ISSN : 2231-0711

Rajath.A.N et al | IJCSET(www.ijcset.net) | June 2015 | Vol 5, Issue 6,158-161

An Adaptive Approach: Text Line Extraction from Multi-Skewed Hand Written Documents

Rajath.A.N

Assistant Professor, Department of Computer Science and Engg, GSSSIETW, Mysuru

***Abstract*—Advancing technology has made document image processing an important feature in automation of office documentation. Digital filing system save space, paper and printing cost. The problem arises when document to be read is not placed correctly in scanner, which leads to the miss interpretation of document and increases the storage space .This paper deals with extraction of text from those skewed document and proper alignment of those texts.**

***Keywords* —Multi-Skewed, Text Line Extraction, Connected Component Labelling, RLS Algorithm, Horizontal and Vertical Projection.**

I. INTRODUCTION

Organizations are moving at a fast pace from paper to electronic documents. However, large amounts of paper documents inherited from a recent past are still needed. Digitalization of documents appear to be as a bridge over the gap of past and present technologies. Scanners tend to be of widespread use for the digitalization of documents. One of the important problems in this field is that very often documents are not always correctly placed on the flat-bet scanner either manually by operators or by the automatic feeding device. This very frequent problem yields rotated images. For humans, rotated images are unpleasant for visualization and introduce extra difficulty in text reading. For machine processing, image skew brings a number of problems that range from needing extra space for storage to making more error prone the recognition and transcription of the image by automatic OCR tools. These reasons make skew detection and correction phases a common place in any environment for document processing.

Very frequently the digitalization process of documents produce images rotated of small angles in relation to the original image axis. The skew introduced makes more difficult the visualization of images by human users. Besides that, 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. Thus, skew correction is an important part of any document processing system being a matter of concern of researchers for almost two decades now. The search for faster and good quality solutions to this problem is still on.

Text Line Extraction from optically scanned document images is one of the major problems of optical character recognition (OCR)*/handwritten* text. Appearance of *skewed lines* in the text makes the problem complex. The problem becomes compounded if the lines in a text image are skewed with different orientations. Such lines are called *multi-skewed lines*. Appearance of multi-skewed lines in text is common to both printed and handwritten texts for various reasons. Lines in a text image get skewed mainly for two reasons. *Firstly*, a few degrees of misalignment of the document with respect to the scanner or copier bed is unavoidable at the time of scanning. For text images of *printed Roman script*, most skew correction. Text line extraction from optically scanned document images is one of the major problems of and line extraction techniques predominantly deal with a single skew angle for an entire document page For dealing with *handwritten text* of *Roman script*, principally determine a skew angle from a page of text lines on the basis of the *base lines* of the text words before skew correction. To extract text lines and words from document images of *handwritten English text* lines **uses** *horizontal and vertical histogram* values of the same. Document images and then performing correction to these lines becomes a trivial problem as in [1].

II. REVIEW OF OTHER METHODS

Numerous methods have been proposed to extract text lines from Hand written Documents. These methods can be classified into the following six major categories: projection profile based, smearing methods, Hough transformation based, clustering or grouping methods, repulsive attractive methods that uses energy minimization systems, and stochastic methods which use stochastic learning algorithms [3]. Some of these methods deal with specific languages such as Chinese or Arabic scripts. Yin and Liu [4] presented a clustering method using Minimal Spanning Trees (MST) to extract lines from both Chinese and Latin-based documents. The results show that their method performs well on multi-skewed and curved text lines in handwritten documents. Kumar et al [3] proposed a graph based approach to extract text lines from Arabic unconstrained handwritten documents.

