Two Phase Data Reduction for Hough Transformation to Detect Skew

Sunil Kumar
Research Scholar, Singhania University
Rajasthan,
India
Email: sunil.mathur10@gmail.com

S. Gopinath
ABB GISL, Corporate Research, Bangalore, India
Email: selgopi@gmail.com

***Abstract*—An important pre-processing task in an automated document image processing system is detection of skew in the image. Skew detection and correction is the first step in the document analysis and understanding processing steps. Correction the skewed scanned document image is very important, because it has a direct effect on the reliability and efficiency of the segmentation and feature extraction stages. It is desirable to have a fast, all-purpose skew detector. This paper proposes a novel two phase method. This two phase method reduces the skewed data before applying the skew detection and correction methods. A combination of the two phase method and the Hough Transformation can quickly and accurately detect skew on a wide variety of document images. Results of a test of the new algorithm and a comparison against the simple Hough transformation algorithm on 100 pages show the new algorithm to be much faster.**

Keywords-Skew detection; Hough Transformation; Two phase data reduction; Projection Profile; Least Squares method; and Principal Component analysis.

I. INTRODUCTION

Skew is the alignment of lines of text with respect to the horizontal normal and it can either be clockwise (positive) or anti-clockwise (negative).In machine printed scripts, skew usually occurs due to incorrect alignment of the page during the scanning process. On the other hand, skew in handwritten scripts can also be introduced by the writer. This causes difficulties in the segmentation and recognition processes if not corrected in an OCR system.

Skew and orientation detection plays an important role in document image preprocessing, especially in noise removal [1] and binarization [2]. This is because the successive processing techniques will fail to work correctly since usually both layout analysis [3] and OCR [4, 5] assume pages to be in the correct orientation.

The orientation and skew detection methods can be broadly classified into three parts, skew-only detection methods, orientation-only detection methods and combined skew and orientation detection methods.

Cattoni et al. [6] and Lu [7] discussed the various skew methods like projection profile, Hough transformation and connected components etc. For orientation detection, a method using the ascender to descender ratio was presented by Caprari [8] in 1999. An up/down orientation detection method for different scripts was presented by Aradhye [9] in 2005. In 2006, Lu et al. [10] present a method for language and orientation detection using the distributions of the number and position of white to black transitions of the components in the line. In 1994, Le et al. [11] presented a system capable of detecting portrait or landscape mode images and then the skew angle. In 1995 Bloomberg et al. [12] presented a method for orientation detection that uses the ratio of the number of ascender characters to the number of descender characters used in English. A more recent method using the ascender to descender ratio for orientation detection is presented by Avila et al. [13].

Hough transform was used by Srihari and Govindaraju V [15] for skew detection. In Hough transformation the Cartesian space (x, y) coordinates are mapped to a sinusoidal curves in (ρ, θ) space coordinates using the transformation

$$P = x \cos \theta + y \sin \theta \qquad (1)$$

Every time a sinusoidal curve intersects another particular value of ρ and θ, the likelihood increases that a line corresponding to that (ρ, θ) coordinates value is present in the original image. An accumulator array is used to count the number of intersect the various ρ and θ values. The skew is then determined by the θ values corresponding to the highest number of counts in the accumulator array. Le, Thoma and Wechsler [14] use bottom pixels of the candidate objects within a selected region for Hough transformation. The hierarchical Hough transformation technique is also adopted by Yu and Jain [16]. The main idea of the above methods is to reduce the amount of Input data, but their computational complexities are still very high. H.K.Chethan and G.Hemantha Kumar [17] presented a method for separating or removal of graphics like logos, animations other than the text from the document and finally textual content skew is corrected and characters are recognized using commercial OCR using horizontal and vertical projection and then Hough Transformation. Atallah Mahmoud Al-Shatnawi and Khairuddin Omar [18] discussed the challenges facing the Arabic skew detection and correction methods and proposed a method to detect and correct the skew using an arbitrary polygon and derivation of the baseline from polygon's centroid.

