Skew Estimation in Digitized Documents: A Novel Approach

D. S. Guru
Dept. of Studies in
Computer Science, UoM,
Mysore
guruds@mailcity.com

P. Punitha
Dept. of Studies in
Computer Science, UoM,
Mysore
punithaswamy@yahoo.com

S. Mahesh
Dept. of Studies in
Computer Science, UoM,
Mysore

Abstract

In this paper, a novel and efficient method for skew detection in document images is proposed. The proposed method is based on connected component blotching and linear regression. The proposed algorithm works well irrespective of the type of script even for a wide range of skew angle (with in ±90º).

1. Introduction

Skew or tilt in images are unavoidable due to their incorrect positioning which is indiscernible when the documents are fed into the scanners/optical sensors to obtain a digital document. Skew angle in digital documents can be defined as the angle made by the text lines of a digital document with that of the direction of the x-axis of the co-ordinate system. The detection of skew angle followed by skew correction helps in making the subsequent steps viz., segmentation, feature extraction, classification, document layout analysis, representation, recognition in document image analysis more intelligent and fast [3], [4]. Hence, skew angle detection is a major and fundamental step in document image analysis.

There have been continuous attempts made by the research community to come out with flexible and intelligent skew detection scheme based on various concepts. The insinuated methodologies can be broadly classified into categories, viz., projection profile based [22], [9], [6], [17], [2], [5], [20], nearest neighbor clustering based [15], [19], [21], transformation based, cross-correlation based [10] and some based on combined application of the others.

Though, projection profile based methods give good results, they are computationally expensive and are highly accurate if the skew angle is limited to ±15°.

The transformation based methodologies can further be classified as Hough Transform based [11], [24], [26], Fourier Transform [22], [14] based and Morphological Transform based [8]. In order to reduce the computational time of Hough transformation, [16] and [18] were proposed. In contrast to [18], [3] gives accurate results for documents skewed up to ±45°. The major drawbacks of all Hough transform based methodologies are that they are seldom used for documents without extensive modifications. Fourier transformation based methodologies require the computation of Fourier transform of the document and hence are computationally expensive. The morphological transforms (opening and closing) based approach [8], though works for synthetic images, it exhibits worse performance on the real document images.

The cross-correlation based method [10] is claimed to be more efficient than Hough transform based [26] method. However, the authors assume that the skew angle is utmost ±15°. Another set of alternative methods [23], [1],[7] were also proposed.

[7] designed a method specifically for Indian scripts like Devanagari and Bangla that have a head line (Shirorekha or Matra). The results of this skew detection methodology are proved to be comparable to those of Hough transform based methods but with less computation. Nevertheless, the method is completely dependent on the head line, which join the characters in the word and makes the word appear as a single component.

Skew detection of scripts having a head line (Shirorekha or Matra) is easier as the headlines are themselves sufficient to detect the skew angle. Detection of skew angle in English documents is much tougher when compared to Devanagari and Bangla scripts as English scripts do not have headline. Hence there are many methods in literature for skew detection in English scripts. Nevertheless, skew detection in English scripts, although they do not have any head line (Shirolekha or Matra) is relatively simpler as English characters are free from vowels or consonant modifiers below a character and in addition, have a predominant base line. Therefore the methods developed for English text are not applicable to all Indian scripts especially for the scripts which are rich in vowels/consonant modifiers viz., Kannada, Telugu scripts. Thus, the core of the problem lies in designing an effective scheme to detect the skew angle irrespective of the script types, which can be used for skew correcting so that the performance of the subsequent steps in document image analysis

would be more accurate. The devised scheme is expected to be capable of detecting any arbitrary angle with out restricting itself to a certain range.

In view of this, in this paper a novel and efficient method for detecting skew angle, an indispensable component in scanned documents is proposed. The proposed method is based on connected component blotching and linear regression. The document image is subjected to connected component blotching to obtain line segmented document image. The line segmented document image is then skeletonised to obtain linear line like structures. Subsequently, all line segments of relatively larger length are chosen and subjected to linear regression analysis to obtain their slope angles. Once the set of candidates for the skew angle are obtained, the values in the set are sorted in ascending order and the candidates with considerably large difference are eliminated from the list. The average of the remaining values in the set gives a more accurate skew angle. The proposed algorithm works well irrespective of the type of script even for a wide range of skew angle (with in ±90º).