Their approach is fast since it is based on connected components. However, it does not perform well in presence of touching components. Shi et al [5] extracted Arabic handwritten text lines by applying a direction filter; then, an adaptive thresholding algorithm was applied to adaptive local connectivity maps to form connected components. Finally, they extracted lines by grouping the connected components using a clustering algorithm. Many script independent text line extraction algorithms have been proposed. Bukhari et al proposed [6] a script independent line extraction algorithm that uses ridges over smoothed images to estimate the central line of text lines parts. An active contour was applied over ridges to segment the lines. Li et al [6] proposed a script independent algorithm that

uses the level set for line segmentation. These two approaches perform well on Arabic handwritten documents. Nevertheless, they suffer from the high computational cost. Ziaratban and Faez [8] used a bottom up algorithm that segments the document into adaptive blocks; then, the skew of each block is estimated. Three parameters were defined to adapt the method to different writers. Different techniques were combined to produce better results or to adapt to scripts of special characteristics. Ouwayed et al [9] implemented a text extraction system using various local techniques including snakes (Repulsive Attractive Methods) to create a contour that segments the lines into local zones. Then, the orientation of each zone was detected using
Special projection profile histograms.

III. Proposed Algorithm

In the author's previous work [1], a hypothetical assumption of water flow based Text line extraction was presented. We propose in this section using RLSA, horizontal and vertical projection algorithm. As shown in figure(1), the extracted features using a recursive algorithm is implemented for connected component labelling (CCL) operation and during this step main geometrical property of Text extraction such as aspect ratio is computed.
The proposed system consists of four main stages: (1) Image Acquisition. (2) Removal of unwanted region by applying Morphological Operation. (3) Detecting Skew Region using Horizontal and Vertical projection algorithm. (4) Segmentation. (5) Skew Correction using RLS Algorithm. General scheme for Text line Extraction is shown in figure (1).

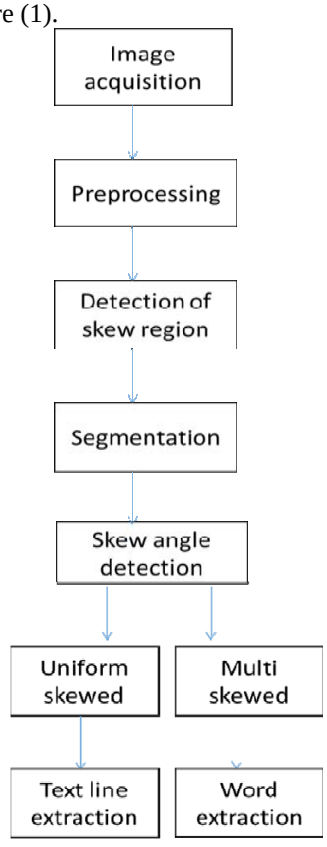


Figure (1): Proposed Model of Text Line Extraction

Typical color images are represented as RGB images. Using the 8-bit monochrome standard as a model, the corresponding color image would have 24-bits per pixel – 8-bits for each of the three color bands (red, green, and blue) but a grey scale image (referred to as monochrome) only contains the brightness information but not color information. In order to improve image processing speed, scanned document image (RGB) is converted to grey-level image.

IV. Text Line Extraction Module

For Uniform and Multi-Skewed Handwritten Documents

A. Image Acquisition

It is the first process in image processing. This could be as simple as being given an image that is already in digital form. The input image may be in .jpg or .png format. Generally, the image acquisition storage involves preprocessing, such as scaling.

B. Removal of Noise

Preprocessing is a method of enhancing the image for better feature extraction. This can be considered as one step which is generally prepares the platform ready for rigorous Pattern Recognition and Image Processing procedures. The choice of preprocessing method to be adopted on a document image depends on the type of application for which the image is used. There are many techniques that are generally available to accomplish preprocessing on images. However several experiments suggest that preprocessing methods have got to be customized to suit the requirements.
Any segmentation method used for Text line extraction image, requires conditioned image input of the image, which implies that the image should be noise free. The preprocessing stage is the second stage in proposed system of text line extraction. This stage is necessary to enhance the quality of an image by noise removal, as well as sharpening.

C. Detection of Skew Region

The next step is skew region based on assumption that Aspect ratio of Skewed region will be more than 10% of original document image, skew region is detected.

