A Painting Based Technique for Skew Estimation of Scanned Documents

Alireza Alaei[1], Umapada Pal[2], P. Nagabhushan[1] and Fumitaka Kimura[3]
[1]Department of Studies in Computer Science, University of Mysore, Mysore, 570006, India
[2]Computer Vision and Pattern Recognition Unit, Indian Statistical Institute, Kolkata-108, India
[3]Graduate School of Engineering, Mie University, Japan
alireza20alaei@yahoo.com, umapada@isical.ac.in, pnagabhushan@hotmail.com, kimura@hi.info.mie-u.ac.jp

***Abstract*—In this paper, we propose an efficient skew estimation technique based on Piece-wise Painting Algorithm (PPA) for scanned documents. Here we, at first, employ the PPA on the document image horizontally and vertically. Applying the PPA on both the directions, two painted images (one for horizontally painted and other for vertically painted) are obtained. Next, based on statistical analysis some regions with specific height (width) from horizontally (vertically) painted images are selected and top (left), middle (middle) and bottom (right) points of such selected regions are categorized in 6 separate lists. Utilizing linear regression, a few lines are drawn using the lists of points. A new majority voting approach is also proposed to find the best-fit line amongst all the lines. The skew angle of the document image is estimated from the slope of the best-fit line. The proposed technique was tested extensively on a dataset containing various categories of documents. Experimental results showed that the proposed technique achieved more accurate results than the state-of-the-art methodologies.**

Keywords: Skew detection; Piece-wise Painting Algorithm (PPA); Regression line; Document Analysis; Skew correction.

I. INTRODUCTION

Several techniques for skew detection are reported in literature [2-11]. These techniques may be grouped into five major categories [3] as follows: (i) Projection-Profile (PP) based techniques, (ii) techniques based on Nearest Neighbor (NN) clustering, (iii) techniques using the Hough Transform (HT) concept, (iv) techniques employing maximum variance of Transition Counts (TC) and (v) techniques using Cross-Correlations (CC) idea. Although great efforts have been done in this particular area of document image analysis to achieve a fast and reliable skew estimation method capable of handling different types of documents, problem of skew estimation is still an open issue for further research [2].

Most of the techniques in the related literature are based on projection profiles analysis [4-5]. In this type of technique, projection profiles are computed at various angles for a given document image. The skew angle of the document is then estimated based on the maximum value of projection profiles. In NN techniques [6] inclination between each connected component and its nearest neighbor is calculated, and subsequently a histogram of the inclinations is created. The peak value in the histogram specifies the estimated skew angle of the input document. The Hough transform is also utilized in the literature to estimate the skew angle of an input document. The peak in Hough space indicates the skew angle of the input document [7-8]. Methods in the fourth category [9] use the maximum variance of transition counts (TC) to estimate skew angle of a given document image. In addition, there are methods that work based on cross-correlations (CC) [10-11] to estimate skew angle of a given document image.

Recently, some techniques have used piece-wise coverings of objects by parallelograms (PCP) for estimating skew angle of a scanned document [2, 3]. In [3] a document image is initially split into a number of non-overlapping slabs and subsequently parallelograms at various angles are generated to cover the components in each slab. The angle at which maximum white space is obtained indicates the estimated skew angle of the document. In [2] an enhanced algorithm based on piecewise covering by parallelogram (e-PCP) has been proposed. In e-PCP, a criterion has been used to classify a document as a landscape or portrait mode. Furthermore, authors have introduced a confidence measure for filtering the estimated skew angles that may not be reliable [2]. They have also mentioned that the e-PCP algorithm has achieved faster and better performance than PCP, PP, HT and NN based methods.

In this paper, we propose a Piece-wise Painting Algorithm (PPA) based skew estimation technique for scanned documents. Applying the PPA on both the directions (horizontal and vertical), two painted images (one for horizontal and the other for vertical) are obtained. Employing some statistical information, regions with specific height (width) from horizontal (vertical) painted images are then selected and top (left), middle (middle) and bottom (right) points of such regions are grouped into 6 separate lists. Utilizing linear regression and line drawing concepts, a few lines are obtained from the lists of the points and their individual slopes are calculated. A voting technique is also proposed to find the best-fit line amongst all the lines and consecutively skew angle of the input document is estimated from the best-fit line. The proposed skew estimation technique is not dependent on writing mode (vertical or horizontal) of a document. Rotation of whole image is an expensive operation used for skew detection in e-PCP and PCP algorithms and our method does not use it for skew angle detection. Hence, our proposed technique improved the results obtained by both the PCP [3] and e-PCP [2] in both the accuracy as well as computing time aspects.