Chiu L. Yu, Yuan Y. Tang and Ching Y. Suen [19] presented an algorithm to detect skew in a totally unconstrained document. This algorithm can detect the skew angle complete page and also in different document blocks which contain different skew angles. Iuliu Konya Stefan Eickeler Christoph Seibert [20] introduces a generic, scale-independent algorithm which is capable of accurately detecting the global skew angle of any document images within the range [-90, 90] degrees and also for Roman script documents for a full range [-180,180) degrees of possible skew angles B.B. Chaudhuri and U. Pal [21] proposes a method to detect the Skew angle of scanned documents containing Indian scripts using the peculiar characteristics of these scripts known as headlines. Zhang Ruilin, Hu Yan, Fang Zhijian, Zhang Lei [22] presented a method which combine the characteristics of fabric images and the weft direction information extracted by Sobel operator detect the skew angle in scanned fabric image using Hough transform. Changyou Li, Quanfa Yang [23] discussed an algorithm to detect the skew angle of a scanned track image based on wavelet transform and linear least square fitting method.

The cross-correlation method proposed by [24] is based on the correlation between two vertical lines in a document image. As the pixels in the two parallel lines are translated due to skew, correlation matrix can be produced, it is defined as

$$R(x_o, s) = \sum_y L_1(x_0, y) L_2(x_0 + d, y + s) \qquad (2)$$

Here, L1(x, y) and L2(x, y) represent the two parallel vertical lines, respectively, d is the space between L1 and L2 and s is the maximum translation. Though this technique can result in accuracy, it is time consuming to calculate the correlation matrix and its projection profile. Furthermore, in certain situations parameter d should be changed and backtracked, which may increase the computing cost.

This paper proposes a novel two phase data reduction method. A combination of the two phase data reduction method and the Hough Transformation can quickly and accurately detect skew on a wide variety of document images. The proposed two phase data reduction method reduces the skewed data before applying the Hough Transformation for skew detection and correction. The approach used in this paper may be described by starting with the figure 1, which represents a skewed binary image. This image is applied to the two phase data reduction algorithm whose final output is shown in figure 3. This figure shows that the remaining pixels are almost all oriented in the same direction. After the two phase reduction, the skew angle of the reduced image is detected by using the Hough transform.

II. Skew, Its Reasons and Detection Approaches

The conversion of paper documents to electronic format is required for many computerized processing's like document archiving, automated document delivery, journal distribution, record management, etc. The stages of document conversion include scanning of the document or capturing the document image with the help of a hand held device like camera or mobile, displaying, image processing, text recognition, image and text database creation and quality assurance. During the image capturing or scanning process, the whole document or a portion of it may be misaligned causing skew.

Human vision can easily detected considerable skew in document and the skew correction can be made by re-scanning the document or re-capturing the image. The human vision system fails to detect the modest skew. Even a smallest skew angle existing in a given document image results in the failure of segmentation of complete characters from words or a text lines, as the distance between the character reduces. Additionally most of the OCRs and document retrieval/display systems are very sensitive to skew in document images.

It is important to detect and correct skewness. In the past, many approaches were proposed for skew angle detection of document images and their correction. Main approaches for skew detection include Hough transform, Projection Profile, nearest neighbor, Principal Component Analysis, Least Squares method, and EM Based Algorithm. All of these approaches require a dominant text area to be present in order to work properly.

The most popular method for skew detection is by using the Hough transform. It is capable of locating disjointed lines in a binary image. Therefore given a group of black pixels, one can find the imaginary line or lines that go through the maximum number of these pixels. Given a binary image with a dominant text area, the detected lines will most probably go along the whole middle zone of the textual lines. Hence these lines have approximately the same skew as the reference lines of the text which the skew of the whole page. Whenever the

Hough transform is used, there is always a tradeoff between accuracy and speed. The more accurate the angles of the detected lines are, the more computation is required. In addition the computation time depends on the number of pixels in the image. Connected component analysis is required. Then several iterations take place in order to connect each component to its nearest neighbor in recursion. This process results in several chains representing the textual lines. A line that goes through the central mass of the components in a chain approximates the skew of the associated textual line.

