International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 14 (2015) pp 34894-34898

A Novel Method for Document Skew Detection and Correction : Application to, Handwritten Document and Bank Documents

Prabhanjan Soukar
Research Scholar
School of Engineering & Technology
Jain University
Ramanagar Dist
prabhan_us@rediffmail.com

Ramegowda Dinesh
Research Supervisor
School of Engineering & Technology
Jain University
Ramanagar Dist
dr.dineshr@gmail.com

Santosh Naik
Research Scholar
School of Engineering & Technology
Jain University
Ramanagar Dist
santoshrcnaik@gmail.com

Abstract- Physical document is scanned for digitization, scanned document inevitably skewed due to physical and manual limitation. Presence of skew in the scanned document affects the accuracy of the subsequent steps of document layout analysis and Optical Character Recognition (OCR) systems. Hence it is highly desirable to detect and correct the skew prior to Document layout analysis and OCR. In this paper, a novel skew detection and correction method based online fitting to the text lines in the scanned document is proposed. Morphological operations with suitable structuring element are used to analyze the text line to form a line structures. These line structures are further converted into thin lines using thinning operations. Lines are fitted to this thinned line segment using regression analysis. The line parameters obtained by the regression lines, and the average slope of all line segments will serve as a skew angle of the document. Further, the document image is rotated by the computed skew angle to correct the skew. The proposed method has been evaluated on a large number of images from handwritten Devanagari scripts images and bank application forms.

Keywords: Connected Component, Document analysis, Linear Regression, Morphological Operations ,OCR, Skew correction, Skew detection, Skew angle.

Introduction

Digitization of paper document is routinely performed in document layout analysis and OCR. They have various practical applications in the areas of office automation, library automation, bank services, postal services and publishing houses. Accuracy of document layout analysis and OCR depends on the preprocessing and skew free document. Skew refers to the text which is not aligned correctly to a specified or implied line. The first is in document scanning; skew is unavoidably introduced into the incoming document image. This may be due to several reasons like placing the document on the scanner without aligning the page to the edges of the platen or may be because of a malfunctioning platen. The second is, skewed handwritten lines & words in the original document. If this skew is not detected and corrected, it will have a negative impact on the recognition accuracy of the analysis and OCR systems. The main steps in the document layout analysis are scanning, binarization, region segmentation, text recognition and document analysis. During the scanning process, the document may not be fed properly into the scanner, thus the text lines in the document images would be skewed and may cause problem in segmenting the document image to extract its layout structure. For example, in the most commonly used methods of profiling, in the presence of skew there would not be valleys in the projection and segmentation would not be possible. Skew detection and correction is, thus, a very important stage in document analysis (Amin & Fischer 2003). Optical character recognition is the process of translating the text material in the scanned document into machine readable codes. If skews in the scanned documents are not detected accurately, it will reduce the accuracy of OCR systems. A small degree of skew in the text document, results in the failure of segmentation of text into lines, lines into words and words into characters. Hence we need a method to detect and correct skew, introduced in the scanned document.

Related Work

There are several techniques for skew detection and correction. These techniques are based on projection profile, Hough transforms, Fourier transforms, nearest Neighbor clustering, and interline cross correlation. In the projection profile technique, projection profiles are obtained along a number of axes and variation is calculated for each of the profiles. The profile which gives maximum variation corresponds to the projection is the actual skew angle of the document Baird (1987) proposes this method and states that the skew angle should be limited to $\pm 15^{\circ}$ to achieve high accuracy. The method proposed is time consuming and its accuracy reduces when the documents contain noisy and containing character fragments. Srihari & Govindaraju (1989), al & Chaudhuri (1996) have proposed skew detection and correction techniques based on the Hough transform (HT). In this technique each point in the original (x, y) plane map to all points in (ρ , θ) Hough space of lines through (x, y) with slope θ and distance ρ from the origin where $\rho = x \text{ Cos } \theta + y \text{ Sin } \theta$ for $0 \leq \theta < \Pi$. The peak value of the Hough space represents the dominant line has been used for skew detection. The draw back this method is computationally expensive and when text becomes sparse it is difficult to choose a peak in the Hough space. Postal (1986) proposed a method based on Fourier transform (FT). In this method Skew angle is detected from the direction for which the density of the Fourier space is the largest. Fourier method is computationally expensive for large images. Hashizume et al (1986) proposed a method which estimates the skew based on