II. PROPOSED SKEW ESTIMATION TECHNIQUE

The skew estimation technique proposed in this paper comprises of different stages: a) Piece-wise Painting Algorithm (PPA), b) selection of candidate bands and

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.68

c) determination of the best-fit line and skew estimation. Each stage is described in detail in the following sub-sections.

A. Piece-wise Painting Algorithm (PPA)

The concept of painting and Piece-wise Painting Algorithm (PPA) are introduced in [13] and are used for handwritten text-line segmentation. In this research work, we extend this concept for estimation of skew. If a rectangle is rotated in any angle (α), the angle, which is made by a side with respect to an axis, is equal to the angle made by its perpendicular side with the perpendicular axis (Figure 1(b)). A skewed line can be considered as a number of piece-wise horizontal lines or piece-wise vertical lines. Considering this concept and the fact that the direction or flow of writing in a document image is unknown (it may be horizontal, vertical or both), the PPA is employed in both the horizontal and vertical directions to get the sides of a rectangle. Application of the PPA in horizontal direction facilitates to find out the skew angle of a document with respect to X-axis and in the same way; vertically applying the PPA provides the skew angle of a document with respect to Y-axis. This novel idea helps us to deal with different types of documents and also to achieve better skew estimation. It is worth mentioning that the PPA can fairly handle both binary and gray-level document images containing different contents [13].

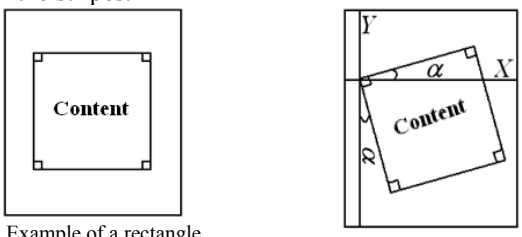
In horizontal painting, initially, the input text-image is decomposed into vertical stripes based on the average width of the connected components from left to right (Figure 2). To compute the average width and height of components the input document image is binarized and component labeling is applied on this binarized image. Subsequent to the division of the input image into stripes, the gray value of each pixel in each row of a stripe is modified by changing it with the average gray value of all pixels present in that row of the stripe. The average gray value (GLM) in each row-stripe is computed using the following formula.

$$\mathrm{GLM}_{ik} = \sum_j I_{ij} \Big/ SW_k \quad \textit{where } k=1 \textit{ to no. of stripes}$$

$$j = (k-1) \times SW_k + 1 \quad \textit{to} \quad k \times SW_k$$

$$i = \textit{row number}$$

The GLM_{ik} is the average gray value of all the pixels placed in the i^{th} row and k^{th} stripe, I_{ij} is the gray value in the i^{th} row and j^{th} column of the input image I, and SW_k is the width of k^{th} stripe. Since, the input document image is divided from the left to right (top to bottom) the last stripe may have a width (height) smaller than width (height) of rest of the stripes.



a) Example of a rectangle containing content portion

b) After rotation by an angle α.

Figure 1. Outlook of a document (a) before, (b) after rotation (by α^o).

The resultant gray-scale image is converted into two-tone image. Both the gray-level and two-tone images are shown in Figure 3 and Figure 4, respectively. The black and white rectangles represent the foreground (text regions) and background, respectively.

The similar procedure is performed on the original document in vertical direction to decompose the image into horizontal stripes based on the average height of components (Figure 5).

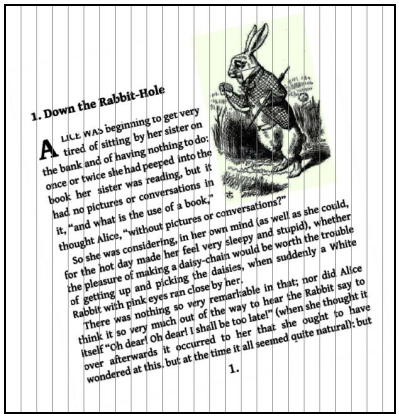


Figure 2. Decomposition of an input document image into stripes based on average width of components

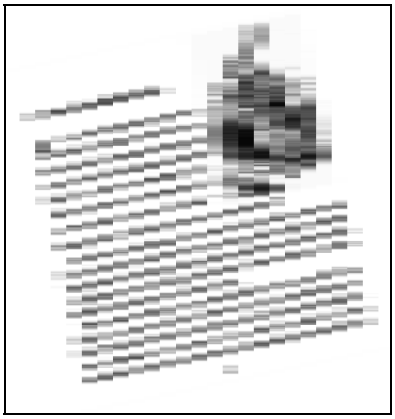


Figure 3. Gray-level painted document image after applying horizontally the Piece-wise Painting Algorithm on Figure 2

