DATABASE DEVELOPMENT OF HISTORICAL DOCUMENTS: SKEW DETECTION AND CORRECTION

S P Sachin[1], Banumathi K L[2], Vanitha R[3]

[1]*UG, Student of Department of ECE, BIET, Davangere, (India)*

[2,3]*Assistant Professor, Department of ECE, BIET, Davangere, (India)*

ABSTRACT

During document scanning, skew is inevitably introduced into the incoming document image. Skew detection is one of the first operations to be applied to the scanned documents when converting data to a digital format. Its aim is to align an image before processing because text segmentation and recognition methods require properly aligned next lines. The first one is based on the Hough transform. Hough transform is performed on the scanned document image and the variance in ρ values is calculated for each value of θ. The angle that gives the maximum variance is the skew angle. The second approach is based on the base-point method. Here a concept of basepoint is introduced. After the successive base-points in every text line within a suitable sub-region were selected as samples for the straight-line fitting. The average of these baseline directions is computed, which corresponds to the degree of skew of the whole document image. The above mentioned algorithms have been implemented and the results have been compared for accuracy.

Keywords: ***Basepoint, Hough Transform, Recognition, Segmentation, Skew***

I. INTRODUCTION

Organizations are moving at a fast pace from paper to electronic documents. However, large amounts of paper documents inherited from a recent past are still needed. Digitization of documents appears as a bridge over the gap of past and present technologies. Scanners tend to be of widespread use for the digitization of documents. One of the important problems in this field is that, very often documents are not always correctly placed on the scanner either manually by operators or by the automatic feeding device. For humans, rotated images are unpleasant for visualization and introduce extra difficulty in text reading. For machine processing, image skew brings a number of problems that range from needing extra space for storage to making more error prone the recognition and transcription of the image by automatic Optical Character Recognition tools (OCR)[1][2][3][4]. These reasons make skew detection[5] and correction phases a common place in any environment for document processing. This very frequent problem yields rotated images. Besides that, it increases the complexity of any sort of automatic image recognition, degrades the performance of OCR tools,

increases the space needed for image storage, etc. So skew detection and correction is very important step that needs to be corrected before segmentation.

Skew estimation and correction are the important preprocessing steps of line segmentation and word segmentation approaches. Skew angle is the angle that the text lines in the digital image makes with the horizontal direction. There are two types of skew in document images. They are single skew and multiple skew[7].

Single skew: In this skew, whole document is skewed to single angle. This work deals with single skew problem. Example for single skew is as shown in Fig 1.1.

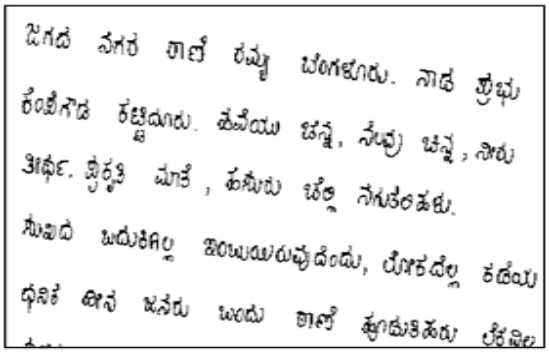

Figure 1.1: Single Skew

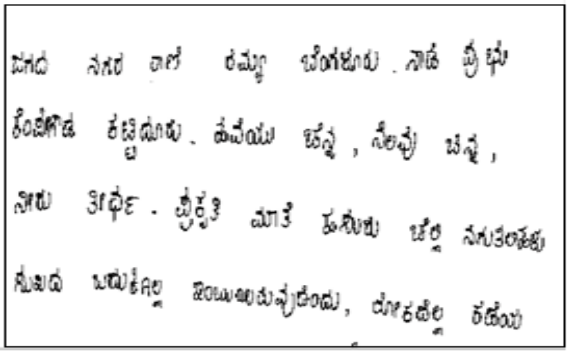

Figure 1.2: Multiple Skew

Multiple Skew: In this skew, scanned document can have many sections; each may be skewed to different angle. Example for multiple skew is as shown in Fig 1.2.