The remaining part of the paper is organized as follows. Section 2 proposes a novel scheme for skew angle detection while section 3 gives the experimental results. The efficacy of the proposed model along with some conclusions is brought out clearly in section 4.

2. The proposed skew detection scheme

The proposed scheme has two stages. The first stage proposes a method of skeletonising the document image, while the second stage deals with the skew estimation.

2.1 Document image skeletonisation

The proposed document image skeletonisation stage finds a minimum blotch, for each connected component present in the document image. The resulting image is then subjected to morphological closing to obtain image with text line blocks. The text line blocks are subsequently subjected to thinning to obtain the skeleton of the document image. Thus, the following are the major steps involved in document image skeletonisation.

2.1.1 Connected component blocking

Instead of processing all black pixels in the document image it is more efficient to process only some selected black pixels to find the skew angle of the document. Rather than choosing black pixels located on the bottom-line of the text or on the top-line of the text, it is more appropriate to choose the pixels lying on the mid-line of the text. However, searching for the pixels lying on mid-line of the text is not only cumbersome but also inefficient as the distance between the resulting pixels pertaining to each character will be considerably large and will be of less help in skew estimation. Thus, we recommend enclosing each connected component with a minimum blotch and transforming the image into text line blocked image so that the resulting text blocked image can be subjected to thinning to obtain line like structures for the text lines useful for skew detection. However, the only assumption made here is that between-line spacing is greater than with-in line spacing. Formally, if I_d is a document image with m connected components then all m connected components are enclosed either by a minimal circular blotch or by a minimum bounding rectangular blotch. If 'B' is a connected component to be blotched, then the component is blotched with a circle of radius d, (the maximum of all the distances from the centroid (x_B, y_B) of 'B' to the boundary pixels of 'B') with (x_B, y_B) being the center if d is less than the threshold T_1, otherwise with a minimum bounding rectangle. This is due to the fact that, if d is large, enclosing 'B' by a circular blotch of radius d connects the components belonging to adjacent lines, which is to be avoided. Enclosing a connected component by a minimum rectangular blotch is subtle as it is expected to preserve the orientation of the connected components. Perhaps one such rectangle could be the minimum sized circumscribing rectangle. Hence, we find the axis of least inertia [13], which helps us in estimating the orientation of the connected component.

2.1.2 Text line blocking through morphological closing and thinning

Once the document image is subjected to the connected component blocking stage, all connected components of the document image are blotched and almost all connected components within a word are merged together to form a single block of word (see Fig. 2). Further to connect and merge the word blocks into lines, morphological closing is performed on the connected component blocked image with a suitable structuring element and all isolated word blocks belonging to the same text line are merged together, producing an image with text line blocks. Indeed, text line blocks are necessary as they are more appropriate and helpful for skew angle detection than the individual word blocks. When Fig. 2 is subjected to morphological closing with the structuring element shown in Fig.3, Fig.4 is obtained as the text line blocked image. It can be observed that morphological closing (with suitable structuring element) connects the blocked words belonging to the same text line. However, the text line blocked image may sometimes contain small holes (single pixel islands)(see Fig.4), which may give rise to breakage in lines during thinning. Hence, these holes are filtered out by using a suitable filter such as median filter. The

filtered image is now free from isolated holes and consists of only blocked solid text lines and rectangular blotches for picture objects present in the document image (See Fig.5). The thinning algorithm [12] is then applied on the filtered image to convert elongated blocks of text lines into lines of single pixel thickness. Most of the thinned lines in this output image represent text lines and they have almost same orientation as the original document image. Selecting some relatively lengthier lines from this skeletonised document image (Fig.6), fitting best lines to these selected lines and then computing their slope determines the skew angle of the document image as explained in the following section.