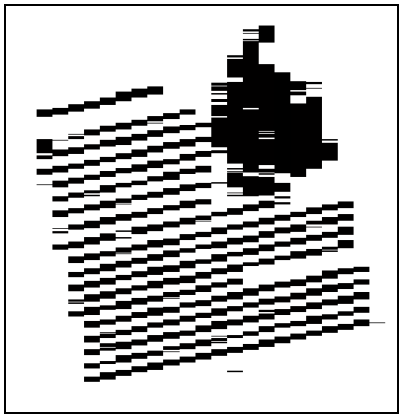


Figure 4. Binarized painted document image obtained applying binarization on Figure 3

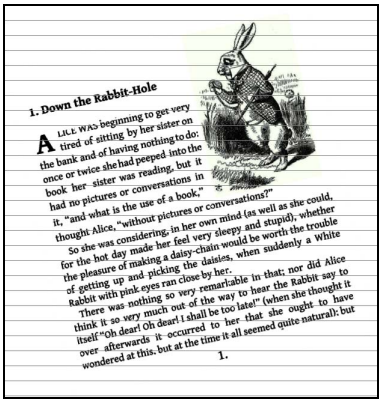


Figure 5. Decomposition of an input document image into stripes based on average height of components

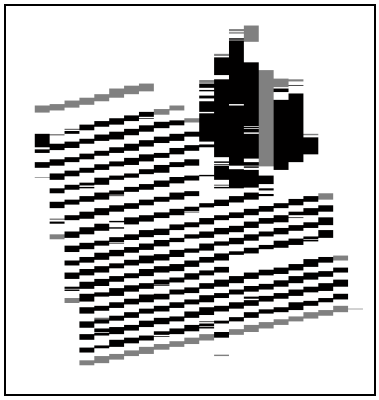


Figure 6. All vertical bands with their top and bottom rectangles are marked by gray

B. Selection of candidate bands

Employing the PPA on the original document in both the horizontal and vertical directions, two separate painted images are obtained. For each stripe of the horizontally painted image, the first and the last rows, which belong to the top and the bottom rectangular black areas of that stripe, are listed. In Figure 6 the horizontally painted image, the first, and the last rows of the rectangular black areas in each stripe is shown. For better representation, white spaces between the top and bottom black rectangles in each stripe is filled (by converting the white pixels into black pixels) to clearly show the length and position of a created long rectangular area in each stripe called band (see Figure 7). As a result, lengths (heights) and positions of all vertical black bands for all the stripes in the horizontally painted image are obtained.

Similarly, for the vertically painted image the first and the last columns of each stripe, which belong to the left and right most rectangular black areas of that stripe, are also grouped in a separate list. Here also the lengths (widths) and positions of all horizontal black bands in the vertically painted document are found.

As it is shown in Figure 1(b), for finding the angle α^o with respect to X-axis, we need to choose a number of vertical bands with a specific height. Therefore, in a horizontally painted image, a group of consecutive vertical bands having a specific height is selected. These bands are called vertical candidate bands. To find the specific height of such bands, statistical mode of heights of all vertical bands is obtained. This value indicates the most frequent value in the first list. The vertical candidate bands are shown in Figure 8. For the second list (widths of horizontal bands), the same process is performed and consequently the horizontal candidate bands are also obtained (see Figure 9).

C. Determination of the best-fit line and skew estimation

Starting, middle and end points of the candidate bands can be listed in 3 separate groups (one for starting point, one for middle point and the other for end points). Since, two groups of candidate bands (horizontal and vertical) are chosen, for each of them 3 separate groups and totally 6 groups of points are obtained. The first group (VT) contains the points belonging to the starting points of vertical candidate bands. The second (VM) and the third groups (VB) contain the points corresponding to the middle and end points of vertical candidate bands, respectively. For the horizontal candidate bands, three groups HL, HM, and HR are considered as starting, middle, and end points, respectively. For each list, we can fit a linear regression line [12] using all the values present in that list. Moreover, a straight line is also drawn by connecting the first and the last points in each group. The result of fitting lines on all the group of points obtained from vertical and horizontal candidate bands are shown in Figure 8 and Figure 9, accordingly. As we have 6 separate groups of points and for each group 2 lines can be fit, totally 12 separate lines and consecutively 12 slopes can be determined (see Figure 8 and Figure 9). One of these slopes can represent the skew of the document image. Experimentally, it was found that a subset of these 12 slopes using only 7 slopes can provide the best accuracy. These 7 angles were belonged to 4 regression lines obtained from VT, VM, HL and HM and 3 straight lines drawn using VT, VM and HL, respectively.