Alternatively, one can calculate the skew by averaging the skews of the lines that connect neighboring components in a chain. This method is based on the fact that inter character and inter word spaces between two consecutive characters or words respectively, are usually smaller than spaces between such neighboring elements that belong to different lines.

The simplest case of Hough transform is the linear transform for detecting straight lines. In the image space, the straight line can be described as y = mx + b and can be graphically plotted for each pair of image points (x,y). In the Hough transform, the main idea is to consider the characteristics of the straight line not as image points x or y, but in terms of its parameters, here the slope parameter m and the intercept parameter b. Based on that fact, the straight line y = mx + b can be represented as a point (b, m) in the parameter space. However, one faces the problem that vertical lines give rise to unbounded values of the parameters m and b. For computational reasons, it is therefore better to parameterize the lines in the Hough transform with two other parameters, commonly referred to as ρ (rho) and θ (theta). The parameter ρ represents the distance between the line and the origin, while θ is the angle of the vector from the origin to this closest point

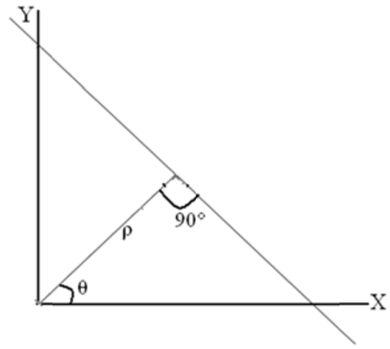

Figure 1: XY plane and ρ, θ plane for line detection

The Hough transform accepts the input in the form of a binary edge map and tries to find edges positioned as straight lines. The idea of the Hough transform is that every edge point in the edge map is transformed to all possible lines that could pass through that point. The line detection in a binary image using the Hough transform algorithm can be summarized as follows:

1. Define the Hough transform parameters ρmin, ρmax, θmin and θmax.
2. Quantize the (ρ,θ) plane into cells by forming an accumulator cell array A (ρ,θ), where ρ is between ρmin and ρmax, and θ is between θmin and θmax.
3. Initialize each element of an accumulator cell array A to zero.
4. For each black pixel in a binary image, perform the following:
    a. For each value of θi from min to max, calculate the corresponding ρi using the equation: $x \cdot \cos\theta_i + y \cdot \sin\theta_i = \rho_i$
    b. Round off the ρi value to the nearest allowed ρ value.
    c. Increment the accumulator array element A (ρi,θi).
5. Finally, local maxima in the accumulator cell array correspond to a number of points lying in a corresponding line in the binary image.

A. Computational Performance of Hough Transform

The computational performance of the Hough transform algorithm can be determined by analyzing the step (4) of the algorithm. In Hough Transformation method the time complexity can be computed as T(n× θi), where n is number of points and θi is number of different values of angles. If N is the number of black pixels in a binary image and K is the number of increments of θi from θmin to θmax, number of computations in the Hough transform is 3NK To achieve higher accuracy we have to take smaller angle intervals so higher numbers of different values for angle are to be computed and hence more running time is required.

Accuracy Hough Transform heavily depends upon how much small is step size for distance parameter ρ. suppose ρ can take values up to10 and we divide it in ten intervals step size of one. Then there can be 11 values for ρ from 0, 1,….up to 10 and space requirement for Hough Matrix H will be A ×11 and if we take step size of 0.5 for ρ , then there can be 21 different values for it and space requirement for Hough Matrix H will be A ×21. Hence space requirement depends upon accuracy desired.

Smallest the step size for ρ, more the accuracy will and more the space for Hough Matrix is required.

