International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 1 Issue 5, July - 2012

Comparison of Skew Detection and Correction Techniques By applying on Gurmukhi Script

.Loveleen Kaur , Lecturer at Guru Teg Bahadur Institute of Engineering and Technology, Chhapianwali, Malout, Muktsar, Punjab, India.
Mandev Singh, Lecturer at Guru Teg Bahadur Institute of Engineering and Technology, Chhapianwali, Malout, Muktsar, Punjab, India.

***Abstract*—** This paper includes the information about the techniques used to detect Skew which are introduced during the scanning of the documents. It also discusses about the tool which have been used to implement the technique. The comparison of the techniques is done on the basis of the angle measured and by applying the algorithm on the Gurmukhi Script. The methods provides a very efficient way to calculate the Skew. Correction in 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. The method deals with an accurate measure of skew, within-line, and between-line spacings and locates text lines and text blocks. The detection and correction of the images are done.

Keywords— Document processing, Gurmukhi Script, Skew angle, Skew Correction, Skew Detection.

I. INTRODUCTION

Organisations are now a days very fastly moving from paper documents to electronic documents. However, large amount of old paper documents of past are still needed. The process Digitalization provides us the way to use the old documents with the help of present technologies. We use scanner for the digitalization of documents. The major problem which we face in the case of scanner is that the documents are not always placed correctly on the scanner by the operator which leads to the rotated images. These rotated images are unpleasant for visualization and also makes the text reading difficult. The skew introduced during the scanning of the documents give rise to various problems such as 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 The problems of skew is to be solved that is why skew detection and correction has become a very important phases in document processing. . Due to the occurrence of various problems many algorithms have been developed for skew detection and correction and still the work is going on.

The Optical Character Recognition (OCR) technique is used to convert the image form of a document into a text in the form of character codes. Basically, OCR system is used to convert the image into text which reduces the space required for storage and transmission of documents.OCR technique is used for various applications such as office automation, automatic data entry in banks, libraries and post offices, etc. When we scan a document, if it is not a well aligned then it results in a skewed digital image. Skew angle is the angle that the text lines in the digital image makes with the horizontal direction[3]. Skewness also refers to the tilt in the bitmapped image of the scanned paper for Optical character Recognition (OCR)[4]. Most of the OCR algorithms are sensitive to the skew of the input document image making it necessary to develop algorithms to perform skew detection and correction automatically. The methods can be mainly categorized into five groups. The one based on Hough transform, cross correlation, Projection profile, Fourier transformation and k nearest neighbor (k-NN) clustering.

Skew detection and correction are important preprocessing steps of document layout analysis and OCR approaches. The Character Recognition process starts with data acquisition and ends up with the skew-correction process. Skew detection is a part of pre-processing stage. It is necessary for both segmentation and recognition as skewed words may cause considerable difficulties in both tasks. This paper has been organized into 4 sections. Literature survey of the existing methods has been discussed in Section 2.Proposed Methodology is discussed in Section 3.Experimental results are shown in Section 4 followed by Conclusion and References.

II. LITERATURE SURVEY

In 1993, O' Gorman finds skew angle by using a nearest neighbor clustering technique. For each component the direction of its nearest neighbor component is detected. A histogram of the direction angle is computed and the peak of which indicates the document skew angle [1]**.** In 1994, Le found connected components in a document and considered only the bottom pixels of each component for Hough transform which also reduces the amount of data to be processed[6]. In 1997, Kim and Govindaraju has employed chain code method of image representation. Chain code is a linear structure that results from quantization of the trajectory traced by the centers of adjacent boundary elements in an image array[7]. Each data node in the structure represents one of eight grid nodes that surround the previous data node. Coordinates of the start and end points of

