International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 355-359
Published Online August 2019 in IJEAST (http://www.ijeast.com)

TEXT EXTRACTION FROM SKEWED TEXT LINES

Ms. Shilpa P Jalli
Department of CSE
KLE Dr. M S Sheshgiri College of Engg & Tech,
Belagavi, Karnataka, India

Prof. Shivanand M Patil
Department of CSE
KLE Dr. M S Sheshgiri College of Engg & Tech,
Belagivi, Karnataka, India

Dr. Virendra S Malemath
Department of CSE
KLE Dr. M S Sheshgiri College of Engg & Tech,
Belagavi, Karnataka, India

Abstract— **The camera caught pictures containing content are having slanted text lines due to contortions by page and the view edge of camera. Hence it is vital while extracting text from the document, the text ought to be horizontal and words are inline appropriately. This proposed method introduces the strategy based on image processing techniques. Text extraction needs to play out some preprocessing tasks for better results. At the beginning, we find gray scale picture of given document image and then after executing the thresholding operation on text document image, we have to perform skew correction. This method, centers around the slope correction of text and we introduce a skew standardization approach which is depends on text skew correction algorithm with respect to x-axis. This method determines the exact text skew angle, and rotate image to correct skew efficiently. The method has been tested on different text document images and accomplishes over 97% accuracy.**

Keywords— **Skewed Text Lines, Image Processing, Optical Character Recognition (OCR), Text Skew Correction Algorithm.**

I. INTRODUCTION

Text image digitization is a significant technique to expand the grade as well as similarity of picture. So document image analysis assumes an imperative job is recognition of data from text pictures. Recognizing horizontal content lines of text picture is simple when contrasted with the skewed content lines. Content of picture is warped due to camera viewpoint and different twists. In this method, we endeavored to handle these snags happen while binarizing text pictures. We utilized image processing approaches for correcting skew and extracting skewed content lines from text pictures.

The camera caught pictures containing content are having slanted content lines due to contortions of page and the camera view edge. Hence this is vital during extracting data from the picture, the content ought to be horizontal and text should be in line appropriately. In any case, content lines segmentation in skewed picture is a troublesome strategy for dewraping strategies. This proposed method introduces strategy dependent on image processing techniques for correcting skew and recognition of text from skewed content picture. Then words in document picture are recognized. The premises of CCs are utilized for segmenting the words.

Content segmentation is significant design analysis ventures of text picture understanding frameworks. This is typically usable prior nourishing content to an optical character recognition (OCR) system. Content data may likewise be utilized for executing the greater part of different document image processing functions, for example, binarization, image cleanup, slant amendment, zone segmentation, and character recognition, dewarping of distorted text pictures. Dewarping is moderately another document picture pre-preparing step when contrasted with others they are referenced here. This is an expert procedure of amending camera-caught text pictures which experience the ill effects of point of view and geometric bends.

II. RELATED WORK

There are many algorithms present in literature of extraction of content from document pictures. Note that the camera caught document images, the distortions are usually caused by page skew and the camera view edge. Hence it is vital while extracting text from the content picture, content ought to be horizontal and text should be inline appropriately. Some delegate text extraction techniques are assessed below.

Liu and Yin [2] proposes content line segmentation technique dependent on MST clustering with distant metric learning. Distance metric initially built using supervised learning carried on database sets of connected components. At that point the connected components of content picture are gathered into tree structure, by this content lines are extricated by continuously cutting edges utilizing another hyper volume

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 355-359
Published Online August 2019 in IJEAST (http://www.ijeast.com)

decrease paradigm and direct estimate. This calculation is vigorous to manage different documents written by hand with multi-slanted and somewhat curve content lines. In any case, the calculation neglects to deal with document images with enormous twists.

Beusecom, Keysers and Breuel [9] introduced document cleanup by utilizing page frame detection. Content picture could be cleanout by the help of geometric matching algorithm. This guide to find the genuine page text area, ignoring minor commotion next to the page outskirt. The existing examination strategies for document can deal with non-content noise sensibly good, while content noise speaks to a noteworthy problem for content analysis frameworks. The content clamor may occurred as unwanted content in OCR yield. Along these lines, it is important to expelled a short time later. For identifying and expelling negligible clamor the current record cleanup techniques are utilized. The technique introduced here defeats the confinements of these current strategies.