As we have discussed above the running cost in Hough Transform is T (n× θi), where n is number of points and θi is number of different values of angles. We have to use more number of angle values to achieve more accuracy. Using finer angle intervals means we require more running time. Also the complexity of the Hough Transform is directly proportional to number of foreground pixels and hence Hough transform is rarely performed directly on input image. So it is always desirous that effort should be put to reduce the number of input pixels. Due to the above facts input image is always preprocessed before applying Hough Transform.

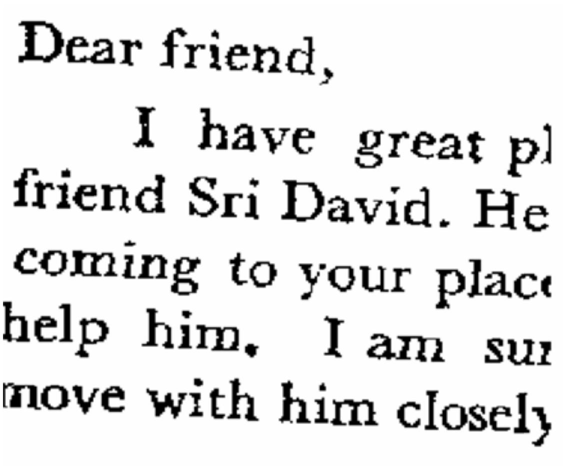

Figure 1: Original Image

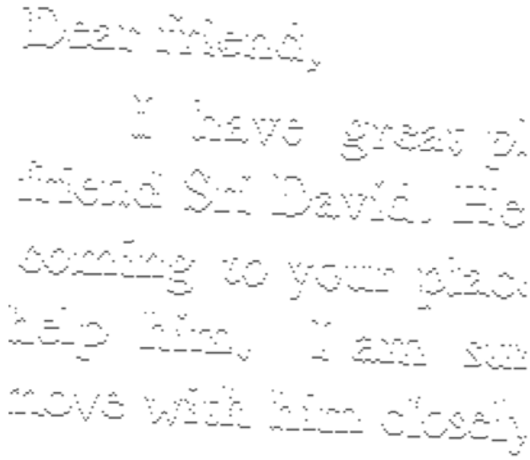

Figure 2: Binary Image after the First Phase

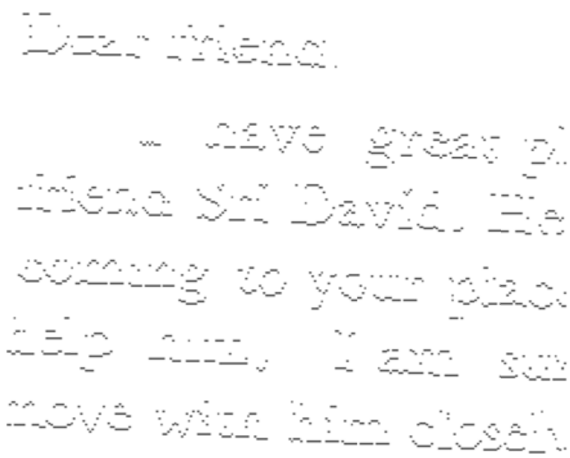

Figure 3: Final Reduced Image after the second phase

III. Two Phase Data Reduction

The preprocessing applied before the Hough Transform is very important and decides how pixels are filters from image document. There have been many algorithms those uses Hough Transform for skew detection. They mainly differ in aspect that how they do preprocessing before applying Hough Transform. For example Amin [25] uses connected component analysis followed by grouping of connected components. Connected components are bounded by rectangular boundary representing connected black pixels Then Hough Transform is applied to resulting document image. M. Aradhya [26] also discusses the skew detection technique based on Hough Transform. It uses Thinning as preprocessing step. Thinning is process of reducing thickness of each line of pattern to just single pixel. This algorithm is iterative and it deletes every point on that lies on outer boundary of symbol. For the purpose of pixel reduction, a new technique is being proposed in this paper. The technique is known as ‘Two Phase Data Reduction’. It takes the original image and reduces it in two phases.