Detecting the skew of a document image and correcting it are important issues in realizing a practical document reader. A variety of algorithms have been devised to tackle these problems which are listed below.

"ManjunathAradhya V N, Hemantha Kumar G and P Shivakumara's Algorithm": M. Aradhya [9] discusses the skew detection technique based on Hough Transform. It uses Thinning as preprocess step.

"Skew Detection Algorithm for Gurmukhi Script by G S Lehal and RenuDhir": Lehal and Dhir [10] proposed algorithm for detection of skew in documents containing Gurmukhi script. This algorithm is based on projection profile.

"Fast Algorithm for Binary Image Rotation": This proposed algorithm [11] for rotating binary images originates from the property that the rotation matrix can be factorized into three skewing matrices. Such a property had been applied to derive a fast three-pass rotation algorithm for general images (gray level or color images). To the rotation of general images, the virtue of decomposing rotation into the three skewing is that all operations are simplified to one-dimensional interpolations and data moves.

Skew correction can be achieved by

- Estimating the skew angle
- Rotating the image by the skew angle in the opposite direction.

There are many approaches for skew detection and correction. This paper presents two approaches to overcome the problem of skew. They are

- Hough transform method.
- Baseline method.

1.1. Document Image Processing Steps

In document analysis the first step is to acquire a digitized raster image of the document using a suitable scanning system. Then it is followed by skew detection and skew correction [8]. Before the structure of the text is obtained, a test is carried out to find out whether the document is skewed. Then skew corrected image is obtained. Now the skew corrected image can be applied for further steps like segmentation. Then the resulting image is given for recognition system.

The document image processing system is as shown in Fig 1.3. It is consisting of three blocks. They are Database, Skew detection and Skew correction.

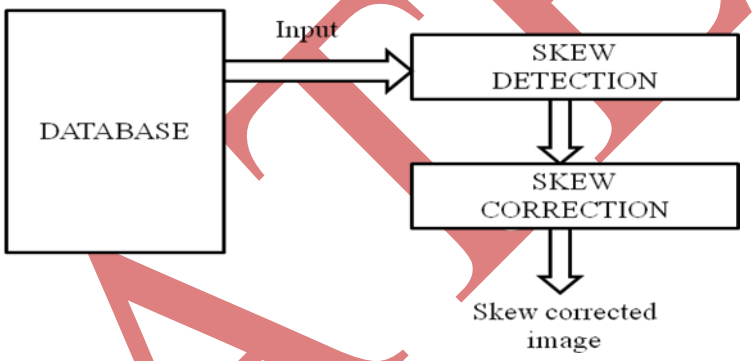

Figure 1.3: Document image processing steps

1.2. Motivation

As per survey, there were no solutions for accurate skew correction. There may be a chance where single document will have multiple skewed lines. It is very hard to segment the skewed text document**.** These problems which we have mentioned motivated us to take up this project. We have introduced the techniques in order to overcome all these problems.

II. IMPLEMENTATION

The implementation of the complete project is divided into three different modules/subprojects which are then connected together to form a complete main module. The three modules are listed below

- Document acquisition
- Skew detection and correction
- User interface

2.1 Document Acquisition

Different handwritten historical documents are collected from different places to build a huge database. Documents over here are of two kinds mainly palm leaf manuscripts and paper manuscripts which are stored in the memory so that it can be accessed with the help of user interface whenever needed. These documents are given as an input for skew detection and correction process which needs to be skew corrected.

2.2 Skew Detection and Correction

We implemented Skew detection and correction using two approaches namely Hough transform method and Baseline method.

1. **Hough Transform Method:** Hough transform is one of the powerful global techniques which can be used to isolate features of a particular shape within an image. This method is used mainly for single skewed document.
2. **Baseline Method:** Skew correction is done by estimating the lower baseline and determining its angle relative to horizontal. This method is used mainly for skew detection of multi skewed document image.

2.3. User Interface

A GUI is a type of computer human interface on a computer. It solves the blank screen problem that confronted early computer user.

Method 1: Hough Transform Method

When Hough transform method is selected, Following three steps are executed

Step1: Binarization

Step2: Skew Detection

Step3: Skew correction

Let us study in detail about each step by considering respective flow charts.