2.2 Skew estimation and detection

All relatively lengthier lines are selected from skeletonized document image obtained as explained in the previous section. Each selected line is subjected to linear regression and a best straight line is fit. For each straight line i, the slope angle θ_i is computed as the candidate for the skew angle. Though, one may feel that the average of all the candidate skew angles is the skew angle of the document image, it is not true, due to the fact that, some of the selected thinned lines corresponding to non-text components (pictures, tables, etc) bear entirely different orientation. In such a situation, the conflict in choosing a unique value for the skew angle is resolved with the following post processing technique.

The angles computed for the selected lines are sorted and the absolute difference between the first two elements is computed. If this absolute difference is greater than T_2, where T_2 is the allowable precision which depends on the user, the first element is removed from the set. Now, the absolute difference between the last two elements is computed. If the difference happens to be larger than T_2, the last element is removed from the list. This process is repeated alternatively from both ends of the sorted list with the next element as the candidate for elimination. This process is repeated as long as applicable. This process not only retains the correct candidates for the computation of actual skew angle but also removes the erroneous candidates. The average of the remaining values in the set gives the most accurate skew angle.

For the example considered, skew angle of the document is estimated by fitting best lines to all thinned lines whose length is greater than the threshold value T_2 = 70. The result of least square fitting (best line fitting) on selected skeletons of Fig. 6 is shown in the Fig. 7. The set of candidate slope angles corresponding to these best fitted lines is {-79.411, -5.917, -6.804, -5.121, -4.916, -6.253, -4.845, -5.512, -4.504, -5.145}. The values in the set are initially sorted in ascending order and is {-79.411,-6.804, -6.253, -5.917, -5.512, -5.145, -5.121, -4.916, -4.845, -4.504}. Now, the absolute difference between first and second value in the set is determined and is found to be 72.607. This value is greater than the chosen threshold value T_3 = 0.4. Hence, -79.411 is removed from sorted set and the set is reduced to {-6.804, -6.253, -5.917, -5.512, -5.145, -5.121, -4.916, -4.845, -4.504 }. The absolute difference between last two values is found to be 0.341. It is not greater than 0.4. Hence, the last value -4.504 remains in the set. The difference between -6.804 and -6.253 is also greater than 0.4 and hence, -6.804 is removed from the set resulting with {-6.253, -5.917, -5.512, -5.145, -5.121, -4.916, -4.845, -4.504}. This process is repeated until all erroneous elements are eliminated from the list. The list finally reduces to {-5.512, -5.145, -5.121, -4.916, -4.845, -4.504}. The average of all the values in the reduced set is found to be -5.007 which in turn is the more approximate value of skew angle of the document image shown in Fig.1.

The following algorithm has thus, been devised to detect the skew angle of a document image

Algorithm : Skew angle detection

Input : I, binary document image.
T_1, T_2, T_3 Threshold.
S_1, Structuring element.

Output : θ, skew angle of document image I.

Method :

Step1: Enclose each connected component in the document image with a minimum enclosing circular/rectangular blotch depending on T_1 and the distance d from the centre of the connected component to a farthest pixel on its boundary as explained in section 2.1.1.

Step2: Perform morphological closing operation on the image obtained from step1 using structuring element S_1.

Step3: Filter out small holes if any in the blocked image.

Step4: Thin the blocked text lines by performing thinning operation.

Step5: For each thinned line whose length is greater than T_2

- **(a)** Fit a best line.
- **(b)** Compute the slope angle and store it in a sorted order.

For end

Step6: Do post processing as explained in section 2.2.

Step7: Compute the skew angle θ as the average of the remaining values in the list.

Algorithm ends.

3. Experimental results

To corroborate the efficacy of the proposed methodology, we have conducted several experiments on document images of various scripts with different skews. Out of them we present only few results. For

this experimentation, we have considered some Kannada documents scanned by flatbed scanner at a resolution of 100 dpi. The documents were tilted by a pre-specified angle lying in the interval [–90°, 90°]. This angle is considered as true skew angle. The results obtained for Kannada document images scanned at different orientations are shown in the Table-1. For this experimentation the thresholds were fixed to be T_1=10 pixels, T_2=70 pixels, T_3=0.5.

