpha, \overline{\varphi} + 135° + \alpha]$ and $A^*_{down} = [\overline{\varphi} - 45° - \alpha, \overline{\varphi} - 45° + \alpha]$ without any further pixels in A_{up} and A_{down}.

The pixels at a common distance from the centre of the

Table 1: Criteria to identify dot-shaped regions of connected black pixels.

Criterion	acceptancy range	ideal value (circle)	used value	value observed in Fig. 3
aspect ratio of bounding box	$\frac{height}{width} \in \left[\gamma_{hw}, 1/\gamma_{hw}\right]$	$\gamma_{hw} = 1$	$\gamma_{hw} = 0.75$	$5/5 = 1.0$
black pixels to bounding box area	$\frac{blackPixels}{boundingArea} \in \left[\gamma^-_{bb}, \gamma^+_{bb}\right]$	$\gamma^-_{bb} = \gamma^+_{bb} = \pi/4 \approx 0.79$	$\gamma^-_{bb} = 0.6$ $\gamma^+_{bb} = 1.0$	$19/25 \approx 0.76$
circumference to area ratio	$\frac{\sqrt{blackPixels}}{circumPixels} \in \left[\gamma^-_{cc}, \gamma^+_{cc}\right]$	$\gamma^-_{cc} = \gamma^+_{cc} = 1/2\sqrt{\pi} \approx 0.28$	$\gamma^-_{cc} = 0.2$ $\gamma^+_{cc} = 1.0$	$\frac{\sqrt{19}}{12} \approx 0.36$
area of concavities	$\frac{concavityArea}{convexHullArea} \in \left[\gamma^-_{cv}, \gamma^+_{cv}\right]$	$\gamma^-_{cv} = \gamma^+_{cv} = 0$	$\gamma^-_{cv} = 0.0$ $\gamma^+_{cv} = 0.3$	$1/13 \approx 0.077$
white pixels inside region	no white pixels allowed	test all pixels in the circle	test selected pixels only	all 9 pixels are black

dot candidate's bounding box are observed. Thrice the length of the diagonal of the bounding box has been used as the radius. This value has been found to be sufficient in most cases (see Fig. 4b for illustration). Every black pixel on the resulting circle (or diamond using the city-block distance) that falls inside one of the four apertures supports the respective hypothesis. This support is accumulated in the variables *up*, *down*, *up** and *down** respectively. If sufficient pixels are found in A_{up} or A_{down}, the dot is interpreted as an i-dot otherwise it is assumed to be a full-stop. If the count of the winning hypothesis is significantly higher, for example if

$$\frac{|up - down|}{\max(up, down)} > 15\%, \qquad (3)$$

the dot as a whole supports the respective hypothesis. Every dot in the document is examined separately and the overall support for each of the hypotheses is calculated. The one with the overall majority is accepted and the directed skew angle is computed as

$$\varphi = \begin{cases} \overline{\varphi} & ; \text{if right way up} \\ \overline{\varphi} + 180° \mod 360° & ; \text{if upside-down} \end{cases}. \qquad (4)$$

This process is demonstrated on two dot candidates in the example of Fig. 4a. Around the i-dot (P_1), three black pixels were found in the aperture A_{up} , while none were found within A_{down}. P_1 is therefore interpreted as a i-dot, voting upside-up. The black pixels neighbouring the full-stop (P_2) are not found within the apertures for i-dots. Under the full-stop interpretation, however, a majority of five for the upside-up hypothesis can be observed. The overall vote therefore totals to 2:0 in favour of the upside-up hypothesis.

4 Experimental Results

In order to evaluate the proposed method, 50 real-life facsimile documents were scanned at 300x300 dpi into bilevel images. They include letters, tables, advertisements and forms with both printed and hand-written text. Text sizes vary from 8 pt up to approximately 40 pt in various fonts. They were scanned in the correct orientation and then rotated electronically through two different angles (20° and 200°), to obtain an upside-up as well as an upside-down version of the document.