nearest neighbor clustering (NNC). In this method, histogram peak gives the skew angle, which is obtained by accumulating direction vector of all nearest neighbors. Noisy sub parts of characters in a connection would reduce the accuracy of this method. Yan (1993) introduced a method for detecting the skew angle of an image using correlation between two vertical lines. It is observed that the correlation between vertical lines in an image is maximized for a skewed document, if one line is shifted relative to the other lines such that the character baseline levels for the two lines are coincident. The proposed method is less expensive and computationally expensive. Gates et al (1997) has proposed a new method to find the skew in the document based on the information existing on a set of equidistant vertical lines. Further, the method considers the entire image Estimation of skew in document image 71 pixels that lie in a set of equidistant vertical lines. Correlation matrix obtained between vertical lines by using these pixels. Lilies et al (2002) has proposed a generic skew detection method for any type of preprinted form which is based on power spectral density (PSD) of the form horizontal projection profile. However, the method is accurate for only small amounts of skew of documents. Tian Jipeng (2011) propose a novel method for skew detection and correction using Hough Transform. In this method the voting procedure used to detect the straight-line. This method can detect the skew angle up to -90 to +90 degrees. Mandip Kaur (2013) has introduced the integrated skew detection and correction using Fourier transform. In this method DCT compression and thresholding technique applied on image to obtain Fourier spectrum. Then spectrum is split into four equal parts and the detected skew angle of all parts is measured. At last the input image is rotated using the Bilinear Interpolation method. From Literature reveals that Hough transform is the most popular technique used in detecting skew angle, but it is computationally expensive. To reduce the computing cost modified methods were proposed. Most of the proposed method will detect skew for the text document, but not be applied to layout analysis. Hence there is a need for research to develop a method used for detection of skew for text documents, for layout analysis and for local skew with high accuracy & less computational cost.

Proposed System

The proposed approach for skew detection is based on fitting a line to the connected components in the document image. Datasets used in our method are unconstrained, constrained, printed Devanagari script, skewed words and scanned bank forms. The overall system architecture of the proposed method is shown in Fig.1.

Binarization

Scanned document may be colored or grayscale image that needs to be converted to bi-level information so that it reduces the computational load and enables the utilization of simplified analysis methods. Understanding document image requires logical and semantic content preservation during thresholding. Documents considered in our experiments are even colored paper, hence global thresholding algorithm Otsu'( ) was used for binarization. An example of binarization is shown in Fig. 2(b).

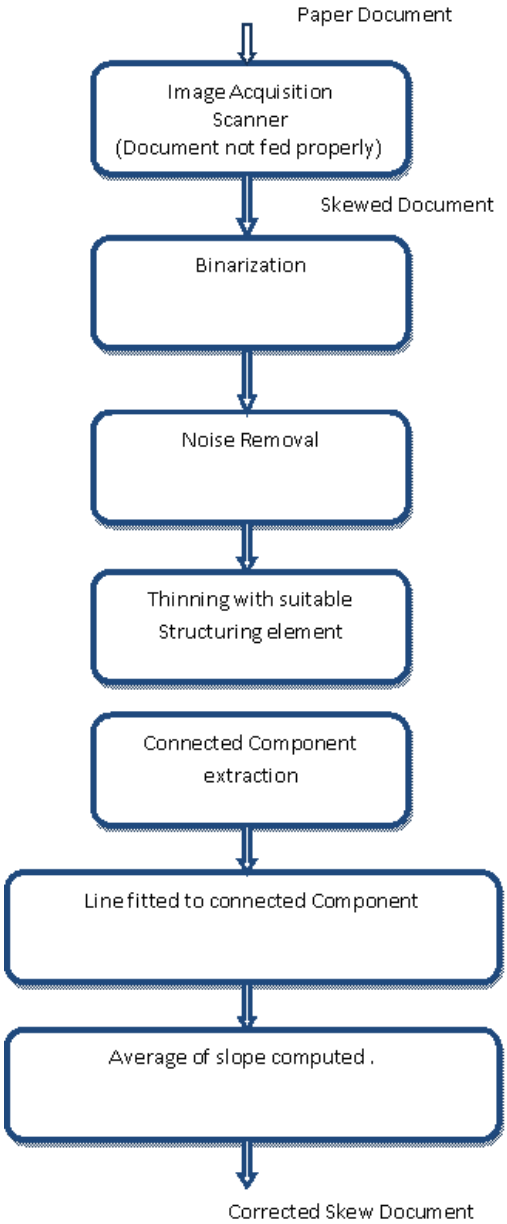
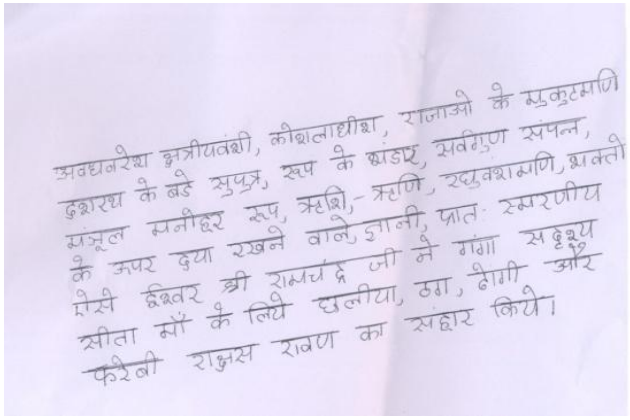