The purpose of this proposed Two Phase Data Reduction technique is to reduce the black pixels while keeping the features of the image intact than extract features from the binary image which are used to detect the document skew angle using the Hough Transformation in a much faster way.

In the first phase, each column of the image is scanned. For each black pixel, the next immediate pixel, i.e. the pixel in the same column but next row, is checked. If the next immediate pixel is white, then this black pixel is retained in the new binary image ‘A1’ else this black pixel is ignored. This image ‘A1’ is the intermediate image obtained at the end of the first phase.

The figure 1, shows the original image which is to be reduced, so that the skew angle can be calculated by Hough Transform in a faster manner. The figure 2 shows the intermediate image at the end of the first image.

In the second phase, the intermediate image ‘A1’ is further reduced. For each row of the image ‘A1’, a count of the black pixels in each row is found out. Based on this count, the mean number of black pixels in a row of the image ‘A1’ is calculated. For each row of the image, check whether the row has a black pixel count greater, equal or lesser than the mean value. If this row has a black pixel count greater than the mean value, then this row is retained in the new binary image ‘A2’, else ignore this row for the new binary image ‘A2’. ‘A2’ is the final, reduced image obtained at the end of the second phase.

The figure 3 shows the reduced image obtained after processing the original image of figure 1.

The process and algorithm of the Two Phase Data Reduction can be recapitulated as below:

1. For each column of the binary image
    a. If the pixel is black, check whether the immediate next pixel, i.e. the pixel in the next row within the same column, is white.
    b. If the immediate pixel is white, then retain this black pixel in the new image ‘a1’ else ignore this black pixel.
2. The image ‘a1’ is the intermediate image, obtained at the end of the first phase.
3. For each row of the image ‘a1’, count the number of black pixels in each row. Based on this count, calculate the mean number of pixels in a row.
4. For each row of the image ‘a1’
    a. If this row has a black pixel count greater than the mean value, then retain this row in the new image ‘a2’.
    b. If the row has a black pixel count less than or equal to the mean value, then ignore this row for the new image ‘a2’.
5. The image ‘a2’ is the final reduced image.

A. Skew Angle Detection Algorithm using Two Phase Data Reduction

The algorithm creates a simplified binary image from the original binary image and applies the Hough transform on the simplified binary image to get the document skew angle. The document skew angle detection algorithm can be summarized as follows:

a. Obtain a simplified binary image by ‘Two Phase Data Reduction’.
b. Apply the Hough Transform on the image.
c. Calculate the peaks of the Hough Transform.
d. Identify lines in the image.
e. Identify the angles of the lines.
f. Calculate the average skew.

IV. RESULTS:

It is well known that the Hough Transform is computationally expensive. The technique of ‘Two Phase Data Reduction’ was tested on a set of 21 images and the reduced images obtained were then used for skew detection using the Hough Transform. These 21 images are of different image formats. Three of the 21 test images are shown in figure 4, figure 5, figure 6 and figure 7. The three test images shown in these figures are of .png, .jpg and .bmp formats respectively.

These images also include handwritten as well as typed text. So this two phase data reduction method is applicable in wide varieties of images. The skew of these two test images are then corrected using the Hough Transformation as shown in figure 8.

A comparison between the Hough Transform and Hough Transform using Data Reduction has been made on the basis of accuracy and speed. The code was run on MATLAB 7.0 and on a standard Pentium IV processor. These results are tabled in table 1.