Step1: Binarization

The flow chart given in Fig 2.1 explains about resizing and Binarization of scanned input image before giving it to the actual skew detection and correction steps.

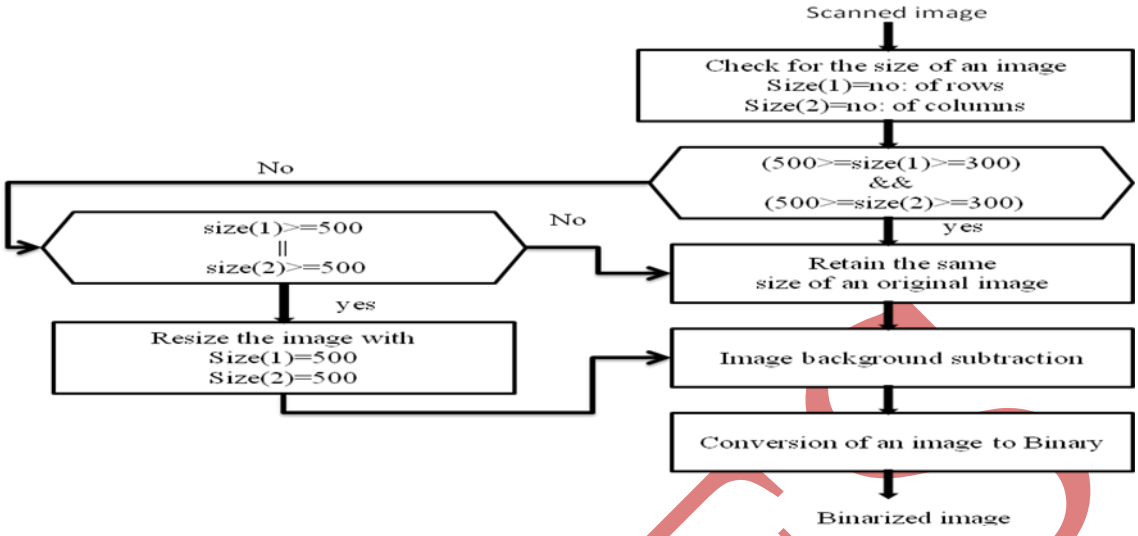

Figure 2.1: Flow chart for Binarization

The size of the scanned input image is determined. If number of rows and number of columns are between 300 to 500 then the same size of original image is retained later is subjected to background subtraction. If number of rows or number of columns is greater than 500, then the image is resized to 500X500. Later is subjected to background subtraction. The resized original image is then subtracted from the resized blank image in order to remove the background noise. The background subtracted image is then converted to binarized image.

Step2: Skew detection

The steps involved in detecting the skew angle are shown in flowchart given in Fig 2.2.

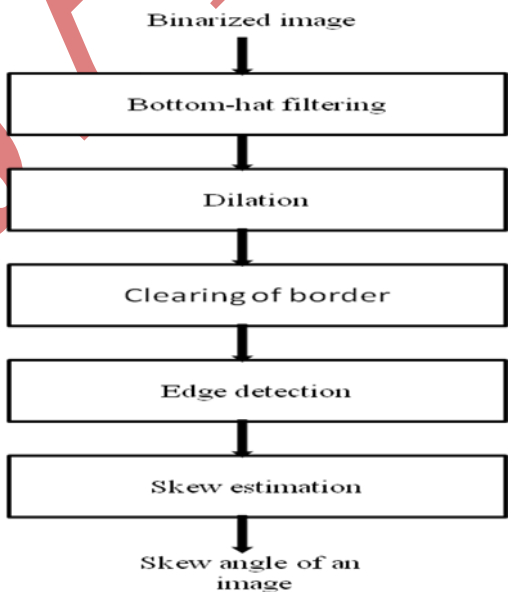

Figure 2.2: Flow chart of Skew angle estimation process

The obtained binary image is filtered by using Bottom-hat filter to enhance the black spots in a white background. 'Imbothat' built in function is used. The filtered image is then subjected to dilation. Dilation is an operation that grows or thickens object in binary image. 'Imclose' inbuilt function is used that performs dilation followed by erosion. The next step in the process is to clear borders. It clears the unwanted borders surrounding the word. 'Imclearborder' inbuilt function is used. The Border cleared image is applied as input to the edge detection process is used to detect the extra edges. The built in function, used over here is

Syntax: edge (o,'sobel','horizontal')

The edge detected image is later given to hough transform function that estimates the parameters of the image such as theta (θ) and rho (ρ). The theta value corresponding to the point at which rho value is maximum is nothing but the skew angle of the text document. This is performed using built in function 'houghpeaks(H, 1)'.