D. Segmentation

This procedure partition an image into its constituent parts or objects. In general, autonomous segmentation is one of the most difficult tasks in image processing. A rugged segmentation procedure brings the process a long way toward successful solution of imaging problems that require objects to be identified individually. In general, the more accurate the segmentation, the more likely recognition is to succeed.
Next step of proposed algorithm is labeling the connected components. Connected components labeling (CCL) is a well-known technique in image processing that scans an image (binary and gray-level) and labels its pixels into components based on pixel connectivity. Once all groups have been determined, each pixel is labeled with a value according to the component to which it was assigned. The skew region is detected using Horizontal Projection Algorithm for Uniform Handwritten Documents.

The Algorithm is as follows
Input : Any Text image
Output: Segmented text

1) Apply Horizontal projection and take the row-wise sum of black pixels.
2) Plot a row-wise sum of black pixels using a histogram.
3) Count the number of continuous sequence of black pixels and store them in an array.
4) Find the biggest value in the array obtained in step 2.
5) Find the position of the biggest value and finally crop that region.

But for the Multi-Skewed hand written documents we use Vertical Projection Algorithm.

The algorithm is as follows:
Input : Any text image
Output: Segmented text and non-text portions
1) Apply vertical projection and obtain the column-wise sum of black pixels.
2) Plot the sum of black pixels using a histogram.
3) Segment the two portions based on a continuous sequence of column-wise white pixels.
4) Identify which is the text portion and which is the non-text portion based on the following two issues:

a. Based on threshold value.
b. Taking the difference of adjacent values stored in array.

E. Skew Angle Detection

In the proposed method, a recursive algorithm RLSA is implemented for skew detection. Before document can be properly analyzed for character recognition, its skew ness has to be detected. The angle of skew is obtained by applying the algorithm. The algorithm is as follows:
The basic RLSA is applied to a binary sequence

Input: logical image
Output: Transformed logical image

The algorithm transforms a binary input sequence into an output sequence according to the following rules:

1) 0's in input are changed to 1's in output if the number of adjacent 0's in input is less than or equal to a predefined limit.
2) 1's in inputs are unchanged in output.
3) When applied to pattern arrays, the RLSA has the effect of linking together neighbouring black areas that are separated by less than limit pixels.

Based on these modules the next step, Text Line is Extraction is done for Uniformly Skewed Handwritten Documents and Word Extraction is done for Multi-Skewed Handwritten Documents.

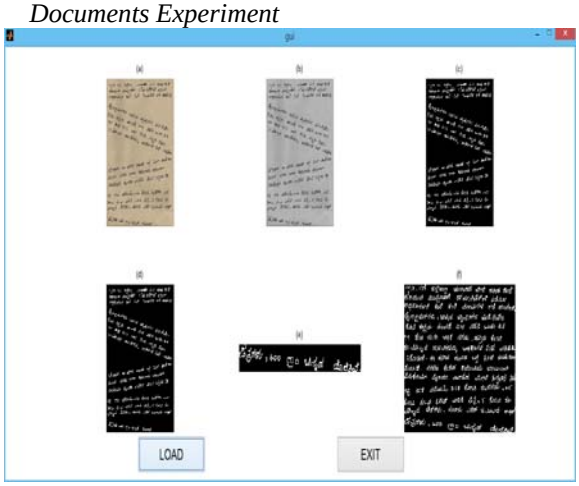
V. Experimental Results

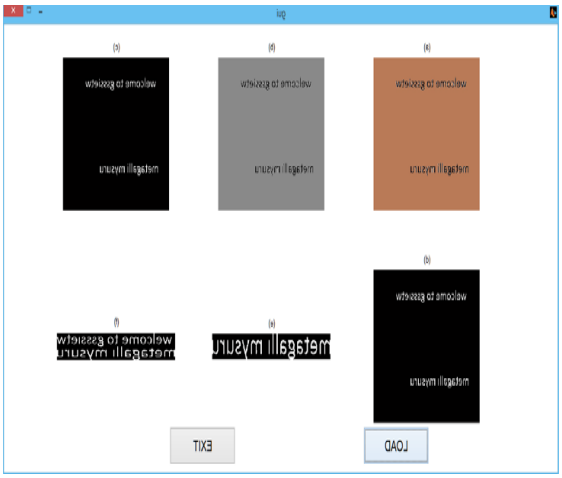
All experiments were done on Pentium Dual Core 2.10 GHz with 2GB RAM under MATLAB R2013a environment. In the experiments, 50 images were employed which is scanned handwritten document .All these documents are scanned using OCR. The satisfactory result has been obtained: the success of text line and skew correction rate of document is up to 85%. This method take images contain text line, it may be uniform skewed line or multi skewed line.

