IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Vol.-6, Issue-7, July 2018 E-ISSN: 2347-2693

Skew Detection and Correction in Text Document Image using Projection Profile Technique

Rubani[1*], Jyoti Rani[2]

[1]Giani Zail Campus College of Engineering & Technology, Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India
[2]Giani Zail Campus College of Engg. & Tech., Maharaja Ranjit Singh Punjab Technical University, Bathinda, Punjab, India

Available online at: www.ijcseonline.org

Accepted: 16/Jul/2018, Published: 31/Jul/2018

Abstract - The detection and correction of document skew is one in all the foremost vital document image analysis steps. Projection profiles have a few applications in record picture process and that they consider flat and vertical lines being adjusted to the tomahawks. The proposed system works in two phases. In the first phase system find the skewed angle from the input text document and in the second phase system correct the sleekness in the given document on line by line basis. At the last every line output line is combined to obtain the final output. The proposed system is experimented on approx 30 text documents for testing. The accuracy of algorithm is around 97%.

Keywords: Skew Detection, Skew Correction, Profile Projection Technique**.**

I. INTRODUCTION TO SKEW RECOGNITION AND REDRESS

The recognition and redress of report skew is a standout amongst the most essential archive picture examination steps. Page format examination and preprocessing activities utilized for character acknowledgment rely upon an upright picture or, in any event, information of the point of skew. [1] One case of a procedure which is spoilt by skew is the utilization of level and vertical projection profiles. Projection profiles have numerous applications in record picture handling and they depend on flat and vertical lines being adjusted to the tomahawks.

Skew discovery of checked report pictures is a standout amongst the most critical phases of its acknowledgment preprocessing. At the point when an archive is checked, skew is unavoidably acquainted into the picture due with different elements. [2] The skew of the examined archive picture determines the deviation of its content lines from the level or vertical hub.

The skew of the archive picture can be a worldwide (every one of report's squares have a similar introduction), various (record's squares have an alternate introduction) or non uniform (numerous introduction in a content line). [3] Generally, measurement diminishment and skew estimation are two stages of the skew identification of checked report pictures. The initial step of skew location is measurement decrease. Each picture is a point in the picture space. Each measurement of the picture space is identified with one of its pixels. The main arrangement of highlights that can be considered for a picture is its pixel esteems. As it were, the estimation of each picture pixel is one of its highlights. The measurement of pictures is high and the work of all picture pixels as its highlights makes many-sided quality and high computational cost on the insignificant highlights. Measurement diminishment is the way toward decreasing the extent of highlights or picture pixels and finding another element with much lower measurements. The second step of skew identification is skew estimation. In this progression, utilizing the capacity characterized in the past advance, the skew is evaluated. The point relating to the most extreme or the base estimation of the capacity is normally considered as the skew. Thus, in this progression, the most extreme or the base of the model capacity is accomplished. Up to this point, numerous strategies for skew location of filtered report pictures have been proposed.

Rest of the paper is organized as follows, Section I contains the introduction of skew recognition and redress, Section II contain the literature review, Section III contain the proposed methodology, Section IV contain the various algorithms used by the proposed system , section V contains the results and discussions of proposed system, Section VI describes conclusion and future scope.

II. LITERATURE SURVEY

Darko Brodić & Zoran N. Milivojević.et.al[2016], used Combined Entropy Algorithm for skew detection. This paper proposes the algorithm for text skew estimation based on the combined entropy calculation. The method consists of three steps. In the first step, it calculates the horizontal and vertical projection profiles. In the second step, the horizontal and vertical entropy for the rough angles is calculated. In the last step, the horizontal and vertical entropy calculation for the smooth angles is performed. The calculated entropy creates the two cost functions: horizontal and vertical. The position where each of cost functions has an absolute minimum represents horizontal and vertical estimated text skew angle. In the last step, it estimates the text skew angle as a mean of horizontal and vertical estimated text skew angles. The functionality, correctness and robustness of the proposed algorithm is investigated by the experiment, which is based on DISEC'13 document database.