Mahmoud, & Rasheed [3] introduces few new procedures for slant correction, corresponding to some current methods. It incorporates 2 novel document slope identification calculations dependent on histogram measurements and CC evaluation. The histogram dependent calculation works more proficient for determining slope edge were we dissect lines as peaks and valleys on histogram. The CC evaluation depends on identifying the CCs inside a solitary line and thinks about them as one mass to gauge slope edge.

Hasan and Karam [8] introduced morphological strategy for content extraction from document images. The proposed morphological method is harsh toward commotion, slant and content direction. This is likewise free from ancient rarities they are generally presented by together ideal global thresholding and fixed-estimate square based local thresholding.

Tian and Narasimhan [4] introduce line tracing dependent content line extraction method which may be actualized legitimately on a gray scale picture. The technique consequently identifies and thickly follows content lines in picture. The extraordinary focal points of this strategy are that it does not depend on few clamor sensitive tasks, for example, picture thresholding and text division, and no predefined information regarding to the text dimension, letters in order is required. Be that as it may, the strategy is regularly difficult to decide a fitting step size for line tracing, and it doesn't work in non-literary areas.

Loo and Tan [5] introduced both word as well as sentence extraction strategy for a content picture which may have a broad assortment of content line directions and formats. Their calculation depends on the unpredictable pyramid structure which is useful in combining text to form words and afterward combine words to form sentences.

Bagdanov and Kanai [6] introduced skew estimation algorithm depends on projection profile. The projection profile at a global skew angle of info picture has restricted peak and profound valleys. Hence with the instinct that a flat projection of content will have restricted peak, the level projection profile is determined for slanted picture having conceivable angle θ and greatest normal estimation of straight projection profile at a specific θ is determined and afterward to slant redress, turn picture by θ.

Rasheed & Sarfraj [7] introduced bounding box method it depends on finding the extraordinary corners of a content picture. On the off chance that the four extraordinary focuses structure an ideal rectangle, the wanted point could be effectively decided. 4 final extraordinary points of a picture are gotten through vertical and flat scaling. Bounding Box Algorithm is that if any 2 of 4 points focuses recognized effectively, it will appraise a precise slanted angle. Then again, might be these 2 of 4 corners do not find out effectively as an issue or shortcoming.

III. PROPOSED METHOD

Figure.1 illustrates the different useful functional blocks of proposed method. This method is partitioned into three sections are as per the following.

a) Preprocessing
b) Slope Detection and Correction of skewed text lines
c) Text Extraction

A. Preprocessing

The dataset is a gathering of both shading pictures as well as gray pictures containing skewed content. Dim scale pictures are utilized all things considered for additional preparing. Shaded pictures are changed over into dark pictures to stay away from huge calculations required for color image processing. Preprocessing is a significant advance in picture processing this is utilized to upgrade picture quality. Within a preprocessing, we utilized binarization of dark scale picture this helps in segmenting skewed content lines. Gray picture $I_G(x, y)$ is changed over into segmented picture $I_B(x, y)$. Changing dim picture into segmented picture is completed utilizing following condition. The limit TH is determined by utilizing the data of histogram of gray scale image.

$$I_B(x,y) = \begin{cases} 1, & if\ I_G(x,y) > TH \\ 0, & if\ I_G(x,y) \leq TH \end{cases}$$

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 355-359
Published Online August 2019 in IJEAST (http://www.ijeast.com)

Where are gray scale picture $I_G(x, y)$ and segmented picture $I_B(x, y)$ individually. *TH* is the limit esteem used to change over the gray scale picture into segmented picture.