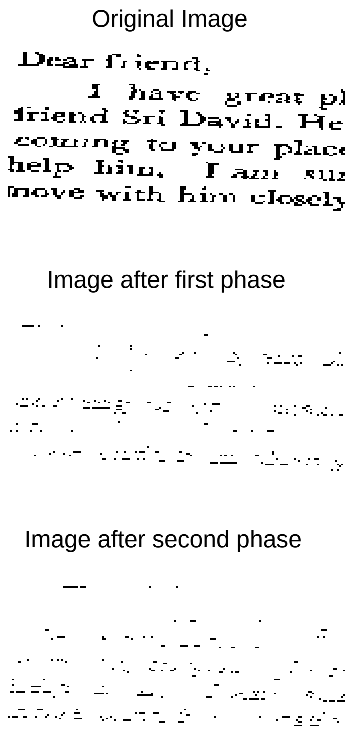

Figure 4: Two Phase Data Reduction

applied on Test Image 1

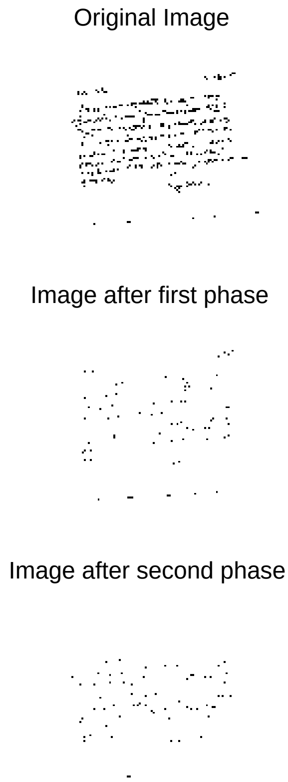

Figure 5: Two Phase Data Reduction

applied on Test Image 2

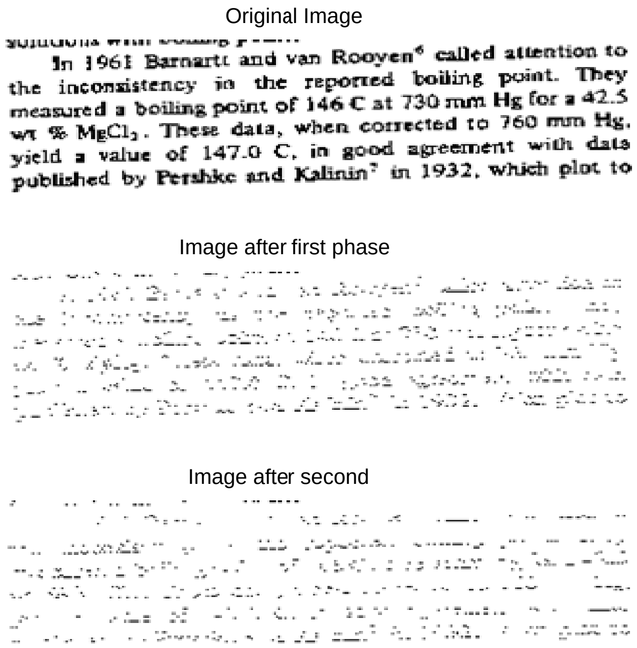

Figure 6: Two Phase Data Reduction applied on Test Image 3

Figure 7: Two Phase Data Reduction applied on Test Image 3

Figure 8: Skew correction after Two Phase Data Reduction applied on Test Image 3

TABLE . TABULATED RESULTS SHOWING THE REDUCTION IN COMPUTATIONAL TIME AFTER APPLYING THE TWO PHASE DATA REDUCTION TECHNIQUE