Skew discovery, adjustment and division of written by hand kannada archive Nov 2012

Mamatha Hosalli Ramappa,et.at., Optical character acknowledgment (OCR) alludes to a procedure of creating a character contribution by optical means, such as examining, for acknowledgment in consequent stages by which a printed or written by hand content can be changed over to a shape which a PC can comprehend and control. A non specific character acknowledgment framework has diverse stages like clamor evacuation, skew location and redress, division, highlight extraction and arrangement. Consequences of the later stages can influence the execution of the resulting stages in the OCR procedure. To make the consequences of the consequent stages more precise, the skew recognition and revision and division assume a vital part. In this paper, writer have proposed plans for skew discovery and redress, division of transcribed Kannada record utilizing jumping box strategy, Hough change and shape identification separately. A normal division rate of 91% and 70% for lines and words is gotten separately. [11]

Skew discovery and redress in record picture in light of straight-line fitting

Yang Cao, et. al., during archive checking, skew is definitely brought into the approaching record picture. Since the calculations for format examination and character acknowledgment are for the most part exceptionally delicate to the page skew, skew location and amendment in archive pictures are the basic strides previously design investigation. In this paper, a novel skew location strategy in view of straight-line fitting is proposed. Also, an idea of eigen-point is presented. After the relations between the neighboring eigen-focuses in each content line inside a reasonable sub-locale were broke down, the eigen-focuses most conceivably laid on the baselines are chosen as tests for the straight-line fitting. The normal of these benchmark bearings is processed, which compares to the level of skew of the entire record picture. At that point a quick skew amendment strategy in light of the examining line show is additionally exhibited. Tests demonstrate that the proposed approaches are quick and precise. [25]

III. PROPOSED METHODOLOGY

The proposed system works in two phases. In the first phase system find the skewed angle from the input text document and in the second phase system correct the sleekness in the given document on line by line basis. The proposed system works in the following phases:

Preprocessing: In this phase input document is entered in the system from user. Threshold operation is performed on the input image to remove the unwanted pixels from the given text document image.

Line extraction: In this phase line is extracted from the input text document.

Skew Angle Detection: Skew angle of every line extracted is calculated independently and that corresponding line is rotated in the reverse direction in which the skew angle is calculated. At the last every line output line is combined to obtain the final output.

IV. VARIOUS ALGORITHMS USED BY THE PROPOSED SYSTEM ARE GIVEN AS BELOW

1.4.1 Skew Detection and Correction techniques:-Skew detection and correction methods are used to align the paper document. For any image taken by scanning the document may have some human mistakes such as while setting the document for scanning, the paper may be placed with some tilt in either of the direction which causes the rotation image. It is also possible that, for hand written document the writer did not have written in perfect alignment causing the rotation. These all techniques are of immense importance in skew detection and its correction process as they help to increase the readability of the document.

Algorithms for skew detection and correction

Step 1: Input the handwritten text documents written in single or multiple scripts.

Step 2: Convert the RGB image into Gray Scale Image.

Step 3: Biniarize the grayscale image and store it into a matrix format.

Step 4: Extract the line from the text document.

Step 5: Convert the extracted lines into independent block with the help of RLSA algorithm.

Step 6: Calculate the corner coordinate points of the block extracted in the step 5.

Step 7: Estimate the skew angle with the help of the corner points extracted in the step 6.

Step 8: Rotate the line in the reverse order at the specified angle.

Step 9: Add the output obtained in the step 8 to the temporary output.

Step 10: Check if it is not the last line then go to the step 4
Step 11: Display the final result to the user.
Step 12: End

1.4.2 Line Extraction technique: - In this phase the text document written in the multiple scripts is scanned from left to right to the find the empty space. Find empty space which will be on the top of the document will be taken on reference and the system scan the document downwards to find the another empty space after the series of line.

The portion discovered in between these empty spaces is considered as the text block which will be further sent to the angle estimation phase.

This technique extracts a line .If it finds white pixel then it means it is a line. If black pixel is found it means the line is finished. Then again some process is followed. This process continues until the document is finished.

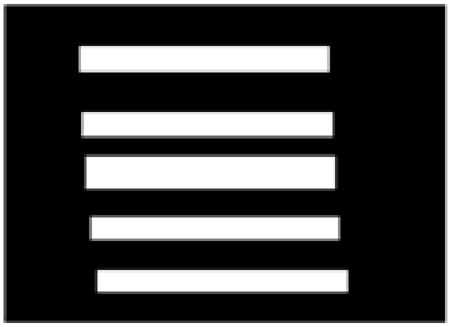


Figure 1.2 Line Extraction with the help of white and black pixels

In the above diagram, white portion represents the text in the original document and black pixels represent the background respectively.