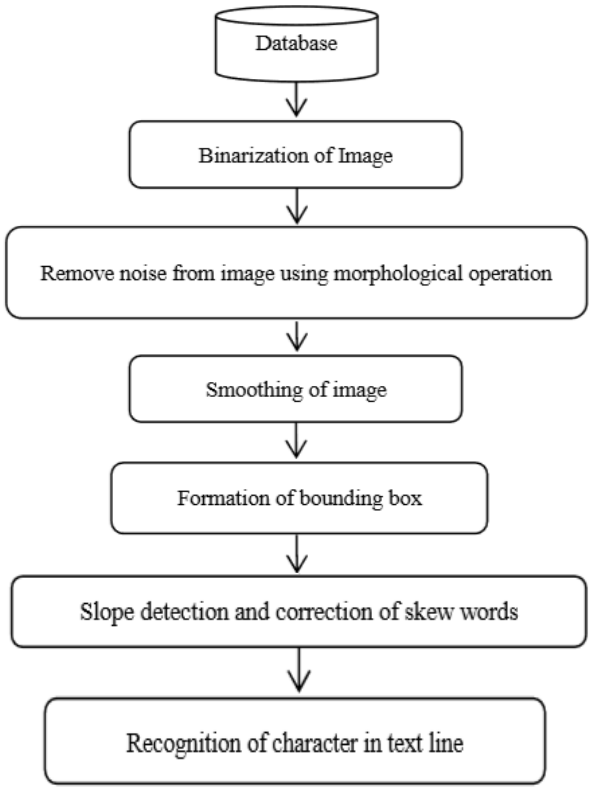

Fig. 1 Proposed method flowchart

In the following stage noise present in segmented picture is expelled utilizing morphological operations. In this progression the little components they are not the piece of the words in skewed text lines are expelled. Morphological operations are some straightforward activities dependent on the image shape. It is typically performed on binary images. It needs two information sources, one is our original picture, second one is called structuring component or kernel which chooses the nature of operation. Two essential morphological administrators are *Erosion and Dilation*.

Smoothing text document image is necessary for legitimate segmentation of the image. OpenCV gives primarily four kinds of blurring procedures. The Bilateral filter is utilized for blurring the content of the picture. Other filters will tend to blur edges. This isn't the situation for the bilateral filter, cv2.bilateralFilter(), which was characterized for, and is profoundly viable at noise removal while saving edges. Picture blurring is accomplished by convolving the document image with a low-pass filter mask.

The example beneath exhibits the utilization of bilateral filtering.

blur = cv2.bilateralFilter(image, 9, 75, 75)

The bounding box is a square shape. This square shape is utilized to recognize the text word utilizing CC. Bounding boxes are conformed to every word utilizing the premises of connected components. This segmentation decides constituents of a picture. It is important to find text locales of document which have printed information and are recognized from figures and designs. From that point onward, the picture is prepared for the text extraction.

B. Slope Detection and Correction of Skew Words

At the point when a text picture is caught by a camera, a little slant is inescapable. This influences the exactness of algorithm for extraction of text. A picture containing a rotated block of text at an obscure angle, we have to correct the text image skew by:

- Determining the angle of rotated text.
- Correcting the text skew by rotating the image.

Here we apply text skew correction algorithms. The objective of our text skew correction algorithm will be to accurately decide the direction and angle of skew, at that point correct for it. We have isolated text in image using thresholding operation, now we would be able to calculate the minimum rotated bounding box that contains the text regions.

Since we have computed the text skew angle, we have to apply an affine transformation to the text image for correcting skew. The rotation matrix is used to play out the real transformation, which is returned by the OpenCV function cv2.getRotationMatrix2D( )

C. Text Extraction

First skewed picture are corrected to a straight image after that recognition of character process is carried out. Character recognition within a document picture is one of the difficult job. Extracting character assumes a significant job for giving data. The characters provide significant data they may be utilized to comprehend the information of a picture. Text picture in the skewed form i.e. picture is rotated with an angle. It is detected first then corrected to a horizontal line using text skew correction algorithm.

Presently, we have horizontal text document in a picture format. It is accomplished by determining the slant word angle and then corrected by turning in that specific angle. After that content extraction is the significant task has been carried. This is accomplished using Optical Character Recognition (OCR). Picture content is comprehensible by OCR method. The OCR

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 355-359
Published Online August 2019 in IJEAST (http://www.ijeast.com)

is utilized for grouping optical patterns related with alphanumeric/different characters.