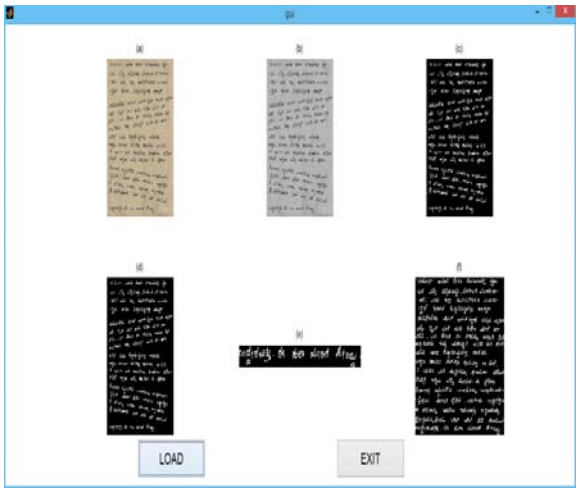
The following figures show the some successful results for Text Line Extraction.

For uniform skewed document (a) Initial Image (b) Gray Scale image (c) Black and White Image (d) Pre-processed Image (e) Segmented Text (f) Corrected Image

A) Successful Text Line Extraction for Uniform Skewed Documents Experiment

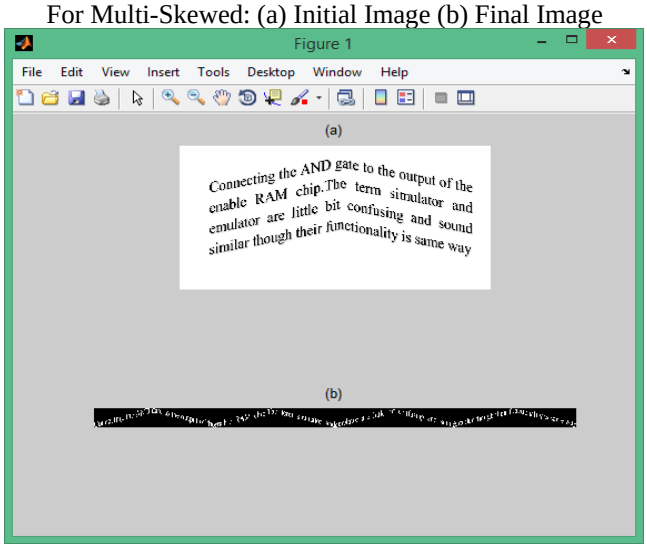


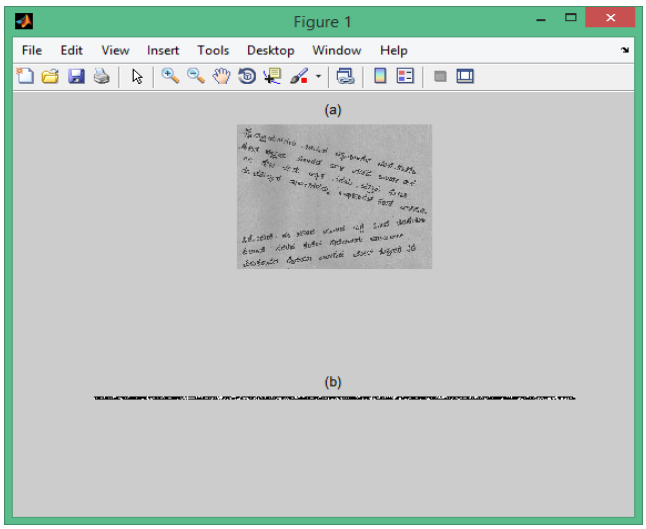


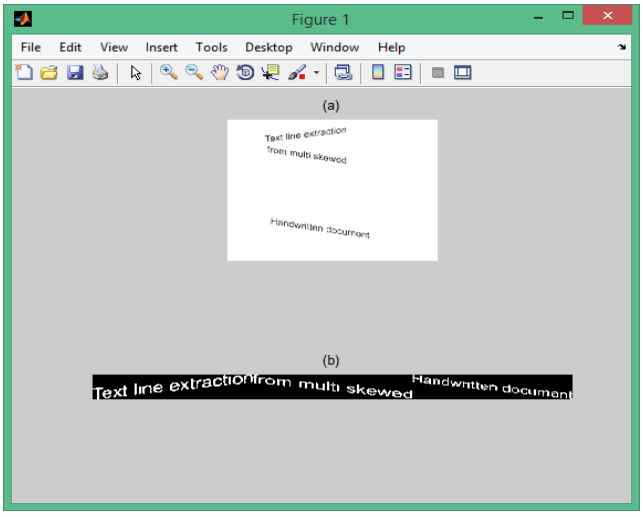


(B) Successful Text Line Extraction for Multi-Skewed Documents Experiment

For Multi-Skewed: (a) Initial Image (b) Final Image







VI. Conclusion

In this paper, we proposed method for text line extraction from multi skewed document. Experimental results show that extraction of line has done for both uniform and multi skewed document and Skew correction of uniformly skewed document. However proposed cannot do skew correction for multi skewed document because word extraction is doing randomly in multi skewed document. so Still this works needs improvement for multi skewed image

References

[1] S. Basu, C. Chaudhuri, M. Kundu, M. Nasipuri∗, D.K. Basu Text line extraction from multi-skewed handwritten documents

[2] Muna Khayyat, Louisa Lam and Ching Y. Suen, Fei Yin and Cheng-Lin Liu, Arabic Handwritten Text Line Extraction by Applying an Adaptive Mask to Morphological Dilation