Step3: Skew correction

If the skew angle obtained is greater than 90 degree, the angle is subtracted from 180 degree and the angle is made negative because line has to be tilted by 180 degree for proper working of skew correction. If not 90 degree, same angle is retained.

With this retained skew angle the image is skew corrected by using Matlab inbuilt function 'imrotate(I, θ)'.

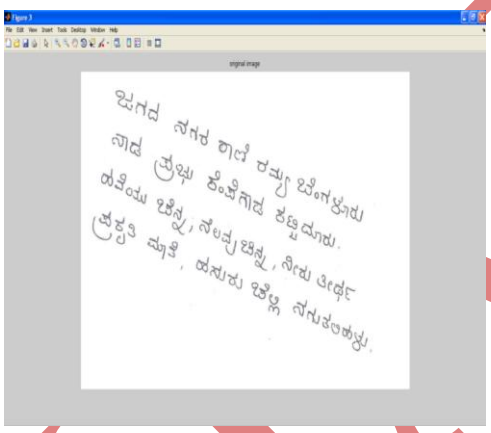

Figure 2.3: Skewed Image

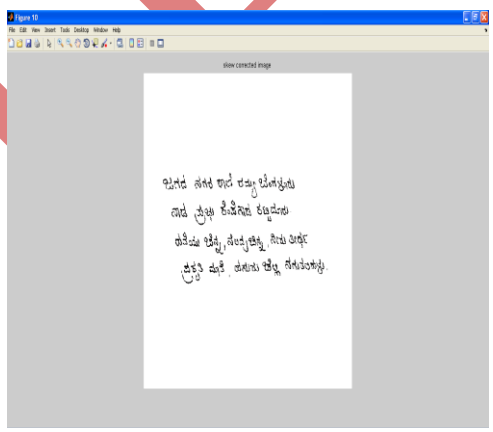

Figure 2.4: Skew Corrected Image

Method 2: Baseline method

When Baseline method is selected, following steps are executed and flow chart for this method is as shown in Fig 2.5. Steps involved in this method are as follows

Original image which is having multiple skewed lines is cropped in order to get required number of lines in the given document. The lines within the selected area as shown in Fig 2.6 are considered for further steps. In the cropped image only single line is considered .The tilted line 1 and line 2 are shown in Fig 2.7. Bottom pixels are found. These pixels are considered to be the points and a line is fitted to these points by using MATLAB inbuilt function 'polyfit(x,y,1)'. Angle at which the line has tilted is found by using parameters (slope and intercept) of the line. By using an inbuilt function i.e. 'imrotate' in MATLAB, the image is skew corrected.

Baseline is drawn for skew corrected line from the first bottom pixel till the width of the line. The skew corrected lines are as shown in Fig 2.8.

