International Journal of Applied Engineering Research ISSN 0973-4562 Volume 10, Number 18 (2015) pp 39071-39073
© Research India Publications. http://www.ripublication.com

An Algorithm for deskewing a Document Image

Prof Akila Victor
Assistant Professor VIT University (India)

Gurleen kaur brar
B.Tech CSE 3rd year VIT University (India)

Sanyam Seth
B.Tech CSE 3rd year VIT University (India)

Abstract
Due to increasing amount of data, the documents these days are digitized and kept in a digital form. This is either done by scanning of an image or capturing an image using digital Camera. During this process the image may sometimes be skewed from its original position. This requires the need to deskew the images in order to store them in proper manner. This paper deals with the problem of skewing of images by introducing an algorithm that helps in deskewing the document images. The algorithm efficiently works for binary document images. The algorithm takes an input image, calculates the skew angle and then performs several mathematical operations in order to perform skew correction. The proposed algorithm was implemented using Matlab and has been tested using a sample of various document images in various languages. The algorithm not only works for Hindi, Bengali, English and Tamil languages but for others as well.

Keywords Image acquisition, skew angle, skew correction, matlab

I. Introduction

The field of digital image processing has been growing tremendously in the past years. It basically deals with the use of computer algorithms to perform several actions such as classification, feature extraction, pattern recognition and signal analysis on a digital image. With the growing technology, there has been a tremendous increase in the volume of information which is now stored electronically rather than using conventional methods such as paper. The information is now stored in the form of document images. The first step in digital image processing is the image acquisition which deals with capturing the images either by scanning or using a camera. The digital processing technique is very helpful in the storing large amount of data. The data is stored by converting the documents into their digital form. The documents kept in the digital form sometimes contain a viable piece of information that needs to be extracted from the images. The methodology of text processing is used to extract the text from the images. Text processing basically refers to the mechanism of creating and manipulating the electronic text. It involves various processes such search and replace, extract, filtering a file and generating a processed report of the content. One of the main problems with text extraction is the problem of skew detection and skew correction. This problem needs to be solved before the image can be used for further analysis like extraction, enhancement of further processing. But it is a challenging problem to deskew the images.The largest class ofmethods for skew and correction is based on Hough transform. The other class of methods is based on Fourier-transform, straight line fitting and moments.

Hough transform provides accuracy and simplicity. But due to its slow speed many researchers have been workingon increasing its speed without compromising the accuracy. So, for improving computational efficiency of Hough transform, there are various variations have been proposed to reduce the computational time for skew angle.

Yang Cao et alproposed a skew detection method using the concept of Eigen-point which is based on straight-line fitting. After the relations between the successive Eigen-points in every text line within a suitable sub-region were analyzed, the Eigen-points most possibly laid on the baselines are selected as samples for the straight-line fitting. This works well, when only a pure text sub-region in the whole image is selected.

George et alused the moments technique for the first time by proposing a simple and fast algorithm that works with any kind of objects, like bend lines (not only straight lines), pictures, columns of text etc. The orientation θ of this object between the principal axis and the horizontal axis gives an exact estimation of the skew angle. The angle range of -30 degrees to +30 degrees is the big disadvantage of this method.

Brodić et al introduced the extension to the moment based method for the text skew estimation. In order to make it useful for the handwritten text, an extension to the moment-based method has been introduced by introducing bounding boxes. The application of the moment-based method to the connected components estimates their local text skew.

In this paper, an improved technique using Hough transformation to detect and correct the skews of an image is proposed. The algorithm works on the same transformation but in a more efficient manner. The algorithm uses the method of information entropy to find the optimal angle using the gradient descent.

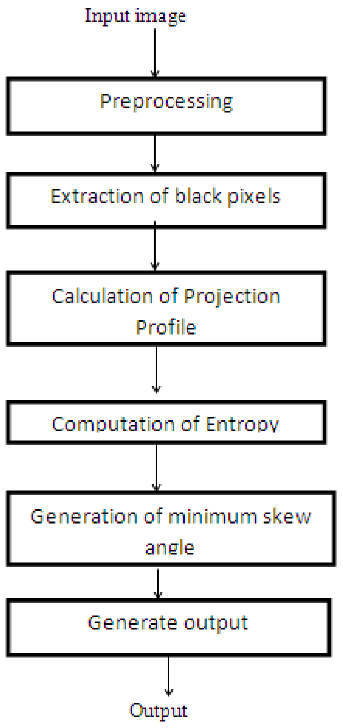
The paper organization is as follows: Section 2 explains the improved algorithm by introducing next step of Skew correction using the method of information entropy. Section 3 describes the experimental results and Section 4 makes the conclusions.