The undirected skew was correctly detected for 90% of the documents. Two documents were composed of two areas of different skew, e.g. an unskewed form with an area for hand-written remarks, written with considerable skew. The method yielded the average of these two skew angles. Two faxed advertisements in portrait orientation contained large portions of reverse text printed along the left hand corner, resulting in a skew error of 90°, effectively interpreting them as landscape documents. One document contained more noise than text, which caused the method to fail completely.

The upside-down detection failed in 9.6% of the test-documents. One document was entirely written in capital letters and contained no dots except a single colon. Many of the hand-written dots did not follow the above constraints. The method therefore often failed in cases

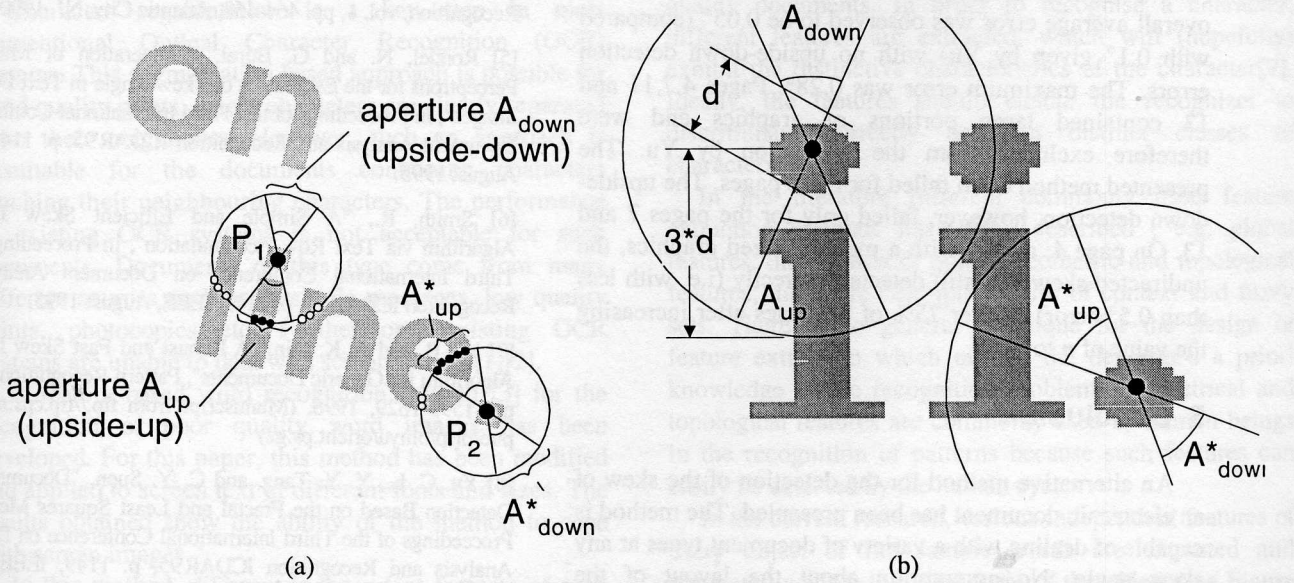

Fig. 4: Upside-down Detection: Testing i-dots and full-stops. a): black pixels marked as filled circles are within one of the hypotheses' apertures and therefore in support. Unfilled circles mark pixels that satisfy the distance constraint but do not appear in any of the apertures. b) the apertures in skewless script.

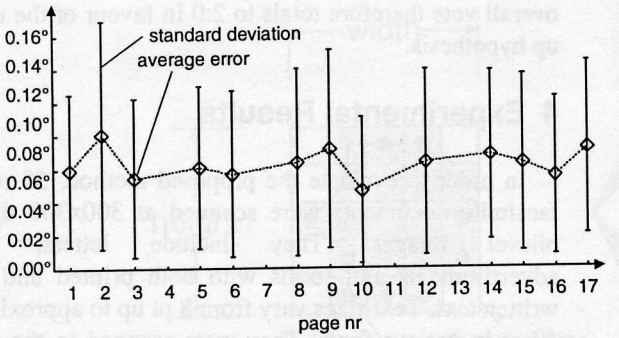

Fig. 5: Average and standard deviation of the skew detection error

were the majority of text was hand-written.