To choose those 7 slopes, 100 documents from the dataset used in [3] were randomly selected. Averages of errors for estimated skew angles using all these 12 lines are listed and they are ranked (Table I). Letters R and D used in Table I indicate errors due to the regression and the straight lines, respectively. We started by the highest ranked results and then keeping adding the next highest result one by one. The results obtained from different subsets of slopes are shown in Table II. From Table II, it is evident that the results obtained by the combination of the 7 higher ranks is the best result. Therefore, we conclude experimentally that these 7 slopes provide the most efficient performance.

TABLE I. RANKING LIST OF AVERAGE ERRORS OBTAINED FOR EACH FIT LINE

Rank	**1**	**2**	**3**	**4**	**5**	**6**
Line	**VT(R)**	**VT(D)**	**VM(R)**	**HL(R)**	**VM(D)**	**HL(D)**
Average errors	0.15921	0.16609	0.17307	0.17496	0.17898	0.17954
Rank	**7**	**8**	**9**	**10**	**11**	**12**
Line	**HM(R)**	**VB(R)**	**HM(D)**	**VB(D)**	**HR(R)**	**HR(D)**
Average errors	0.18246	0.18846	0.18954	0.18998	0.19645	0.19785

TABLE II. AVERAGE ERRORS OBTAINED FROM EACH SET OF FIT LINES

Different subsets of lines	**1**	**1,2**	**1,2,3**	**1,2, 3,4**	**1,2,3, 4,5**	**1,2,3, 4,5,6**
Average errors	0.15921	0.16972	0.16730	0.15196	0.15898	0.16954
Different lines	**1,2,3, 4,5,6,7**	**1,2,3,4, 5,6,7,8**	**1,2,3,4, 5,6,7,8, 9**	**1,2,3,4, 5,6,7,8, 9,10**	**1,2,3,4, 5,6,7,8, 9,10,11**	**1,2,3,4, 5,6,7,8, 9,10,11,12**
Average errors	**0.14846**	0.17846	0.17954	0.19836	0.21645	0.22785

To obtain the best fitting line and consecutively the best slope among the 7 slopes as estimated skew angle for the input document, a majority voting criterion is proposed as follows.

$S_r = Round\ (Absolute(S))$

$S_M = StatMode\ (S_r)$

$S_d = Absolute(S - S_M) = Absolute(s_1 - S_M, s_2 - S_M, s_3 - S_M, s_4 - S_M, s_5 - S_M, s_6 - S_M, s_7 - S_M)$

$Estimated\ skew\ angle = S\ (Minimum\ (S_d))$

where $A = (s_1, s_2, s_3, s_4, s_5, s_6, s_7)$ is the list of 7 slopes (skew values) obtained for a given document, $S_r = (s_{r1}, s_{r2}, s_{r3}, s_{r4}, s_{r5}, s_{r6}, s_{r7})$ is rounded absolute values of the list S, S_M is statistical mode of the values in Sr and $S_d = (s_{d1}, s_{d2}, s_{d3}, s_{d4}, s_{d5}, s_{d6}, s_{d7})$ is absolute of values of differences between the values of S and S_M. The skew angle having minimum difference with S_M represents the skew angle of the given document.

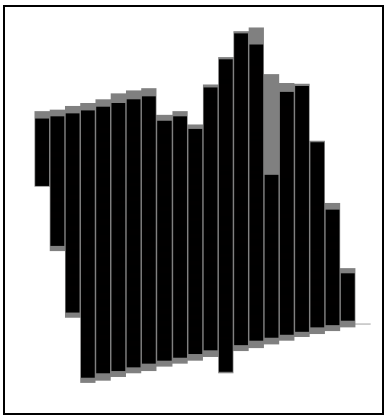


Figure 7. Vertical bands obtained after filling the white space between two black rectangles from Figure 6.