each vertical line extracted provide the slant angle. Global slant angle is the average of all the angles of the lines, weighted by their length in the vertical direction since the longer line gives more accurate angle than the shorter one. In 1996, Pal and Chaudhuri applied a clustering technique on candidate points for skew estimation. In 1997, Chaudhuri and Pal suggested a skew detection approach for the printed characters of Devanagari script. This approach is based upon detecting the inherent feature (head line) in the Devanagari word[7] In 1998, Lehal and Madan have used the physical properties of Gurmukhi script to device a range free skew detection scheme for Gurmukhi script, which detects skewness in range -180° to 180°. In 1999, Lehal and Dhir proposed a range free skew detection technique for machine printed Gurmukhi documents. This approach can easily be extended to other Indian language scripts such as Devanagari and Bangla[9]. Most characters in these scripts have horizontal lines at the top called headlines. The characters forming a word are joined at top by headlines, so that the word appears as one single component with headline. The ratio of pixel density above and below the headline of any word in Gurmukhi script is always less than 1. These inherent characteristics of the script have been employed and a new algorithm based on projection profile method has been devised. In 2001, Pal discussed about the documents having multi-skew text of Devanagari and Bangla scripts. The connected components have been labeled and then selected. The upper layer of the selected components is found by the column-wise scanning from the top of the component[6]. Portions of the upper envelop which look similar to DSL is detected. They are then clustered into groups belonging to single text lines. Estimates from these individual clusters give the skew angle of each text line. In 2003, Cao Yang, Shuhua Wang and Li Heng proposed skew detection and correction in document images based on straight-line fitting. The bottom center of the bounding box of a connected component is regarded as an Eigen-point[11]. According to the relations between the successive Eigen-points laid on the baseline are extracted as sample points. Then these samples are adapted by the least squares method to calculate the baseline direction. In 2004, Kapoor proposed a new algorithm for skew detection and correction which exploits the inherent Shiro-Rekha of the Devanagari script and it does skew detection and correction of a document in a single attempt without any intermediate step[9]. In 2007, Manjunath Aradhya proposed method for skew detection in binary document images. The method considered the some selected characters of the text which may be subjected to thinning and Hough transform to estimate skew angle accurately[10].In 2009,Omar proposed skew detection and correction technique for Arabic document images based on centre of gravity[11].. It involved inscribing the text in the document by an arbitrary polygon and derivation of the baseline from polygon's centroid.

III. PROPOSED METHODOLOGY

This section presents the proposed methodology to determine the skew angle accurately. The technique is implemented using Visual C++ language. When a document is scanned, the output image of the document may get skewed (scanned at an angle) if the document is not properly placed on the scanner. If the scanned image is a text written in any language, the skew in the image can be detected and corrected by a software program. This program contains a new algorithm called topline algorithm which finds out skew angle from the scanned text image. A second algorithm rotates the image in a direction to remove the skew in the image.

The topline algorithm does not operate directly on the skewed image. First the skewed image is converted to a segment file or a thin segment file, and then the algorithm operates on one of these files to find skew angle. Now which file (segment file or thin-segment file) gives us more accurate skew angle when compared to the actual skew angle is part of the research work.

This application program uses a third party program (ocr.exe) which runs from within main application, takes skewed image file as an input and generates two files, segment (seg.dat) and thin-segment (thinseg.dat) files as output. The topline algorithm then finds skew angle theta1 from seg.dat file and theta2 from thinseg.dat file. The original skewed image (say a.bmp) is then rotated by angle –theta1 and is copied to file a_segrotated.bmp. Next a.bmp is rotated by angle –theta2 and is copied to file a_thinsegrotated.bmp. Next all the three files, a.bmp (original), a_segrotated.bmp and a_thinsegrotated.bmp can be displayed one at a time. The two later files (with skew removed) can be compared with the original skewed image file.

Segment file: This is basically a text file containing a matrix of 0's and 1's, which is similar to the matrix of pixels in an image file. This file is generated from the image file. When generating segment file from image file, each pixel in an image file is converted to either 0 or 1. If the pixel is dark, it is converted to 1, otherwise it is converted to 0. An average value of a pixel can be set (assumed) below which the pixel can be considered as dark pixel.

Thin-Segment file: This is a text file and is generated from a segment file. A further algorithm applied on the segment file converts it into thin-segment file.

The technique is developed by using the concept of Hough Tranform. Hough Transform technique is an approach used for fitting lines and curves[11]. This approach is preferred when the objective is to find lines or curves formed by groups of individual points on an image plane. The method involves a transformation from an image plane to a parameter space. Consider the case in which lines are the objects of interest. The line is expressed as $\rho = X \cos\theta + Y \sin\theta$. There are two line parameters namely, the distance (ρ) and the angle (θ) which defines transformation space. Each coordinate (x, y) of ON pixel in the image plane is mapped onto the locations in the transformed plane for all possible straight lines. It has been assumed that the images taken should be noise free.