IV. EXPERIMENTAL RESULTS

Experimental results are performed on text document images. Fig. 2 shows the input database image and binary inverted picture of gray scale image shown in Fig. 3. Fig. 4 shows the bilateral filtered database picture. The rotated text document image is shown in Fig. 5. Fig. 6 shows the result of final step i.e., text extracted from rotated text document image. Performance of this proposed method is determined by calculating the Accuracy (A) as follows.

$A = W_A / W_D$

Where W_A refers to the count of words extracted by this proposed method and W_D refers to the count of words occurs in content picture. Below Table.1 demonstrates the outcome of accuracy calculation.

TABLE I PERFORMANCE MEASUREMENT

Dataset Images	Total count of words in content picture	Count of words extracted by this method	Accuracy(A)
Img 1	72	72	100%
Img 2	80	79	98%
Img 3	94	93	98%
Img 4	77	77	100%
Img 5	79	78	98%
Img 6	51	48	94%
Img 7	56	54	96%

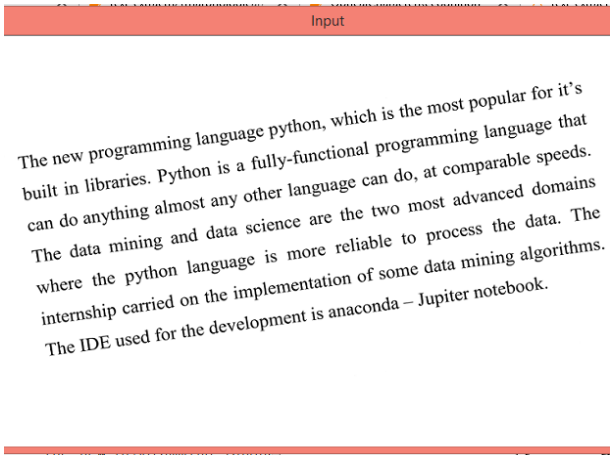

Fig. 2 Original input picture

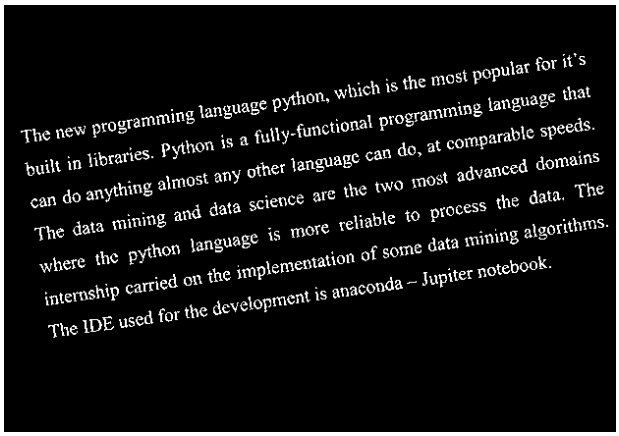

Fig. 3 Binary picture

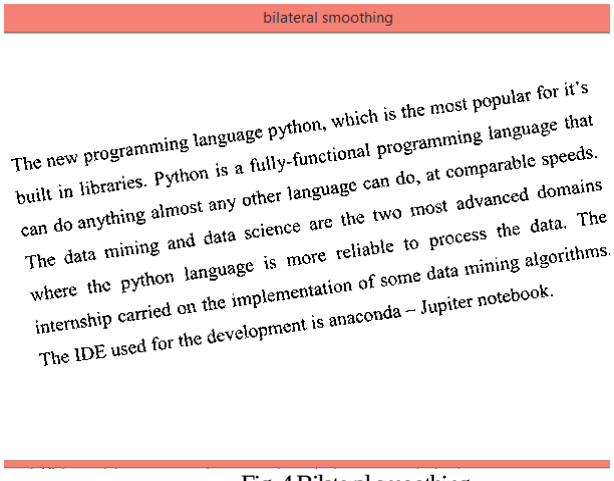

Fig. 4 Bilateral smoothing

International Journal of Engineering Applied Sciences and Technology, 2019
Vol. 4, Issue 4, ISSN No. 2455-2143, Pages 355-359
Published Online August 2019 in IJEAST (http://www.ijeast.com)