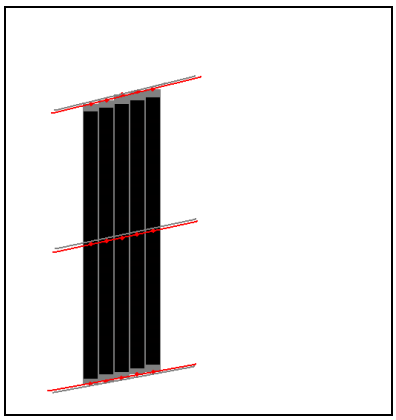


Figure 8. Fitting 2 lines on each group of points obtained from vertical candidate bands (totally 6 lines)

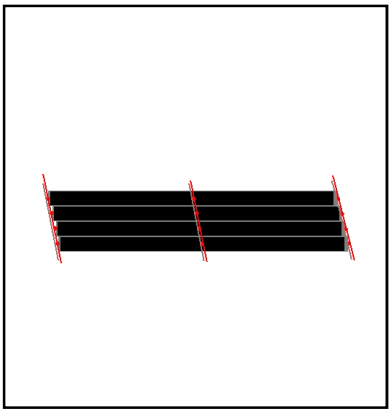


Figure 9. Fitting 2 lines on each group of points obtained from horizontal candidate bands (totally 6 lines)

III. EXPERIMENTAL RESULTS

A. Performance evaluation and results

To evaluate the proposed technique automatically, a similar evaluation strategy that was used in [2, 3] is utilized. The proposed technique generates two evaluation factors (average error and variance of errors) for each set of data. These factors are calculated as follows:

$$\text{Average Error} = \mu = \frac{1}{N}\sum_{i=1}^{N}(\beta_i^* - \beta_i)$$

$$\text{Variance} = \sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(\beta_i - \mu)^2$$

where β_i^* is the ground truth skew angle, β_i is the estimated skew angle by the proposed technique and N is the number of documents in the dataset.

To have an appropriate comparative analysis with the state-of-the-art methodologies particularly the methods presented in [2, 3], in addition to the above error rate factors, we also computed another two types of error rate by considering only 80% of the best results amongst all the results obtained from all the images (similar to paper [3]).

The proposed skew estimation technique is tested on 500 scanned documents [3] containing five categories (English, Chinese and Japanese, Figures, Tables and Multilingual documents). The documents were collected from newspapers, books, magazines, journals and scanned at 300dpi [3]. While testing the proposed algorithm with 500 documents [3], the average error of 0.1404 and a variance of 0.07 were obtained (Table IV). The results obtained from different categories of documents used in [3] are tabulated in Table III. From the experimental results, it can be noted that the proposed algorithm is capable of estimating skew angles of documents with different contents, languages, and layouts.

B. Analysis of errors

From the experiments, we noted that most of the errors occurred due to the obtaining insufficient number of horizontal/vertical candidate bands and consequently candidate points for line fitting from vertically/ horizontally

painted image. We noted that most of the error occurs when the number of candidate bands was less than 3. To get the idea of the distribution and frequencies of errors obtained during our experiments on 500 images, we noted that more than 90% of results obtained from the proposed technique using the dataset used in [3] have the error less than 0.5º.

TABLE III. RESULTS OF THE PROPOSED TECHNIQUE ON DIFFERENT CATEGORIES OF THE DATASET USED IN [3] AND IMAGES WITH DIFFERENT RESOLUTIONS

Dataset	All images		Top 80%	
	Average Error	Variance	Average Error	Variance
1st 100 images of [3]	0.1427	0.0830	0.0455	0.0040
2nd 100 images of [3]	0.1237	0.069	0.0170	0.0006
3rd 100 images of [3]	0.2680	0.0980	0.2014	0.0375
4th 100 images of [3]	0.0600	0.0155	0.0150	0.0004
5th 100 images of [3]	0.1075	0.0657	0.0240	0.0006

IV. COMPARATIVE ANALYSIS

To perform an evaluation with the state-of-the-art results in the literature, we compare the results of the present work with the results reported in two recently published papers [2, 3]. Since the results achieved by the e-PCP [2] and PCP [3] have outperformed the other techniques in the literatures, only these two techniques are considered for the comparative study. The same dataset (500 documents) used in [3] is utilized for the comparison of results. A detailed comparison of different methods is shown in Table IV. From Table IV, it is evident that the average error obtained from the proposed technique is better than the state-of-the-art methodologies. Moreover, the variance of errors calculated based on the results obtained by the proposed technique is much better than the techniques presented in [2, 3]. We obtained some errors during our experimentation. These errors occurred because of getting insufficient number of horizontal/vertical candidate bands and consequently candidate points for fitting different lines in some images. In overall, the results reported in the proposed technique outperformed the results of the recent skew estimation methodologies such as PCP and e-PCP.

