TEXT DOCUMENT DE-SKEWING TO FIND THE ORIGINAL TEXT

[1]Bhabani Prasad Mishra
[1]Asst. Prof.
[1]Dept. of CSE ,
[1]GATE, Berhampur(O)

Abstract
This paper presents a method for the skew correction of text documents in digital image format, using primarily the Hough Transform. The Hough Transform provides a good estimate of line angles in an image and this can be extended to lines of text. Positive angles between 0 and 90 degrees are estimated correctly in both test images while negative angles are estimated correctly only after implementing a solution found empirically. This routine also fits the result-ing corrected image to a given bounding box.

Keywords
Text de-skewing, Hough Transform, Morphology

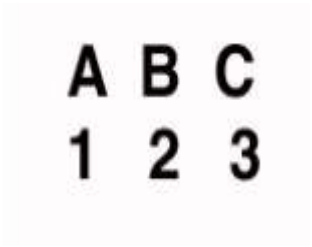
I.INTRODUCTION
The skew correction of text documents is the starting point for digital document analysis, because such analysis cannot proceed when a document is skewed even at a small angle. This algorithm therefore seeks to detect a document's skew angle based on the angle at which the text lies. Assuming that a text's angle is the same as that of the entire docu-ment, the Hough transform is used to detect the skew angle.

HOUGH TRANSFORM

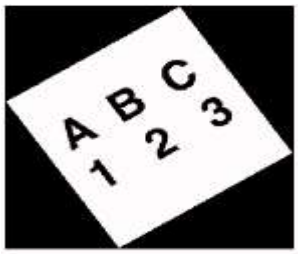
The Hough Transform is a mapping of lines in Cartesian x-y space to radii and angles in - space. Lines in x-y space are of the form,
$y = ax + b$
while each point in Hough line space (,) describes a line through (x,y) with slope
and distance from the origin
[1]. Lines in Hough space are sinusoids of the form,
xcos + y sin =.Hough sp
ace is separated into angle bins on one axis and
radius bins on the other. Each point , then is an accumu-lator counting the number of Hough space sinusoids that pass through it as the above equation is evaluated at each x,y and . Ultimately, the largest accumulators correspond to a line that exists in x,y space with slope and a distance from the origin. It is easy to see that when applied to lines of text, the angle of skew may be found by looking at the - location of the highest accumulator(s) [1][2]

II.ALGORITHM [2]
1)Find gradients in image by using a Sobel mask.
2)Threshold the Sobel result to create a binary im-age.
3)Perform morphological 'closing' and 'thickening' so text is thinned out and resembles many points on a line. The Hough Transform is more accurate,
if it operates on clear lines in the document im-age[1]. A pre processed image is shown below.
4)Split image in half width-wise, in order to get bet-ter resolution in angle estimation for each half of the image.
5)Perform Hough Transform on each part of the im-age. Return accumulator results.
6)Perform some simple statistical analysis on the re-sults of the Hough operation, throwing out any re-turned angles equal to zero or greater in magnitude than 90o.
7)The number of times a given angle is returned by the Hough routine is tabulated, and the angle re-turned the most is the skew estimate.
Pre-processed mesh text.jpg after Rotation

III.RESULTS
Experimentation was done using skew angles between but not equal to -90 and 90 degrees. For positive angles the algorithm returns perfect results, giving an exact estimate of the skew angle and the skew is corrected by rotation. An example of the algorithm's implementation is shown here for a skew of +33 degrees.

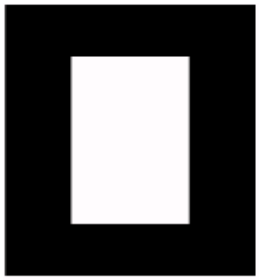


Original Image



Skewed Image

Bounding Box



Bounded Image

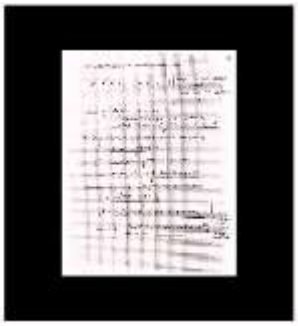
Corrected Result

When running the algorithm with a negative skew angle, the returned estimation was consistently off by 90 degrees.
For example, starting with a skewed image at -70 degrees returns an estimate of the skew angle at +20 degrees. The workaround for this is to simply correct the image using the incorrect estimate such that the image now lies with a skew of -90 and run the algorithm again. The new estimate is exactly -90 degrees and the image can be now be corrected accordingly. Running the algorithm twice with a positive angle also results in an estimate of -90 degrees so in the interest of keeping it as automatic as possible, the user should enter a guess of -1 if the skew angle is thought to be negative and a guess of +1 if thought to be positive. Using another image, this one with a lot of text and grid