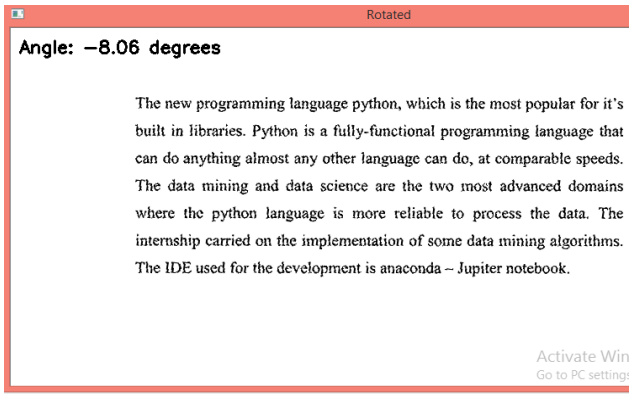

Fig. 5 Diskewed picture

```
--- Start recognize text from image ---
[INFO] angle: -8.062
Angle: -8.06 degrees

The new programming language python, which is the most popular for it's
built in libraries. Python is a fully-functional programming language that
can do anything almost any other language can do, at comparable speeds.
The data mining and data science are the two most advanced domains
where the python language is more reliable to process the data. The
internship carried on the implementation of some data mining algorithms.

The IDE used for the development is anaconda ~ Jupiter notebook.
```

Fig. 6 Text extracted

V. CONCLUSION

In this method, characters extracted from skewed content document images. This method can be applicable to both dark scale picture and shaded picture. This approach computes the exact angle of skew and is computationally very efficient contrasted with existing methods. Using this skew angle rotated the image to correct the skew. After that characters are extracted from content picture utilizing Optical Character Recognition (OCR). The accuracy of this proposed method is around 97.75%.

VI. REFERENCE

[1]. Huang, C., & Zhu, Y. (2009). New Morphological Filtering Algorithm for Image Noise Reduction. 2009 2nd International Congress on Image and Signal Processing. doi:10.1109/cisp.2009.5303495

[2]. F. Yin and C.-L. Liu. Handwritten chinese text line segmentation by clustering with distance metric learning. Pattern Recognition, 42(12):3146–3157, 2009. 1

[3]. Sarfraz, M., Mahmoud, S. A., & Rasheed, Z. (2007). On Skew Estimation and Correction of Text. Computer Graphics, Imaging and Visualisation (CGIV 2007). doi:10.1109/cgiv.2007.63

[4]. Y. Tian and S. G. Narasimhan. Rectification and 3d reconstruction of curved document images. In Computer Vision and Pattern Recognition, Jun 2011. 1, 4

[5]. Loo, P.K., Tan, C.L.: Word and sentence extraction using irregular pyramid. In: Document Analysis Systems V, vol. 2423 of Lecture Notes in Computer Science, pp. 307–318. Springer, Berlin (2002)

[6]. A. Bagdanov and J.Kanai, "Projection profile based Skew Estimation Algorithm for JBIG Compressed Images ", ICDAR pp 401–405, 1997.

[7]. M Sarfraj and Z Rasheed "Skew Estimation and Correction of Text using Bounding Box", Fifth IEEE conference on Computer Graphics, Imaging and Visulization, pp. 259–264, 2008.

[8]. Hasan, Y. M. Y., & Karam, L. J. (2000). Morphological text extraction from images. IEEE Transactions on Image Processing, 9(11), 1978–1983. doi:10.1109/83.877220

[9]. V.Beusekom,J.Keysers and D.Breuel, "Document cleanup using page frame detection",Int. J. Doc. Anal.Recognition,Vol.11, No. 2, pp.81–96, 2008.

[10]. Panwar, S., & Nain, N. (2012). A Novel Approach of Skew Normalization for Handwritten Text Lines and Words. 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems. doi:10.1109/sitis.2012.51

[11]. Shejwal, M. A., & Bharkad, S. D. (2017). Segmentation and extraction of text from curved text lines using image processing approach. 2017 International Conference on Information, Communication, Instrumentation and Control (ICICIC). doi:10.1109/icomicon.2017.8279138