Line Extraction Algorithm
Step 1: Input the text document image.
Step 2: Starting from the first line of the text document scan the entire document to find the row in which all the pixels are black.
Step 3: mark this row in which all the pixels are black as end index.
Step 4: Copy and remove the pixels from the text document image into an array matrix.
Step 5: repeat the step 2 to step 4 until end of the document reached.
Step 6: End.

1.4.3 Angle Estimation technique: - Angle estimation is a technique which is used to measure an angle from a document. Once we estimate an angle, then we rotate that document so that its skewed ness is removed. This whole process occurs line to line which means that one by one a line is taken, angle is estimated and then it is rotated to remove skewed ness.

The main idea of angle estimation is to find the right angled triangle then find the corresponding skewed angle.

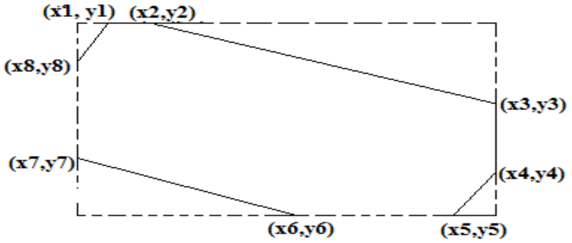


Figure 1.3 Example of angle estimation

In the above figure, the extracted points represent the skewed ness in the text document. With the help of these extracted points proposed system detect the skew.

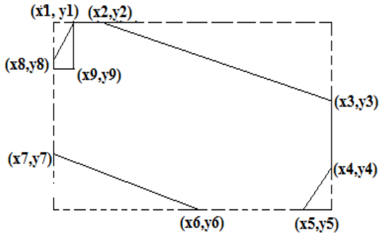


Figure 1.4 Angle estimate using right angle triangle

In the above figure, (x1,y1)(x2,y2)(x3,x4)(x5,x6)(x7,y7)(x8,y8)(x9,y9) are extracted as the rectangle points. These points represent the skewed ness in the text document. From these extracted points proposed system evaluates the three points of the triangle as (x1, y1), (x8, y8,) (x9, y9) with the help of existing coordinate points of the rectangle. The triangle formed by these points right angled triangle. Now with the help of sin formula proposed system evaluate the skewed angle of the text document.

From these extracted points proposed system evaluates the three points of the triangle as (x1, y1), (x8, y8), (x9, y9) with the help of existing coordinate points of the rectangle. The triangle formed by these points right angled triangle.

Angle Estimation Algorithm
Step 1: Input the line of which skew angle is to be estimated.
Step 2: transfer the image into matrix form.
Step 3: Calculate the x and y coordinates from the corners of the matrix.
Step 4: form the triangle from these calculated coordinate values .
Step 5: Calculate the skew angle from this triangle.
Step 6: End.

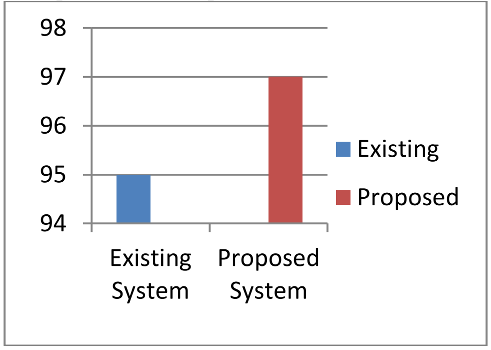
V. RESULTS AND DISCUSSION

The proposed system is experimented with more than 30 text documents for testing. These text documents are written in various languages like Punjabi, Hindi and English. The direction of skew presented in these documents is both upward and downward. The algorithm has provided satisfactory results. Proposed system is implemented and tested using MATLAB. The use of projection profile for determination of skew angle is really simple and efficient. The accuracy of the proposed system is defined as the difference in skew angle calculates to the original skew angle present in the document. The accuracy of algorithm is around 97%.

These differences are given in the following table:

Accuracy (%)	**Value**
Existing System	95%
Proposed System	**97%**

Comparison of Proposed method with existing techniques



1) Existing system works on the whole documents at once i.e. detect and correct the skew from the entire text document.

Proposed system works on the document in line by line basis i.e. detect and correct the skew of every line independently.

2) Existing system only works on the documents containing single script.

Proposed system can work on the documents having multiple scripts on it.

3) Existing system could not detect and correct the skew if there is an unequal space between the adjacent lines.