Image	*Image Size*	*Total Black Pixels*	*Black Pixels (after reduction)*	*Skew (in degrees)*	*Skew (with data reduction)*	*Time (in seconds)*	*Time (with data reduction)*	*% Reduction in time*
Test Image 1.png	708 x 414	33878	4416	2	2	1.57	0.93	40.76
Test Image 2.jpg	1734 x 1397	86706	6712	-6	-3	5.62	4.24	24.55
Test Image 3.bmp	908 x 579	36569	1765	3	3	2.12	1.08	49.05
Test Image 4.bmp	884 x 552	36743	1966	-7	-6	2.04	0.98	51.96
Test Image 5.jpg	720 x 133	3455	720	-1	-1	0.65	0.54	16.92
Test Image 6.jpg	596 x 720	24213	4696	2	2	1.29	0.81	37.20
Test Image 7.jpg	405 x 410	15283	2119	0	0	1.32	0.64	51.51
Test Image 8.jpg	610 x 198	21830	1500	-2	-2	1.45	0.57	60.68
Test Image 9.jpg	266 x 76	2326	564	0	0	0.56	0.39	30.35
Test Image 10.jpg	894 x 442	58613	3168	0	0	2.65	0.87	67.16
Test Image 11.jpg	456 x 357	7244	1130	-4	-3	0.79	0.55	30.37
Test Image 12.jpg	523 x 350	20487	3476	2	2	1.50	0.68	54.66
Test Image 13.jpg	1696 x 347	27793	3079	-1	-1	1.70	1.41	17.05
Test Image 14.png	404 x 226	14776	4192	-1	-1	1.26	0.65	48.41
Test Image 15.png	401 x 226	14781	4007	2	2	1.21	0.63	47.93
Test Image 16.jpg	633 x 649	55277	6676	-1	-1	3.04	1.23	59.53
Test Image 17.jpg	457 x 189	5741	764	-2	-2	0.68	0.57	16.17
Test Image 18.jpg	443 x 185	9947	1808	1	1	0.85	0.52	38.82
Test Image 19.jpg	608 x 291	13429	2376	1	1	0.92	0.63	31.52
Test Image 20.jpg	503 x 269	4082	386	-8	-7	0.67	0.54	19.40
Test Image 21.jpg	301 x 225	7676	1220	-2	-2	1.01	0.49	51.48

V. CONCLUSION

The Hough Transform is a robust method for skew detection. However, it is computationally not so efficient. Thus, speed up algorithms can be used for it. For speed up purposes, we have used the ‘Two Phase Data Reduction’ technique. This technique helps in reducing the computational time required for Hough Transform considerably with less compromise on the accuracy front.

The Two Phase Data reduction is also be applicable for 'Principal Component Analysis’. But there would be a trade off on the accuracy as this method depends on the no. of black pixels for better accuracy.

The technique of Projection Histogram is found to be less accurate but much faster than the Hough Transform. The Hough Transform is better suited to handwritten scripts than Projection Histogram.

When this data reduction technique is applied to improve the speed of skew detection and correction with Projection Histogram, Least Squares method, and EM Based Algorithm we find that it does not improve these methods.

REFERENCES

[1] F. Shafait, J. van Beusekom, D. Keysers, T. M. Breuel, “Document cleanup using page frame detection” Int. Jour. on Document Analysis and Recognition 11 (2) (2008) pp. 819-823..

[2] F. Shafait, D. Keysers, T. M. Breuel, “Efficient implementation of local adaptive thresholding techniques using integral images” Proc. of SPIE Document Recognition and Retrieval XV [22], pp. 681510.

[3] F. Shafait, D. Keysers, T. M. Breuel, “Performance evaluation and benchmarking of six page segmentation algorithms” IEEE Trans. on Pattern Analysis and Machine Intelligence 30 (6) (2008) pp. 941-954.

[4] R. Smith “An overview of the Tesseract OCR engine” Proc. of the 9th Int. Conf. on Document Analysis and Recognition [21], pp. 629-633.

[5] T. M. Breuel, “The OCRopus open source OCR system” Proc. of SPIE Document Recognition and Retrieval XV [22], pp. 0F1-0F15.

[6] R. Cattoni, T. Coianiz, S. Messelodi, C. M. Modena, “Geo-metric layout analysis techniques for document image understanding: a review” Tech. Rep. 9703-09, IRST, Trento, Italy (1998).

[7] Y. Lu, C. Tan, “A nearest-neighbor chain based approach to skew estimation in document images” Pattern Recognition Letters 24 (14) (2003) 2315{2323.