Table-1: Actual and estimated skew angles (in degrees) for Kannada documents

True skew angle (in degrees)	Estimated skew angle (in degrees)	True skew angle (in degrees)	Estimated skew angle (in degrees)	True skew angle (in degrees)	Estimated skew angle (in degrees
3	3.190	-20	-19.173	60	60.453
-3	-3.148	30	30.485	-60	-59.293
5	5.321	-30	-29.850	70	70.543
-5	-4.598	40	40.281	-70	-69.555
10	10.322	-40	-39.788	80	80.279
-10	-9.762	50	50.276	-80	-79.458
20	20.243	-50	-49.794		

4. Discussion and Conclusion

Skew detection and correction indeed helps the subsequent stages in document image analysis. Devising schemes which can detect skew angles accurately, irrespective of the scripts and range of skew angles is a challenging task in the field of document image analysis. In fact these are the shortcomings existing in almost all the existing methodologies. Though many models were devised for skew detection, they are applicable to specific types of scripts and are not capable of detecting a wide range of skew angles. In view of this, in this paper we have made a successful attempt in exploring a model that overcomes the aforementioned shortcomings. The method transforms the document image into a text blocked image through connected component blocking and morphological closing. The image is then subjected to thinning to obtain a skeletonised document image. Line like structures of considerable length are chosen and best lines are fit. The slopes of the line segments are the candidates for skew angle and the skew angle is finally decided by eliminating the erroneous candidates and taking the average of the remaining.

The experimental results shows that the proposed algorithm works well for all types of documents containing both body-text and non-body text (graphics, tables, drawings, etc). It is evident that the proposed algorithm can detect large skew angles (between $\pm 90°$) in a document. The proposed method yields more accurate values than document spectrum method [21] in almost all cases. Although Hough transform based methods [18] work for any range of skew angles, they are less accurate for documents having pictures, drawings, etc. Thus, our method is better than even Hough transform based methods. The proposed method is also more efficient than methods proposed by [26, 7] as these methods are not suitable for complex documents like Kannada and Telugu scripts.

A comparison of the proposed method with some of the existing methodologies is given in Table-2.

In summary, the paper presents a novel approach to detect skew angles in digitized document images. The proposed method is based on morphological closing. The proposed method is more efficient than any other skew detection methods as it works well irrespective of the type of script and also for wide range of skew angles (within $\pm 90°$). Even though, we have considered only Kannada and English scripts for experimentations, the study made in this paper reveals that the method works well for any type of script, which contains more number of text lines when compared to that of non-body text

Table-2: Comparison of accuracy of the proposed skew detection scheme with that of some of the existing methodologies

True skew angle	Estimated skew angles by different methods and by the proposed method				
	[18]	[7]	[21]	[26]	Proposed method
3	2.80	3.14	3.82	3.69	3.10
5	5.24	5.03	5.74	5.01	5.16
10	10.67	9.96	10.71	10.46	9.83
20	19.25	19.52	19.16	19.53	19.93
30	29.40	29.85	29.11	29.88	30.06

(pictures, head lines, tables, etc). Our future work concentrates on dynamic selection of threshold value.

References

[1] H.K. Aghajan and T. Kailath, SLIDE: subspace based line detection, PAMI, Vol. 16, No. 11, 1057-1073, 1994.

[2] T. Akiyama and N. Hagita, Automated entry systems for printed documents. 23(2), 1141-1154, 1990.

[3] A. Amin and S. Fisher, A document skew detection method using the Hough Transform. Pattern Analysis and Applications-3. 243-253, 2000.

[4] Avanindra and S. Chaudhuri, Robust detection of skew in document images, IEEE Trans. On image processing – 6(2), 1997.

[5] A. Bagdanov and J. Kanai, Projection profile based skew estimation algorithm for JBIG compressed images, Information sxience research Institute, University Nevada, Las Vegas, USA, 1997.

[6] H. S Baird, The skew angle of printed documents. In proceedings of SPSE 40th Conf. Symp. Hybrid imaging systems, Rochester, Newyork, 21-24, 1987.

[7] B. B. Chaudhari and U. Pal, Skew angle detection of digitized script document, PAMI, Vol. 19, No. 2, 182-186, 1997.