Comparison of the observed accuracy with the results reported for other methods is difficult, as different sets of test images have been used that were acquired under different circumstances. Furthermore, it is not always possible to manually determine the accurate skew angle as a reference. Following the method proposed by Yu[9], the manuscript of that paper has been converted into its electronic version without intermediate steps for printing and scanning. Using the ghostscript software package, 17 images have been produced at a 200x200 dpi resolution. These synthetic documents have then been rotated by -180 to +180 degrees in steps of 9°. The radius r was allowed to vary between 600 and 1000 pixels (i.e. 7.6cm to 15.2cm); n=1000 black pixels were observed per page; the opening of the aperture was set to α=30°. The average error of the undirected skew angles is shown in Fig. 5 for each page together with the standard deviation. The overall average error was observed to be 0.05° (compared with 0.1° given by Yu) with no upside-down detection errors. The maximum error was 0.28°. Pages 4,7,11 and 13 contained large portions of graphics and were therefore excluded from the evaluation by Yu. The presented method often failed for these pages. The upside-down detection, however, failed only for the pages 7 and 13. On page 4, a page with a medium sized graphics, the undirected skew was still detected correctly (i.e. with less than 0.5° error) in over 75% of the cases after increasing the value of n to 2000.

5 Conclusion

An alternative method for the detection of the skew of an electronic document has been presented. The method is capable of dealing with a variety of document types at any skew angle. No assumption about the layout of the document has to be made nor has the detectable skew to be restricted to a given interval. The accuracy of the method has been demonstrated to be similar if not superior to alternative approaches, even though the complexity of the algorithm is considerably lower. The proposed method for the detection of upside-down documents, however, rests on the assumption of detectable dots, which does not hold for all types of documents, especially hand-written ones. Alternative methods have to be developed to address this problem, exploiting additional properties of text such as left-alignment of paragraphs or the relationship between the number of ascenders and descenders in the text.

6 References

[1] Amin, A. and R. Shiu, "New Skew Detection and Correction Algorithms", in Proceedings of the Fifth International Workshop on Frontiers in Handwriting Recognition IWFHR-5, p. 251, Essex, England, September 1996.

[2] Baird, H. S., B. Yu, and A. K. Jain, "A Robust and Fast Skew Detection Algorithm for Generic Documents", Pattern Recognition, vol. 29, pp. 1599-1629, 1996.

[3] Hennig, A., N. Sherkat, and R. J. Whitrow, "Zone Estimation for Multiple Lines of Handwriting Using Approximating Spline Functions", in Proceedings of the Fifth International Workshop on Frontiers in Handwriting Recognition IWFHR-5, p. 325, Essex, England, September 1996.

[4] Hinds, S. T., J. L. Fisher, and D. P. d'Amato, "A Document Skew Detection Method Using Run-Length Encoding and the Hough Transform", in International Conference on Pattern Recognition, vol. 1, pp. 464-468, Atlantic City, NJ, 1990.

[5] Rondel, N. and G. Burel, "Cooperation of Multi-Layer Perceptrons for the Estimation of Skew Angle in Text Document Images", in Proceedings of the Third International Conference on Document Analysis and Recognition ICDAR'95, p. 1141, IEEE, August 1995.

[6] Smith, R., "A Simple and Efficient Skew Detection Algorithm via Text Row Accumulation", in Proceedings of the Third International Conference on Document Analysis and Recognition ICDAR'95, p. 1145, IEEE, August 1995.

[9] Yu, B. and A. K. Jain, "A Robust and Fast Skew Detection Algorithm for Generic Documents", Pattern Recognition, vol. 29, pp. 1599-1629, 1996. (Manuscript from ftp://ftp.cps.msu.edu/-pub/prip/binyu/orient.ps.gz)

[8] Yu, C. L., Y. Y. Tang, and C. Y. Suen, "Document Skew Detection Based on the Fractal and Least Squares Method", in Proceedings of the Third International Conference on Document Analysis and Recognition ICDAR'95, p. 1149, IEEE, August 1995.