[8] R. S. Caprari, “Algorithm for text page up/down orientation determination” Pattern Recognition Letters 21 (4) (2001), pp.311-317.

[9] H. Aradhye "A generic method for determining up/down orientation of text in Roman and non-Roman scripts" Pattern Recognition 38 (11) (2005) 2114-2131.

[10] S. Lu, C. L. Tan "Automatic document orientation detection and categorization through document vectorization" Proc .of the 14th ACM Int. Conf. on Multimedia, New York, NY, USA, 2006, pp. 113-116.

[11] D. Le, G. Thoma, H. Wechsler "Automated page orientation and skew angle detection for binary document images" Pattern Recognition 27 (10) (1994) 1325-1344.

[12] D. S. Bloomberg, G. E. Kopec, L. Dasari, "Measuring document image skew and orientation" Proc. of SPIE Document Recognition and Retrieval II, San Jose, CA, USA, 1995, pp. 302-316.

[13] B. T. Avila, R. D. Lins "A fast orientation and skew detection algorithm for monochromatic document images" Proc. of the 5th ACM symposium on Document engineering, New York, NY, USA, 2005, pp. 118-126.

[14] Le D S, Thoma G R and Wechsler H, "Automatic page orientation and skew angle detection for binary document images" Pattern Recognition 27, 1994, pp 1325-1344.

[15] Srihari SN and Govindaraju V "Analysis of textual images using the Hough Transform" Machine Vision and Applications, vol 2, 1989, pp. 141-153.

[16] Yu, B., Jain, A.K. "A robust and fast skew detection algorithm for generic documents" Pattern Recognition 29 (10), pp 1599-1629, 1996.

[17] H.K.Chethan and G.Hemantha Kumar "Graphics Separation and Skew Correction for Mobile Captured Documents and Comparative analysis with Existing Methods" International Journal of Computer Applications (0975 – 8887),Volume 7– No.3, September 2010

[18] Atallah Mahmoud Al-Shatnawi and Khairuddin Omar "Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity" Journal of Computer Science 5 (5): 363-368, 2009, ISSN 1549-3636 © 2009 Science Publications

[19] Chiu L. Yu, Yuan Y. Tang and Ching Y. Suen "Document Skew Detection Based on the Fractal and Least Squares Method" 0-8186-7128-9/9$54 .00 0 1995 IEEE

[20] Iuliu Konya Stefan Eickeler Christoph Seibert "Fast Seamless Skew and Orientation Detection in Document Images" 2010 International Conference on Pattern Recognition 1051-4651/10 $26.00 © 2010 IEEE DOI 10.1109/ICPR.2010.474

[21] B.B. Chaudhuri and U. Pal " Skew Angle Detection of Digitized Indian Script Documents" IEEE Transactions On Pattern Analysis And Machine Intelligence, Vol. 19, No. 2, February 1997

[22] Zhang ruilin, Hu yan, Fang zhijian, Zhang lei "Skew Detection and Correction Method of Fabric Images Based on Hough Transform" 2009 Second International Conference on Intelligent Computation Technology and Automation978-1-4244-3608-8/09/$25.00 © 2009 IEEE

[23] Changyou Li, Quanfa Yang "Skew Detection of Track Images Based on Wavelet Transform and Linear Least Square Fitting" Proceedings of the 2009 IEEE International Conference on Information and Automation June 22 -25, 2009, Zhuhai/Macau, China.

[24] Yan, H. Skew correction of document images using interline crosscorrelation, Computer Vision, Graphics, and Image Processing 55, 1993, pp 538-543.

[25] A Amin and S. Fischer, A Document Skew Detection Method Using the Hough Transform, Pattern Analysis & Applications (2000)3:243-253

[26] Manjunath Aradhya V N, Hemantha Kumar G, and Shivakumara P, Skew Detection Technique for Binary Document Images based on Hough Transform, International Journal of Information Technology Volume 3 Number 3