Script Recovery from Scanned Document Image

[1] *Dr.Srinivasan. K.S.*
Professor, Department of Electronics and Communication Engineering Easwari Engineering College, Chennai 600089, India

[2] *Rajesh Kumar.T*
Asst. Professor (Sr.G), Department of Electronics and Communication Engineering
Easwari Engineering College, Chennai 600089, India

Abstract: Document digitization with scanner in text document images which have distortions that deteriorate the quality of the document. We propose a goal-oriented rectification methodology to recover the document from distorted document image. Our approach relies upon a coarse-to-fine strategy. First, a coarse rectification is accomplished with the projection of the curved surface on the plane which is guided by the textual content's appearance in the document image while incorporating a transformation which does not depend on specific model primitives or scanner setup parameters. Secondly, normalization is applied on the word level aiming to restore all the local distortions of the document image. Experimental results on various document images with a variety of distortions demonstrate the robustness and effectiveness of the proposed rectification methodology that improves OCR accuracy. It finds its application widely in de-warping of document images, images captured from sculptures, from cursive handwritten text, text from palm leaves and so on...

I. Introduction

Document image acquisition by a flatbed scanner or a digital camera often results in several unavoidable image distortions (see Fig. 1) due to the form of printed material (e.g. bounded volumes), the camera setup or environmental condition
(e.g., humidity that causes page shrinking). Text distortions not only reduce document readability but also affect the performance of subsequent processing such as clarity and optical character recognition (OCR).
There are many different techniques that have been proposed for text document image recovery .They can be classified into two main categories based on
1) 3-D document image processing
2) 2-D document image processing.
On the other hand, techniques in the latter category do not depend on auxiliary hardware or prior information but they only rely on 2-D information.
In this paper, we propose a goal-oriented algorithm to compensate for undesirable distortions of document images captured by flatbed scanners. The proposed technique is directly applied to the 2-D image space without any dependence to auxiliary hardware or prior information. It first detects words and text lines to rectify the document image in a coarse scale and then further normalize individual words in finer detail using baseline correction. Experimental results on several document images with a variety of distortions show that the proposed method produces rectified images that give a significant boost in OCR performance. The proposed rectification methodology is shown in Fig. 2.

I. Proposed System

The proposed system uses only scanned document images without any dependence on auxiliary hardware or prior knowledge. It involves two stages. First, is preprocessing. Secondly goal oriented algorithm which consist of coarse rectification and fine rectification. The purpose of coarse rectification is to remove large distortions in document image. This process is applied after the completion of preprocessing. The purpose of fine rectification is to remove the local distortions in order to achieve an optimal rectification of the document image. Thus the output from the fine rectification process is the recovered script.

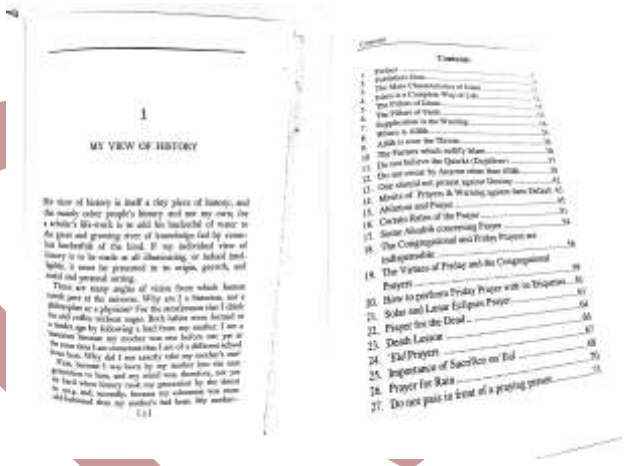


Fig .1 These are some of the Distortions occurs in the scanned document image

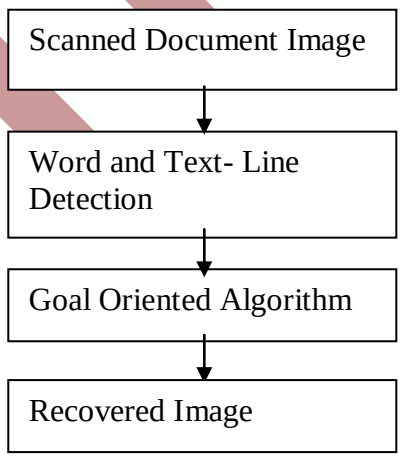
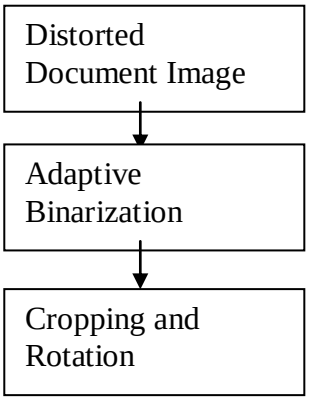