[8] S. Chen, R. M. Haralick and I. T. Phillips, Automatic text skew estimation in document images, Proceedings of 3rd Int. Conf. on DA&R (ICDAR-95), 1995.

[9] G. Ciardiello, G. Scafur, M. T. Degrandi, M. R. Spada and M. P. Roccoteli, An experimental system for office document handling and text recognition for office document handling and text recognition. In proceedings of ninth conference on PR, 739-743, 1988.

[10] B. Gatos, N. Papamarkos and C. Chamzas, Skew detection and text line position determination in digitized documents, PR vol 36(9), 1505-1519, 1997.

[11] B. Gatos, S. J. Perantonis and N. Papamarkos, Accelerated Hough transform using rectangular image decomposition, Electronic Letters 32(8), 730-732, 1996.

[12] R. C. Gonzalez and R. E. Woods, Digital image processing, Pearson Edu., Inc., 2001.

[13] D. S. Guru, Towards accurate recognition of objects employing a partial knowledge base: some new approaches, Ph. D. Thesis, Department of Studies in Computer Science, University of Mysore, Manasagangothri, Mysore, India, 2000.

[14] M. Hase and Y. Hoshini, Segmentation method of document images by two-dimensional Fourier transformation, Systems and Computers in Japan, Vol. 16, No. 3, 1985.

[15] A. Hashizume, P. S. Yeh, A. Rosenfeld, A method of detecting the orientation of aligned components, PRL-4, 125-132, 1986.

[16] S. C. Hinds, J. L. Fisher and D. P. D' Amato, A document skew detection method using runlength encoding and the Hough transform. In Proc. 10th Int. Conf on PR, 464-468, 1990.

[17] Y. Ishitani, Document skew detection based on local region complexity, IEEE-7, 49-52, 1993.

[18] D. S. Le, G. R. Thoma and H. Wechsler, Automated page orientation and skew angle detection for binary document images, PR 27(10), 1325-1344, 1994.

[19] J. Liu, C. M. Lee, R. B. Shu, An efficient method for the skew normalization of a document image. Proc. of 11th International conference on Pattern Recognition, ICPR, 152-155, 1992.

[20] Y. Nakano et al., An algorithm for the skew normalization of document images. Proc. Of 10th ICPR, Atlantic city, NJ, 8-11, 1990.

[21] L. O'Gorman, The document spectrum for page layout analysis, IEEE trans on PAMI, 15(11), 1162-1173, 1993.

[22] W. Postl, Detection of linear oblique structures and skew scan in digitized documents. Proc. 8th Int. Conf. on PR, 687-689, 1986.

[23] R. A. Smith, A simple and efficient skew detection algorithm via text accumulation. Proceedings of 3rd ICDAR, 1145-1148, 1995.

[24] S. N. Srihari and V. Govindaraju, Analysis of textual images using the Hough transform, Machine vision and Applications 2, 141-153, 1989.

[25] D. M. Tsai and M. F. Chen, Object recognition by a linear weight classifier. Pattern Recognition Letters 16, 591-600, 1995.

[26] H. Yan, Skew correction of document images using interline cross correlation, Computer Vision graphics image process: Graphical models and image processing, 55(6), 538-543, 1993.