Fig.1. System Architecture

Noise Removal

Image acquisition processes, which converts optical image into continuous electronical signal using imaging sensors. Image sensors having electro mechanical limitations and hence it is impossible to eliminate the noise at the source. Hence an effective mechanism required to handle the noise at preprocessing level. The salt and pepper noise is most common type of noise occur in the image as isolated dots. These isolated dots need to be removed. In this work we have used various image filtering techniques to remove the noise. Noises are removed using median filtering operation and to eliminate the small isololated components morphological opening and closing operations are performed.



(a)

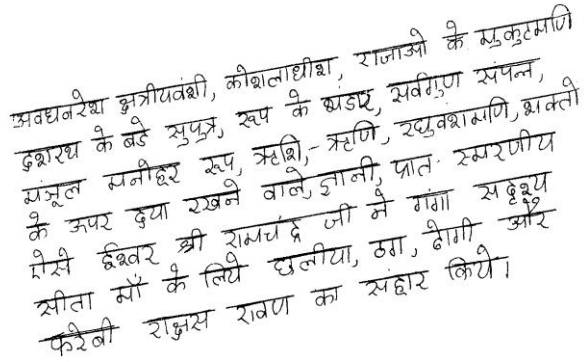


Fig.2. (a) Original Image (b) Binarized Image

Thinning

To estimate the skew of a document, line need to be fit to each skewed line. For this purpose, in this work, we have performed the morphological operation close using a structuring element whose width is equal to width of word and then thinned using algorithm[14] until each line of text into thin line using suitable structuring element depending on the type of document. This smears the image horizontally and one connected component produced for each line of text Fig. 3 shows the thinned document.

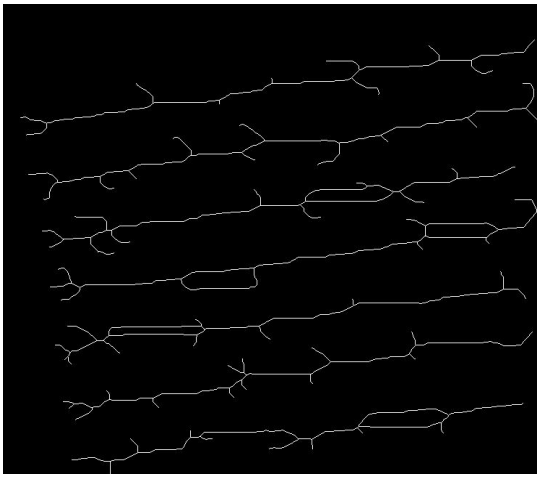


Fig.3. Thinned document

Extraction of connected component

Connected components representing each line are extracted using algorithm[15]. Lines are fitted to each connected component whose length is greater than the predefined threshold using least square method.