Proposed system can detect and correct the skew from the documents even if there is an unequal space in between the two adjacent lines.

4) Existing system Rotate the whole document at once to correct the skew from the given text document.

Proposed system rotates the document on line by line basis to correct the skew from the given document.

5) Existing system is not capable of detecting correcting the upward and downward skew together in a single text document. Proposed system can detect and correct the upward and downward skew together in the single text document.

Table 1.5.1 Statistics of the proposed system

Parameters	**Value**
No. of Images Tested	40
Skew detected and corrected	38
Overall System accuracy	97
Avg. Time	2.13
Scripts	Punjabi , Hindi and English
Type	Printed and Handwritten

VI. CONCLUSION AND FUTURE SCOPE

1.6.1 Conclusion

In the proposed system we have implemented a technique based on advanced profile projection to detect and correct the skew angle from the image. Existing systems detect the skewed angles and rotate the documents in the reverse side of the angle detected but Proposed system is very efficient to detect and correct the skew in handwritten and printed text documents having unequal spaces in between the adjacent lines. Existing system works on whole documents at once but proposed system works line by line on the document and gives the accurate results. A combination of various techniques is used to detect and correct the skew in the given text image. These techniques are vertical profile projection technique, horizontal profile projection technique and line extraction algorithm. Proposed system is tested on various types of inputs containing text in various scripts written in the multiple languages. System is evaluated on three parameters which are skew angle detected, time to correct the skew angle, and type of skewed angle. Proposed system shows good results on various images tested than that of existing system.

1.6.2 Future scope

- In future system can be extended to perform skew correct on text document in doc., pdf format containing text as well as images.
- System can be further improved to take lesser time to correct the skewed angle detected in the text document images.

REFERENCES

[1] Darko Brodić & Zoran N. Milivojević 2016 Text Skew Detection Using Combined Entropy Algorithm, ITC-Journal of Information Technology And Control, 46: PP 308-318.

[2] A. Papandreou, B. Gatos, G. Louloudis and N. Stamatopoulos 2013 Document Image Skew Estimation Contest, ICDAR-Published in the 12th International Conference on Document Analysis and Recognition, PP 1444-1448.

[3] B. Aditya Vighnesh,Abhishek Kumar,Abhishek Kumar 2013 Skew Detection in Handwritten Documents IJCA- International Journal of Computer Applications, 69:PP 17-18

[4] Bishakha jain, Mrinaljit Borah 2014 A Comparison Paper on Skew Detection of Scanned Document Images Based on Horizontal and Vertical Projection Profile Analysis, IJSRP-International Journal of Scientific and Research Publications 4: PP 1-4.

[5] Danil Korchagin 2012 Automatic Time Skew Detection and Correction, IJCEE- International Journal of Computer and Electrical Engineering , 4: PP 684-687.

[6] Deepak Kumar, Dalwinder Singh 2012 A Review of Scanned Document Skew Detection and Correction Techniques, IJCST-International Journal of Computer Science Trends and Technology, 3: PP 251-254

[7] Jinal Patel,Anup Shah,Dr. Hatal Patel 2015 Skew Angle Detection and Correction using Randon Transform, IJEECS-International Journal of Electronics, Electrical and Computational System, 4: PP 1-6

[8] Jung Gap Kuk,Nam Ik Cho 2009 Feature Based Binarization of Document Images Degraded by Uneven Light Condition, IEEE-Published in 10th International Conference on Document Analysis and Recognition, PP 1-5

[9] Lipi Shah, Ripal Patel, Shreyal Patel, Jay Maniar 2014 Skew Detection and Correction for Gujarati Printed and Handwritten Character using Linear Regression, IJARCSSE- International Journal of Advanced Research in Computer Science and Software Engineering, 4: PP 642-648

[10] Loveleen Kaur, Simpel Jindal 2011 Skew Detection Technique for Various Scripts, IJSER-International Journal of Scientific & Engineering Research 2: PP 1-3

[11] Mandip Kaur, Simpel Jindal 2013 An Integrated Skew Detection and Correction Using Fast Fourier Transform and DCT, IJSTR-International Journal Of Scientific & Technology Research 2: PP 164-169

[12] Mamatha Hosalli Ramappa and Srikantamurthy Krishnamurthy 2012 Skew Detection, Correction and Segmentation of Handwritten Kannada Document, IJAST-International Journal of Advanced Science and Technology 48: PP 71-88