[3] Jayant Kumar, Le Kang, David Doermann and Wael Abd-Almageed, "Handwritten Arabic Text Line Segmentation Using Affinity Propagation," *Document Analyis Systems*, pp. 135-142,2010.

[4] Fei Yin and Cheng-Lin Liu, "Handwritten Text Line Extraction Based on Minnimum Spanning Tree Clustering," *Pattern Recognition*, vol. 42, issue 12, pp. 3169-3183, 2009.

[5] Zhixin Shi, Srirangaraj Setlur, and Venu Govindaraju, "A Steerable Directional Local Profile Technique for Extraction of Handwritten Arabic Text Lines," *In Preceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009)*, vol. 1, pp.176,180,2009.

[6] Syed Saqib Bukhari, Faisal Shafait and Thomas M. Breuel, "Script-Independent Handwritten Textlines Segmentation using Active Contours," *In Preceedings of the 10th International Conference on Document Analysis and Recognition (ICDAR2009)*, vol. 2, pp. 446-450, 2009.

[7] Li Yi, Yefeng Zheng, David Doermann, and Stefan Jaege, "Script-Independent Text Line Segmentation in Freestyle Handwritten Documents," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, val. 30, no. 8, pp. 1313-1329,2008.

[8] Majid Ziaratban and Karim Faez, "An Adaptive Script-Independent Block-Based Text Line Extraction," *In Procceedings of the 20th International Conference on Pattern recognition(ICPR '10)*, pp. 249-252, 2010.

[9] Nazih Ouwayed, Abdel Bela, and Francois Auger, "GeneralText Line Extraction Approach based on Locally OrientationEstimation," *In Procceedings of the 17th Document Recognition and Retrievall Conference (DRR 2010)*, 2010.