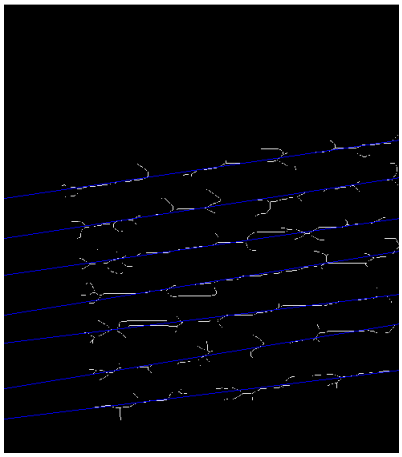


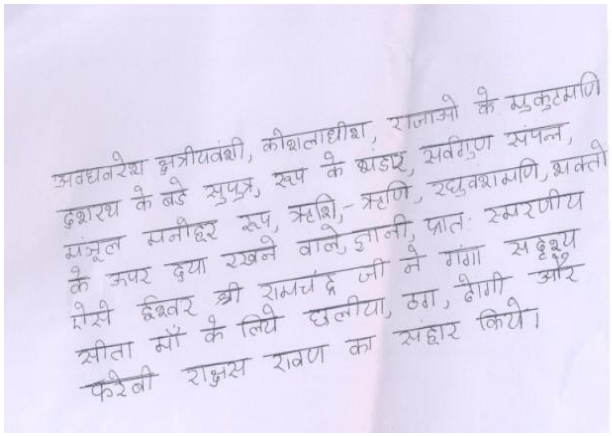
Fig.4. Lines fitted to connected component

Skew Angle Detection and correction

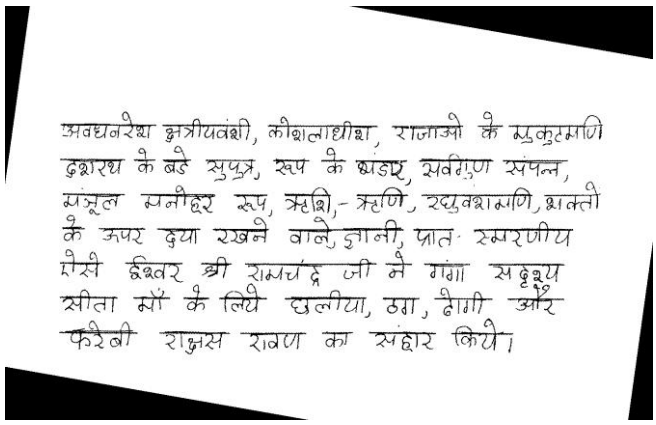
Average slope of the lines fitted are calculated which gives the angle to be corrected. Rotate the image by angle found using bilinear interpolation method.

Experimental Results

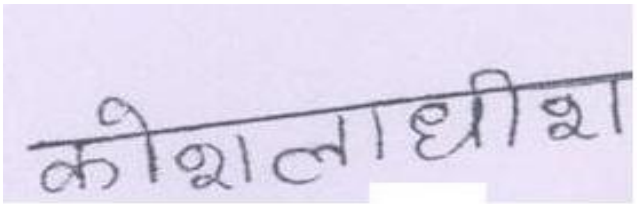
In this section, we present the detailed experimental analysis of the proposed approach on two datasets, derived from the unconstrained handwritten Devanagari script and bank forms collected from various banks. Overall, the number of images considered for this experimentation is 25 images. We have developed the algorithms using Matlab. Fig. 5. shows the result of document skew detection and correction of the proposed system. Table 1 shows the evaluation of existing method.



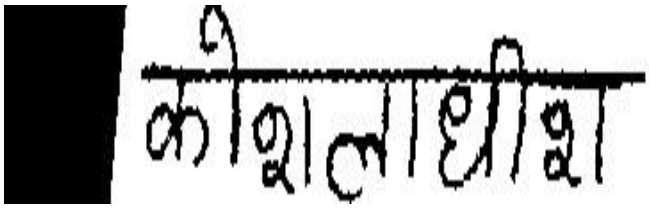
(a)



(b)



(c)



(d)

(e)

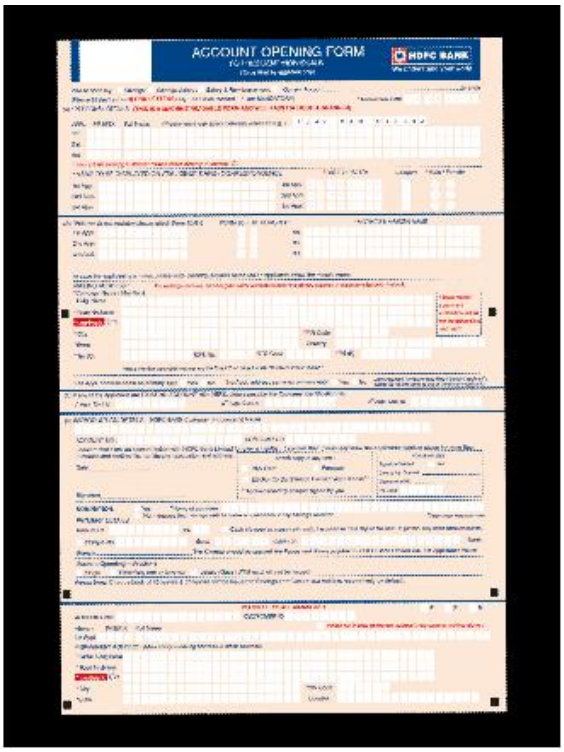


Fig. 5. (a) Original Document (b) Deskewed Document (c) Skewed handwritten word (d) Deskewed Handwritten Word (e) skewed bank form (f) deskewed bank form