V. CONCLUSION

In this paper, we presented a novel technique for skew estimation of scanned documents. The proposed method is applied to the entire document irrespective to the flow of writing and content. Moreover, rotation as an expensive operation has not been used in the proposed technique. We achieved improvement in both the accuracy and computing time without performing any additional procedure for finding flow of writing, rotation, or separation of textual and non-textual parts from the input documents. The efficacy of the proposed technique is proved by conducting an extensive set of experiments on three different datasets containing text, text with large-scale figures, text with large tables, multi-script and complex documents. Based on the results it can be noted that the performance of the proposed technique is better than the most recent works [2, 3] in the literature.

TABLE IV. COMPARATIVE OF RESULTS OF THE DIFFERENT TECHNIQUES ON DIFFERENT CATEGORIES OF THE DATASET USED IN [3]

Dataset	Algorithms	All images	
		Average Error	Variance
the first-category (100 images) of [3]	**e-PCP [2]**	Not reported	Not reported
	PCP [3]	0.1490	0.1290
	Proposed	**0.1427**	**0.0830**
the second-category (100 images) of [3]	**e-PCP [2]**	Not reported	Not reported
	PCP [3]	0.1390	0.1430
	Proposed	**0.1237**	**0.0690**
the third-category (100 images) of [3]	**e-PCP [2]**	0.4290	Not reported
	PCP [3]	**0.2310**	0.1350
	Proposed	0.2680	**0.0980**
the fourth-category (100 images) of [3]	**e-PCP [2]**	0.2546	Not reported
	PCP [3]	0.0770	0.0750
	Proposed	**0.0600**	**0.0155**
the fifth-category (100 images) of [3]	**e-PCP [2]**	0.1946	Not reported
	PCP [3]	0.1110	0.1270
	Proposed	**0.1075**	**0.0657**
All 500 images from [3]	**e-PCP [2]**	Not reported	Not reported
	PCP [3]	0.1414	0.1200
	Proposed	**0.1404**	**0.0700**

REFERENCES

[1] J.J. Hull, Document image skew detection: survey and annotated bibliography, in: J.J. Hull, S.L. Taylor (Eds.), Document Analysis Systems 2, World Scientific, Singapore, 1998, pp. 40–64.

[2] P. Dey, S. Noushath, e-PCP: A robust skew detection method for scanned document images, Pattern Recognition 43, 2010, 937-948.

[3] C.-H. Chou, S.-Y. Chu, F. Chang, Estimation of skew angles for scanned documents based on piecewise covering by parallelograms, Pattern Recognition 40, 2007, pp.443–455.

[4] G. Nicchiotti, C. Scagliola, Generalized projections: a tool for cursive handwriting normalization, Proc.of 5th ICDAR, pp. 729–732, 1999.

[5] S. Li, Q. Shen, J. Sun, Skew detection using wavelet decomposition and projection profile analysis, Pattern Recognition Letters 28, 2007, 555–562.

[6] X. Jiang, H. Bunke, D. Widmer-Kljajo, Skew detection of document images by focused nearest-neighbor clustering, Proceedings of the 5th ICDAR, pp. 629–632, 1999.

[7] U. Pal, B.B. Chaudhuri, An improved document skew angle estimation technique, Pattern Recognition Letters 17, pp. 899–904, 1996.

[8] C. Singh, N. Bhatia, A. Kaur, Hough transform based fast skew detection and accurate skew correction methods, Pattern Recognition 41, 2008, pp. 3528–3546.

[9] Y.K. Chen, J.F. Wang, Skew detection and reconstruction based on maximization of variance of transition-counts, Pattern Recognition 33, 2000, 195–208.

[10] T. Akiyama, N. Hagita, Automated entry system for printed documents, Pattern Recognition 23, 1990, pp.1141–1154.

[11] M. Chen, X. Ding, A robust skew detection algorithm for grayscale document image, Proc. of 5th ICDAR, pp. 617–620, 1999.

[12] P. Shivakumara, G. Hemantha Kumar, D. S. Guru, P. Nagabhushan, An Efficient Skew Estimation Technique for Binary Document Images Based on Boundary Growing and Linear Regression Analysis, Proceedings of the ICONIP, pp. 659-665, 2004.

[13] Alireza Alaei, Umapada Pal and P. Nagabhushan, "A New Scheme for Unconstrained Handwritten Text-line Segmentation", Pattern Recognition,Vol. 44, No. 4, 2011, pp. 917-928.