Algorithm topline

Step 1: Input .bmp image to the OCR.exe and convert the image to seg.dat and thinseg.dat.
Step 2: Apply both seg and thinseg algorithm on the images.
Step 3: Calculate the h1 and the h2 coordinates.
Step 4: Calculate the width and the margin.

Step 5: Input these values to the formula for detecting the Skew angle.
Step 4: The values are entered into the formula:

Θ= a tan (diff /(w-margin)).

Step 5: Then this angle is used to rotate the image to get the skew freed images.
Step 6: The difference among both the angles can be seen easily.
Step 7: Stop.

IV. EXPERIMENTAL RESULTS

This section presents the results of the experiments conducted to study the performance of the proposed method. I have considered different skewed documents from different sources like journals, textbooks, newspapers and the like. Obtained Skew Angle by the proposed methodology for these documents are reported in Table 1. The scanned text document images are shown in Fig. 4.1 to Fig. 4.5

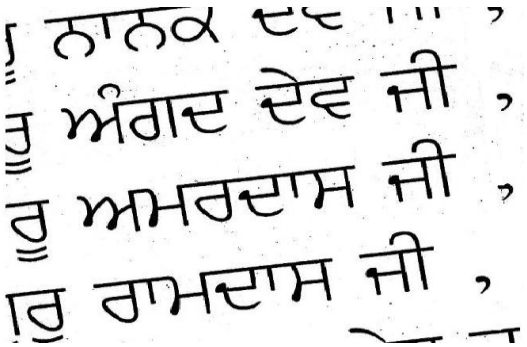

Fig. 1 Skewed Gurmukhi Printed text document

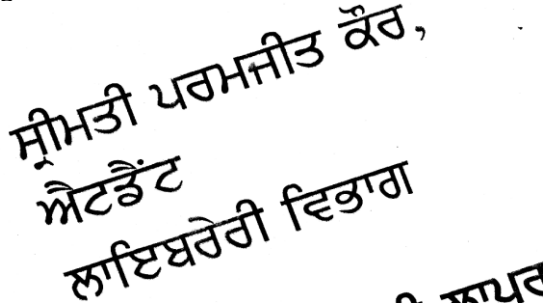

Fig. 2 Skewed Gurmukhi Printed text document

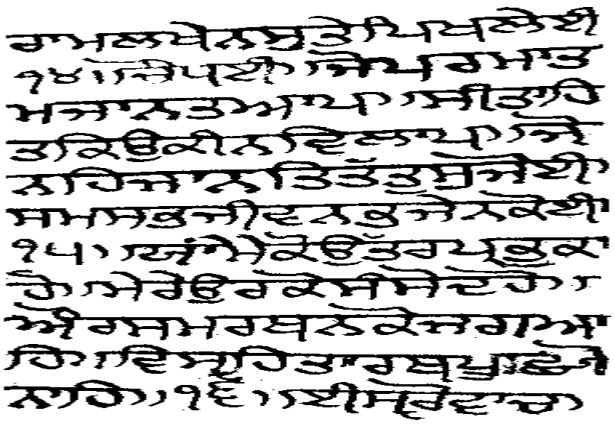

Fig. 3 Skewed Handwritten Gurmukhi document

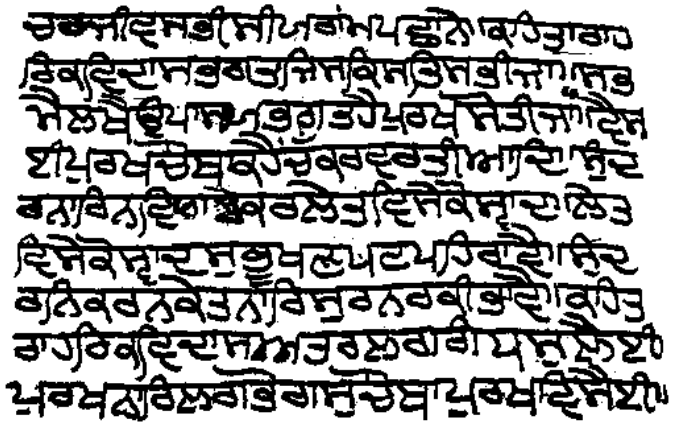

Fig. 4 Skewed Handwritten Gurmukhi document

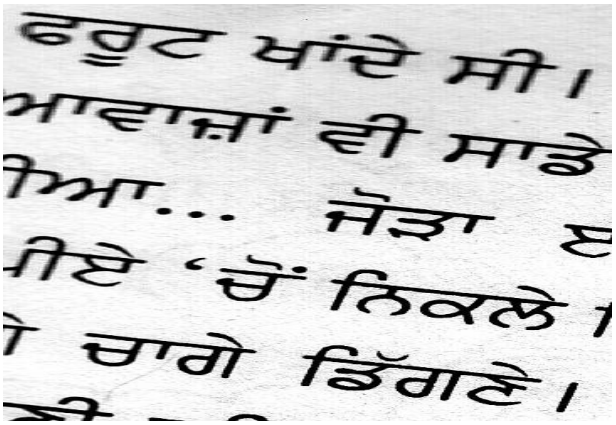

Fig. 5 Skewed Gurmukhi Printed text document

Table 1: Performance of the Proposed method with different scanned images

Figure	Detected Angle By Seg Method	Detected Angle By Thinseg Method
1	1.50	1.50
2	10.42	10.58
3	5.19	5.54
4	2.01	1.86
5	9.53	9.75

The fig 6 to fig 10 shows the corrected images.