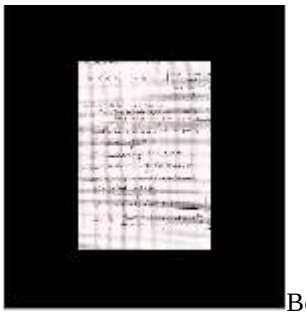
lines, the results are the same. The example below is for a skew angle of 75 degrees.

Original Image

Skewed Image



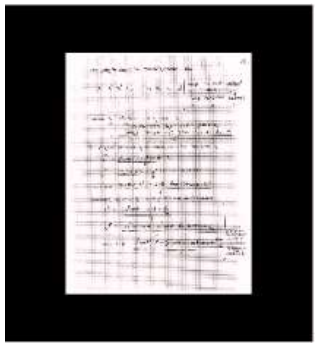
Corrected Result

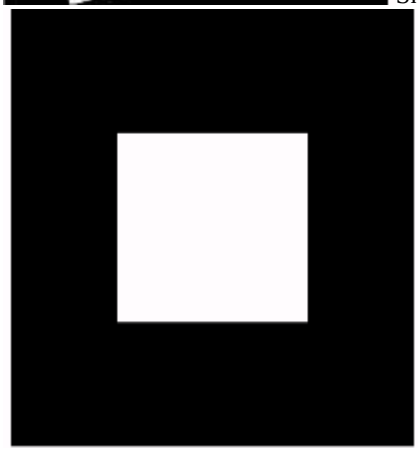


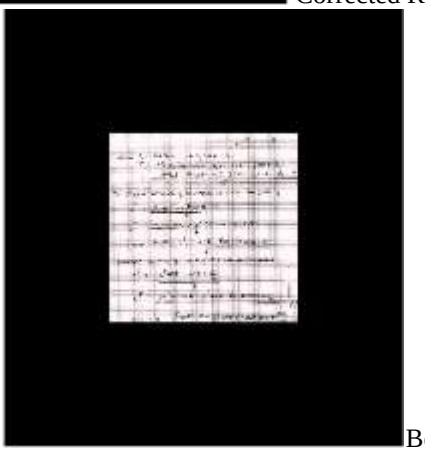
Bounded Image

The negative angle approach is shown implemented here for an angle of -75 degrees and with the help of a user guess of -1.


Skewed Image


Corrected Result


Bounding Box


Bounded Image

III.DISCUSSION

For both sparse text images and dense text image this algo-rithm works great for ositive skew angles. The Hough routine is presented with preprocessed images whose line
detail has been enhanced, so the routine returns some very good skew angle estimates and some that aren't so good. It is easy then to automatically throw out those that aren't
within the known skew angle range between -90 and 90 degrees. Of those that remain, it is then a simple matter to count up the number of occurrences of each. The angle that
occurs the most in that tabulation is the good skew angle estimate.

IV.SUMMARY

As shown before, this algorithm works very well for posi-tive skew angles and can be made to work for negative skew angles with a little help from the user indicating the
sign of the skew. With this "guess" from the user being the only interaction, this algorithm lends itself to an almost completely automated implementation for use in document analysis devices.

ACKNOWLEDGMENTS

The free version of Gonzalez's, Eddins' and Woods' DIPUM toolbox obtained from the "Digital Image Processing Using Matlab" website was used for the Hough transform routine
and test images were obtained from Prof. Merat's course website.

REFERENCES

[1] Chengming Sun, Deyi Si, "Skew and Slant Cor-rection for Document Images Using Gradient Di-rection". Pgs 1-2.
[2] R. C. Gonzalez, Richard E. Woods, "Digital Image Processing, "2ndEdition, Prentice Hall, Upper Saddle River, NJ, 2002.