Table .1 Evaluation of existing method

Document Type	Actual Angle	Detected Angle
Handwritten	5	5.12
	10	10.02
	15	14.98
	20	20.11
	25	25.13
Printed	5	4.98
	10	10.11
	15	15.07
	20	20.12
	25	25.1
Handwritten word	5	5.01
	10	10.05
	15	14.1
	20	20.01
	25	25.08
Bank Form	5	5.1
	10	10.1
	15	15.01
	20	19.05
	25	25.09

Conclusion

In this paper, we have proposed a novel approach for skew detection and correction of skewed documents. The proposed approach employed morphological operation to identify different skewed lines. Subsequently, to identify the slopes of these lines, connected component whose length is greater than the threshold are fitted with a line using least square analysis. Average slope of these fitted lines will be the skew angle of the document. Further, the derived skew angle is used to correct the skew of the document. The extensive experimentation reveals that the proposed method can be used to detect and correct skew up to ±25° for printed text, handwritten text of any language and printed form using the suitable structuring elements.

References

[01] MandipKaur, Simpel Jindal 2013.“ An Integrated Skew Detection And Correction Using Fast Fourier Transform And DCT”, International journal of science & technology research volume 2, issue 12, December ISSN 2277-8616

[02] Amin A, Fischer S 2003,”A document skew detection method using the Hough Transform”, Pattern analysis and applications (London: Springer-Verlag) pp 243–253

[03] Baird H S 1987,”The skew angle of printed documents”,. Proc. Soc. Photogr. Sci. Eng. 40: 21–24

[04] Srihari S N, Govindaraju V 1989 Analysis of textual images using the Hough Transform. Machine Vision Appl. 2: 141–153

[05] Pal U, Choudhuri B B 1996,” An improved document skew angle estimation technique”. Pattern Recogn. Lett. 17: 899–904

[06] Hashizume Yeh, P S, Rasenfeld A (1986),”Method of detecting the orientation of aligned components. Pattern Recogn”. Lett. 4: 125–132

[07] Yan H 1993,”Skew correction of document images using interline cross-correlation. Comput. Vision, Graphics Image Process”. 55: 538–543

[08] Gates B, Papamarkos N, Chamzas C 1997,”Skew detection and text line position determination in digitized documents”. Pattern Recogn. 30: 1505–1519

[09] Tian Jipeng, G. Hemantha Kumar and H.K. Chethan 2011, "Skewcorrection for Chinese character using Hough Transform”, International Journal of Advanced Computer Science and Applications, Specital Issue on Image Processing and Analysis, Vol 22,pp. 33-36,May

[10] Mandip Kaur and Simpel Jindal 2013,“ An Integrated Skew Detection and Correction Using Fast Fourier Transform and DCT", *International Journal of scientific & technical research*, vol. 2, pp. 164-171, Dec.

[11] Prakash K Aithal, Rajesh G and U Dinesh Acharya 2013, "A Fast and Novel Skew Estimation Approach uses Radon Transform", International Journal of Computer Information Systems and Industrial Management Applications, Vol. 5, pp. 337-344, 2013.

[12] PostlW 1986, “Detection of linear oblique structure and skew scan in digitized documents”, In Proc. Int.Conf. on Pattern Recognition, pp. 687-689.

[13] Liolios N, Fakotakis N, Kokkinakis G 2002 On the generalization of the form identification and skew detection problem. Pattern Recogn. 35: 253–264

[14] Lam, L., Seong-Whan Lee, and Ching Y. Suen, "Thinning Methodologies-A Comprehensive Survey," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol 14, No. 9, September 1992, page 879, bottom of first column through top of second column.

[15] Haralick, Robert M., and Linda G. Shapiro, Computer and Robot Vision, Volume I, Addison-Wesley, 1992, pp. 28-48.