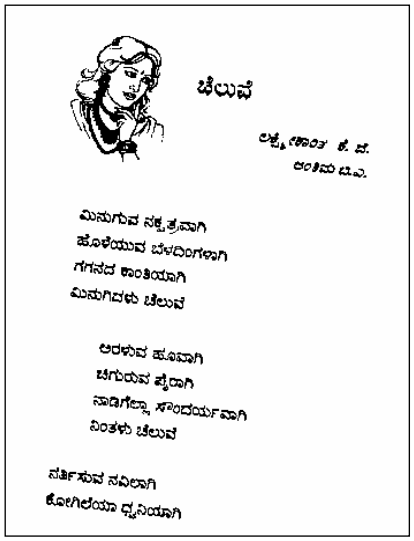

Fig. 1 A Document image tilted by -5°

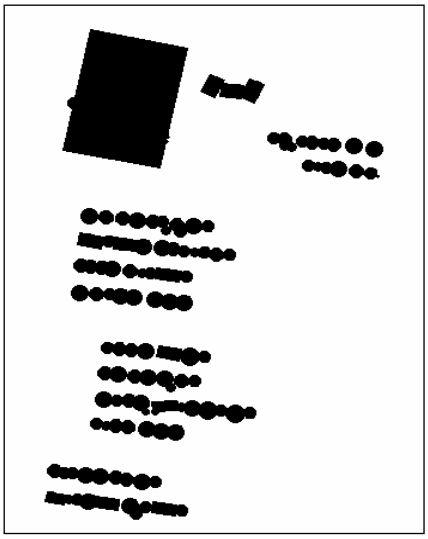

Fig. 2 Word Blocked image of Fig. 2.1

0	1	0
1	1	1
0	1	0

Fig.3 A 3×3 structuring eler

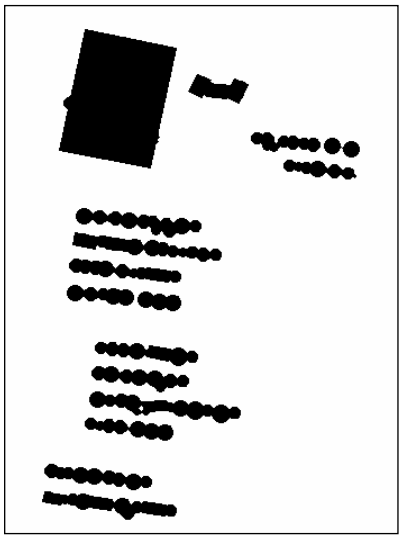

Fig. 4 Result of morphological closing on the image in Fig. 2.1

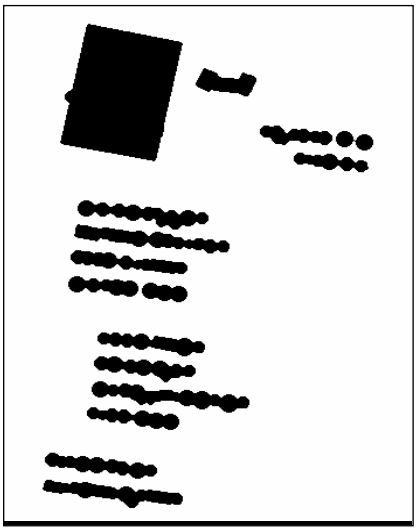

Fig.5 Result of filtering on Fig.4

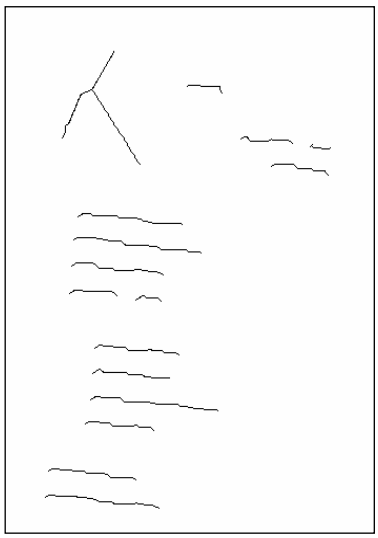

Fig.6 The result of thinning on image shown in Fig.5

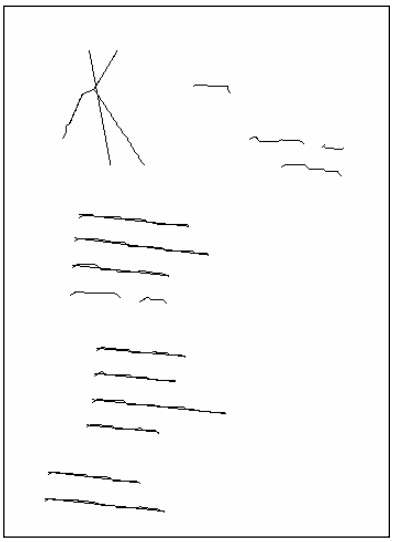

Fig. 7 The least square best fit lines in Fig. 6