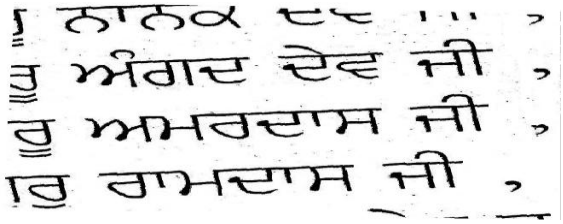

Fig.6 Corrected image of Fig.1

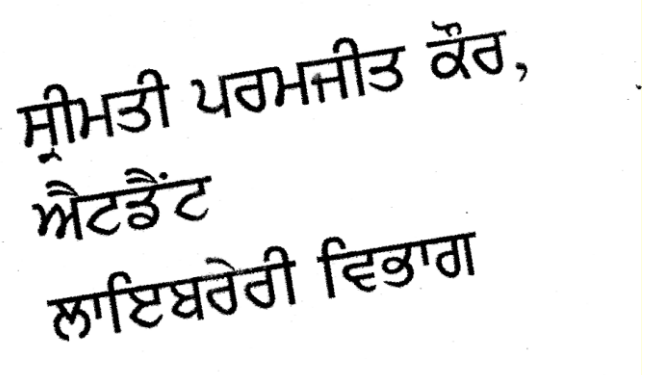

Fig. 7 Corrected image of Fig.2

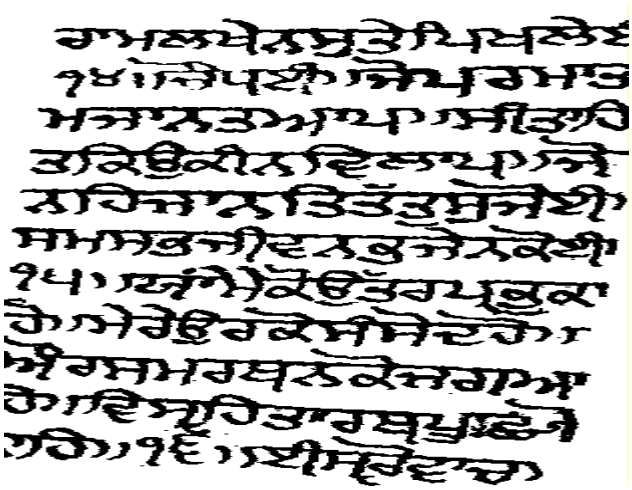

Fig. 8 Corrected image of Fig.3

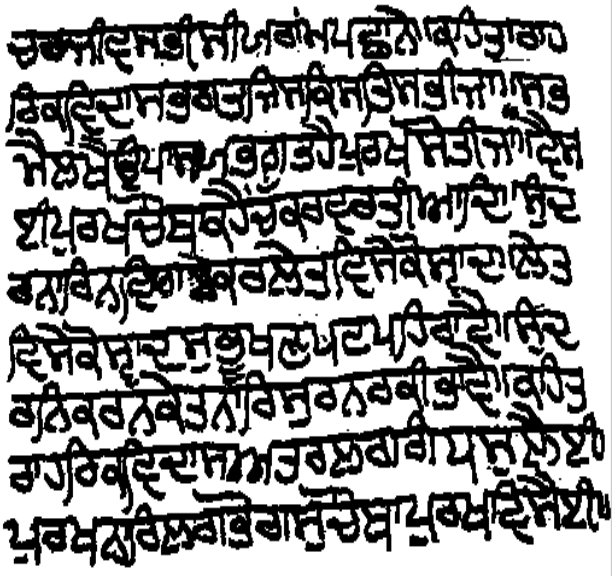

Fig. 9 Corrected image of Fig.4

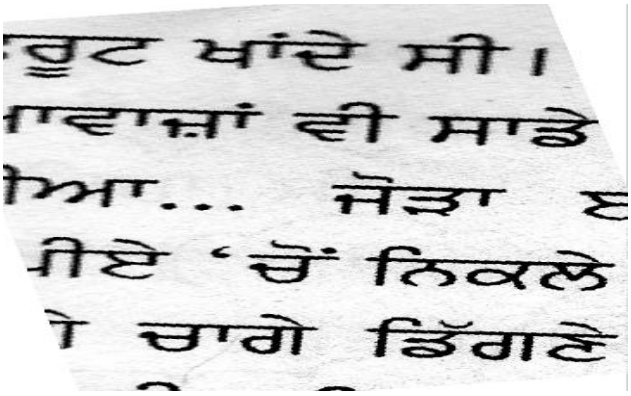

Fig.10 Corrected image of Fig.5

V. Conclusion

In summary, an efficient, novel and accurate methodology to estimate skew angle is presented in this paper. The proposed method work is based on the image top. The proposed method is fast compared to other HT based methods. We have shown the comparison among both the methods. It has been shown that among them thinseg provides more accurate results. The performance of the proposed method is shown by applying it on the various scanned documents. The main advantage is that the method works on all types of scripts. However, the proposed method fail for document images containing text with picture.

References

[1] O'Gorman L.,"The document spectrum for page layout analysis", IEEE Transactions on Pattern analysis and Machine Intelligence", vol.15,no. 11, pp. 1162-1173,1993.

[2] Chaudhury, B.B., Pal, U.,"OCR Error Detection and Correction of an Inflectional Indian Language Script",Proceedings of ICPR.,1996.

[3] Chaudhury, B.B., Pal, U.," Skew angle detection of digitized Indian script documents", IEEE Trans. PAMI.,vol. 19,pp.182-186,1997.