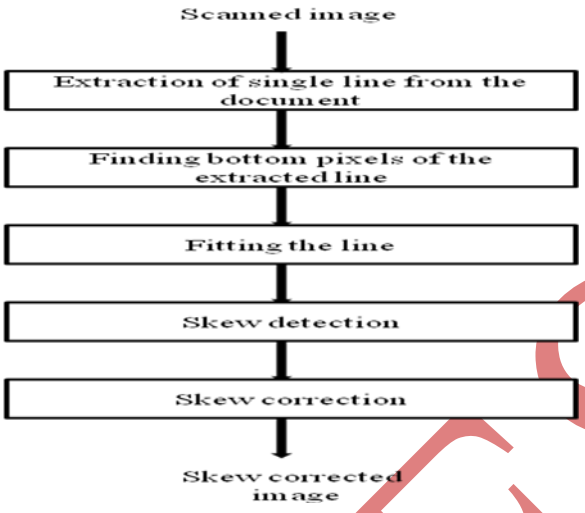

Figure 2.5: Flow chart of Baseline method implementation

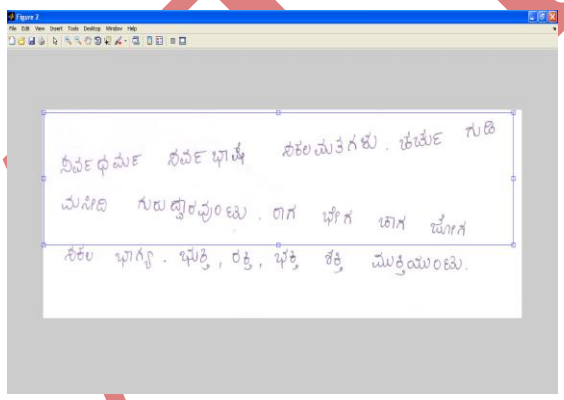

Figure 2.6: Original image

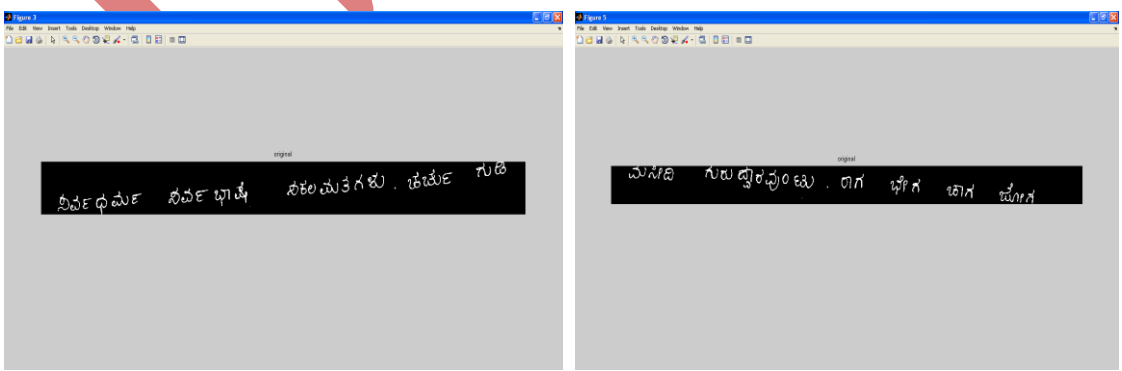

Figure 2.7: Tilted line1 and line 2

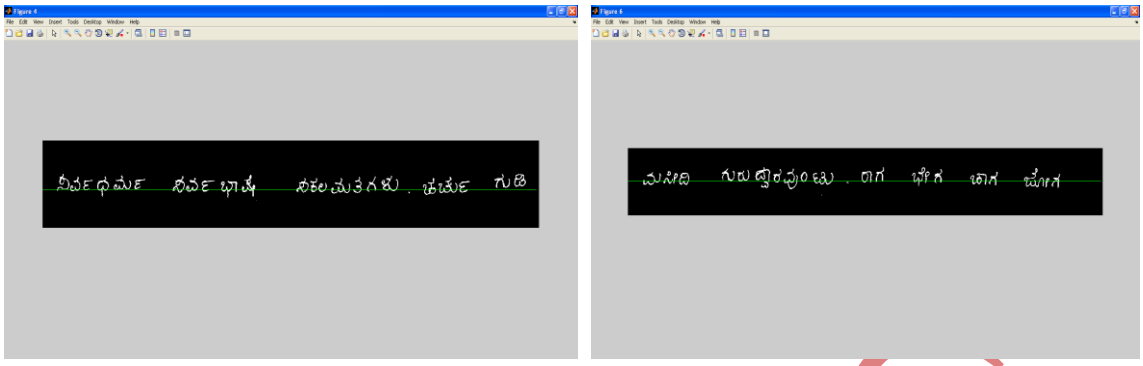

Figure 2.8: Skew corrected lines

III. APPLICATIONS

The various applications of the Skew Detection and Correction approach are as listed below

- In improving the visual output of facsimile machines and duplicating machines
- In signature verification system
- In banks
- Library automation post-offices
- Defense organizations
- Reading aid for the blind
- Library automation
- Language processing
- Multimedia design

IV. RESULT ANALYSIS

The experiment was carried out over three sets of handwritten documents. Result analysis for Hough Transform method is shown in Table 4.1.