II. Proposed algorithm

The overall goal of this paper is to propose a more efficient algorithm which reduces the amount of time required to detect and correct the digital images. It also deals with the process of collecting, analyzing and evaluating the document images in order to improve the efficiency of the algorithm. The new proposed algorithm includes 7 steps in which the first three steps shows the technique to find the skew angle and later steps shows the technique for skew correction.

1. Inputs- The algorithm takes three inputs. The first is a two dimensional binary image, the second is the maximum angle for deskew and the third is the resolution for angle to be searched.
2. Setting of parameters- Set the input parameters according to the margin of the image.
3. Preprocessing – Then the image is preprocessed, that means it converts the truecolor image RGB to the grayscale intensity image with the help of predefined rgb2gray function.
4. Extract black pixels: The black pixels are extracted and stored in another array for further analysis.
5. Calculate the projection profile: The projection profile is calculated by plotting a histogram between the various indices of the black pixels and the unique value of the pixels.
6. Compute the entropy: The histogram values are converted to probability values of the corresponding frequencies of the indices at which the black pixel occurs. Now compute the information entropy.
7. Generate the minimum skew angle: the minimum skew angle is generated in the end and the input image is rotated by the value of the min skew angle.

The algorithm steps 4, 5, 6 calculate the information entropy of a projection profile in a more efficient manner as it calculates the value of the entropy in O(N) operations. The algorithm is very helpful in long data sets where the image may have a lot of data.


Input image
Preprocessing
Extraction of black pixels
Calculation of Projection Profile
Computation of Entropy
Generation of minimum skew angle
Generate output
Output

Flowchart 1

The above flowchart explains the algorithm step by step in a sequential manner

III. Experimental results

The above algorithm has been coded and implemented in Matlab and has been tested using different images in different languages such as Hindi, Bengali, English, Latin and Tamil. The experiments show the results on the various tested images. Figure 1 represents an input document image in Bengali language. Figure 2 shows another input image in Latin. the input images are two binary document images. Figures 3 and 4 show the outputs of the both Bengali and Latin document images. The outputs represent the images after deskewing.

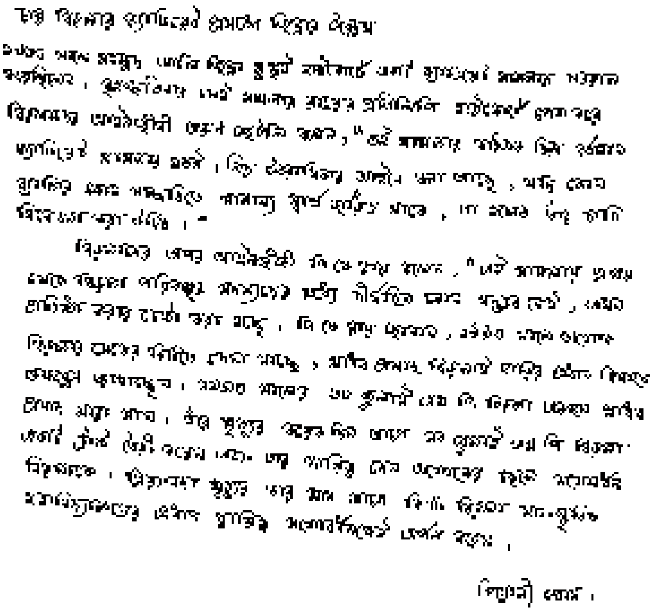

[illegible]

Figure1: Skewed Bengali document image

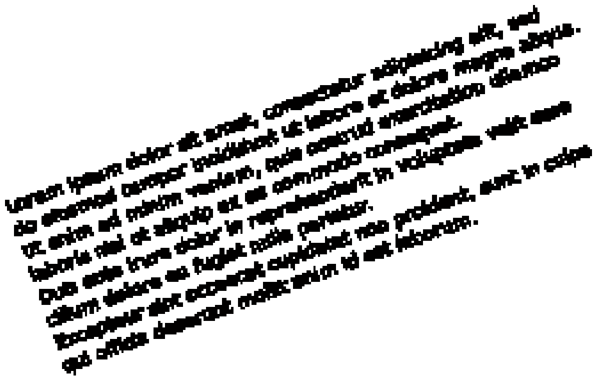

Lorem ipsum dolor sit amet, consectetur adipiscing elit, sed
do eiusmod tempor incididunt ut labore et dolore magna aliqua.
Ut enim ad minim veniam, quis nostrud exercitation ullamco
laboris nisi ut aliquip ex ea commodo consequat.
Duis aute irure dolor in reprehenderit in voluptate velit esse
cillum dolore eu fugiat nulla pariatur.
Excepteur sint occaecat cupidatat non proident, sunt in culpa
qui officia deserunt mollit anim id est laborum.