[4] G **S** Lehal and Renu Dhir, **"A** Range Free Skew Detection Technique for Digitized Gurmukhi Script Documents" ,Proceedings 5th International

Conference of Document Analysis and Recognition, Bangalore, pp. 147-152, 1999.

[5] G S Lehal and Chandan Singh,"A Gurmukhi Script Recognition System",Published byIEEE,2000.

[6] Rajiv Kapoor , Deepak Bagai , T.S. Kamal,"Skew Angle Detection of a cursive handwritten Devanagari script character image", published by J. Indian Inst. Sci.,pp.161–175,2002.

[7] Davessar, N. M., Madan, S., and Singh, H., "A Hybrid Approach to Character Segmentation of Gurmukhi Script Characters", Proceedings of the 32nd Applied Imagery Pattern Recognition Workshop,2003.

[8] Rajiv Kapoor,Deepak Bagai,T.S.Kamal,"A new algorithm for skew detection and correction",publishedin Pattern Recognition Letters 25,pp. 1215–1229,2004.

[9] Bo Yu Li and Yun Wen Chen,"Classification Using the Local Probabilistic Centers of k-Nearest Neighbors",18th International Conference on Pattern Recognition,2006.

[10] Manjunath Aradhya V N, Hemantha Kumar G, and Shivakumara P,"Skew Detection Technique for Binary Document Images based on Hough Transform", published in International Journal of Information Technology,2007.

[11] Atallah Mahmoud Al-Shatnawi and Khairuddin Omar," Skew Detection and Correction Technique for Arabic Document Images Based on Centre of Gravity", published in Journal of Computer Science 5 (5):363-368, 2009.