Table 4.1: Result analysis for Hough Transform method

Document No	Author's name	Accuracy (In %)	Problems encountered
1	M K Gandhi	100.00	
2	Rabindranath Tagore	100.00	
3	Valmiki	00.00	Since document is palmleaf

Upon conducting experiment on the documents mentioned in Table 4.1, we got 100 percent result for documents 1 and 2. But zero percent of it for document 3 (Valmiki). Because document 3 is palm leaf and it has high background noise with itself and ornamentation.

For Baseline method we have considered our own handwritten documents as we did not find any poet's document having multiple skew. Palm leaf manuscripts are also been considered. This method works for both single skew and multiple skew.

Table 4.2: Result analysis for Baseline method

Document No	Number of lines in the document	Number of lines corrected	Accuracy (In %)	Problems encountered
1	4	2	50.00	Insufficient spacing between the successive lines
2	6	6	100.00	
3	8	6	75.00	

Table 4.2 shows the result analysis for Baseline method. Document 1(Valmiki) is a palm leaf manuscript. It had four lines among which only two lines got skew corrected. So we got 50 percent accuracy. For the document 2, all the lines got skew corrected. So we got 100 percentage of accuracy. Document 3 has 8 lines among which 6 lines got skew corrected with 75 percentage of result.

V. CONCLUSION

For efficient Text-line Segmentation of Historical documents, the preprocessing steps are necessary such as Skew detection and Correction. Since different authors may write with different skew (either Single skew or Multi skew). In order to overcome this we have implemented two approaches, namely Hough transform and Baseline method. Hough transform method can be applied only for single skewed document and Baseline method can be applied for both single skewed and multi skewed documents.

5.1. Scope for Future Enhancement

- Converting handwritten document into printed format.
- Improving the complexity of baseline method.
- Language Identification.

REFERENCES

[1] B. B. Chaudhuri and U. Pal, "A complete printed BanglaOCR system", Pattern Recognition, vol.31, pp.531-549.

[2] R. Plamondon and S. Srihari. Online and offline handwriting recognition: A comprehensive survey. IEEE Trans. Pattern Anal. Machine Intell., 22(1):63–84, 2000.

[3] SEETHALAKSHMI R "Optical Character Recognition for printed Tamil text using Unicode", Thanjavur, TamilNadu.

[4] R SANJEEV KUNTE and R D SUDHAKER SAMUEL"A simple and efficient optical character recognition systemfor basic symbols in printed Kannada text".

[5] Amin A, Fischer S. "Fast algorithm for skew detection". IS&T/SPIE Conference on Real-Time Imaging, USA, 1996; 65–76 [6]

[6] L. Likforman-Sulem, A. Hanimyan, and C. Faure, "A Hough Based Algorithm for Extracting Text Lines in Handwritten Documents", Proc. 3rd Int'l Conf. on Document Analysis and Recognition (ICDAR'95), 1995, pp. 774-777.

[7] U. Pal, M. Mitra, and B. B. Chaudhuri, "Multi-skew detection of Indian script documents", in Proc. 6th Int.Conf. Document Analysis Recognition, pp. 292-296.

[8] ManjunathAradhya 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.

[9] Ahmed Maher and Ward Rabab, "A Rotation Invariant rule-Based Thinning Algorithm for Character Recognition", IEEE Transaction on pattern analysis and machine Intelligence, Vol 24, No 12, December 2002.

[10] Lehal G S and Dhir R, "A Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents", Fifth International Conference on Document Analysis and Recognition, 1999. ICDAR '99.

[11] A. Amin, S. Fischer, T. Parkinson, and R. Shiu. Fast algorithm for skew detection. IS&T/SPIE Symposium on Electronic Imaging, San Jose, USA, 1996.