Figure 2: Skewed Latin document image

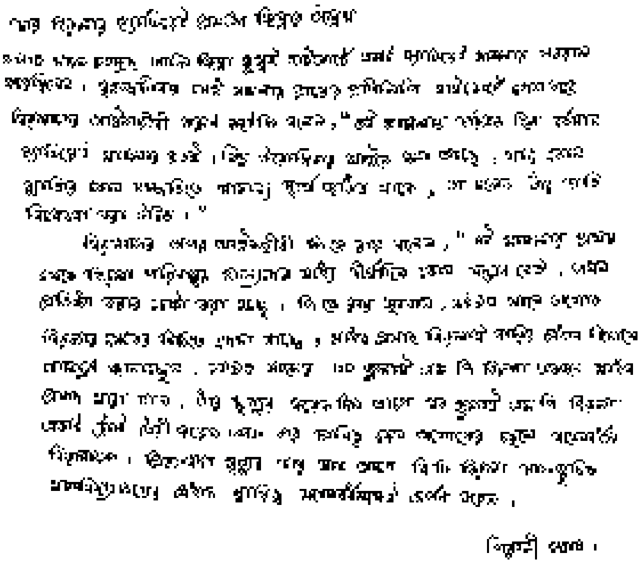
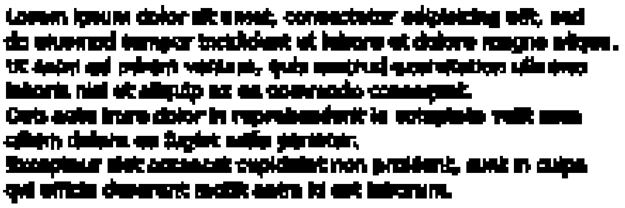


Figure 3:Deskewed Bengali document image



Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed
do eiusmod tempor incididunt ut labore et dolore magna aliqua.
Ut enim ad minim veniam, quis nostrud exercitation ullamco
laboris nisi ut aliquip ex ea commodo consequat.
Duis aute irure dolor in reprehenderit in voluptate velit esse
cillum dolore eu fugiat nulla pariatur.
Excepteur sint occaecat cupidatat non proident, sunt in culpa
qui officia deserunt mollit anim id est laborum.

Figure 4: De-skewed Latin document image

IV. Conclusion

The image skew detection and correction has many methods and approaches like Hough transformation, Fourier-transformation, nearest neighbor, cross- correlation and moments. But each has its own limitations. The limitations may be based on the speed or the accuracy of the computation. The above proposed algorithm based on Hough transformation though takes inputs for resolution and max angle but it can yield high precision of about 1% positive or negative.The code has been developed using Matlab and has been tested using document images in various languages like English, Bengali, latin, Tamil etc. The algorithm can be modified in a way so that it can be used for colored images by extracting the first colored pixel(except white) and then approach in the similar manner as explained in the algorithm above. The algorithm is very helpful in deskewing the document imagesas it produces results with the efficiency of 1 percent.

References

[1] Yang Cao, Shuhua Wang, Heng Li ―Skew detection and correction in document images based on straight-line fitting‖, Pattern Recognition Letters, Volume 24, Issue 12, August 2003, Pages 1871- 1879.

[2] K. George,K. Nicholas., ―A Fast High Precision Algorithm for the Estimation of Skew Angle Using Moments‖, Department of Informatics and Telecommunications University of Athens, (2002).

[3] Brodić, Darko, and Z. Milivojević. "Estimation of the Handwritten Text Skew Based on Binary Moments." Radioengineering 21.1 (2012): 162-169.

[4] Gonzalez, Rafael C. and Woods Richards E., (1999),"Digital Image Processing", Addison Wesley.

[5] S. Lowther, V. Chandhran, et al. ―An accurate method for skew determination in document images‖ .In Proceedings of the Digital Image Computing Techniques and Applications.Melbourne ,Australia, (2002)

[6] K. Iuliu, E. Stefan et al., ―Fast Seamless Skew and Orientation Detection in Document Images‖, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) SchlossBirlinghoven, 53754 Sankt Augustin, Germany, (2010).

[7] Chandan Singh, Nitin Bhatia, Amandeep Kaur, ―Hough transform based fast skew detection and accurate skew correction methods‖, Pattern Recognition, Volume 41, Issue 12, December 2008, Pages 3528-3546. Improved Skew Detection and Correction approach using Discrete Fourier Algorithm

[8] Yang Cao, Shuhua Wang, Heng Li ―Skew detection and correction in document images based on straight-line fitting‖, Pattern Recognition Letters, Volume 24, Issue 12, August 2003, Pages 1871- 1879.

Authors:

Sanyam Seth, currently pursuing B.Tech Computer Science and Engineering 3rdyear at VIT University. Fields of interest are Image processing, Web Development, Big Data and Database Management

Gurleen Kaur, currently pursuing B.Tech Computer Science and Engineering 3rd year at VIT University. Fields of interest are Image processing, Web Development, Software Development and Database Management