Fig .2 Methodology



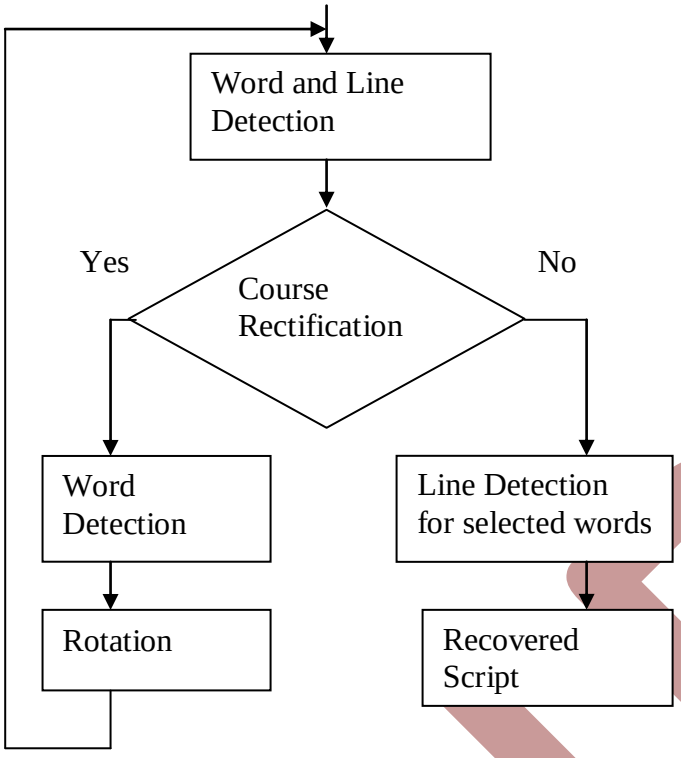


Fig3. Flowchart for Methodology

A. Preprocessing:

In this system, first we apply a preprocessing step at the original distorted document image which consists of an a noise removal, adaptive binarization, border removal, cropping and rotation. The noise removal step in preprocessing is applied only if the image has noise such as salt and pepper noise, Gaussian noise etc. otherwise image can be directly proceed with adaptive binarization.

Goal Oriented Algorithm:

As mentioned earlier, it consist of two steps
i.e.
1. Coarse Rectification
2. Fine Rectification.

1. Coarse Rectification:

In this step, a word detection technique for distorted document images is introduced. In this the two curved line segments at the top and bottom area, upon which the mapping from the projection of a surface to a rectangle is applied. The result of coarse rectification is shown in fig 3. In this stage, we do not care whether the text line detection is accurate, since we just need some specific points in order to model the curved surface projection on the plane and we will not use each detected text line to correct the distortions of the document. For the sake of clarity, we will report on some possible errors of the detection step.

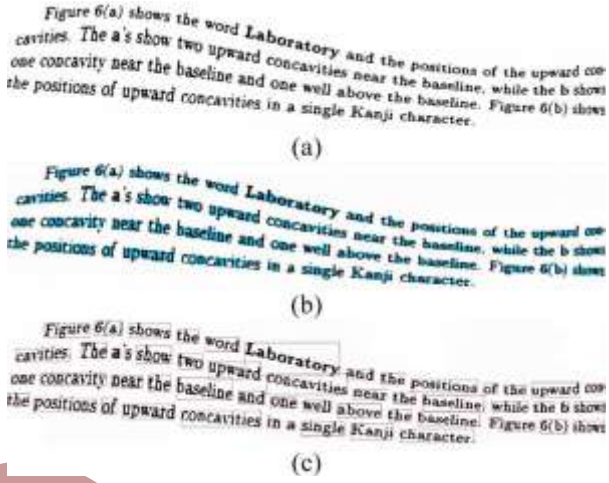


Fig4. (a) Original input image (b) Selection or Highlighting the text in the image (c) Detected text

2. Fine Rectification:

In this step all detected words are rotated followed by the line detection in order to obtain the final rectified document image. First, we proceed with the rotation of the words which is shown in fig 5 (a). After the completion of rotation, we proceed with the line detection. In this all the text lines are detected. (See fig 4(b)). Recovered image is shown in fig 4 (d).

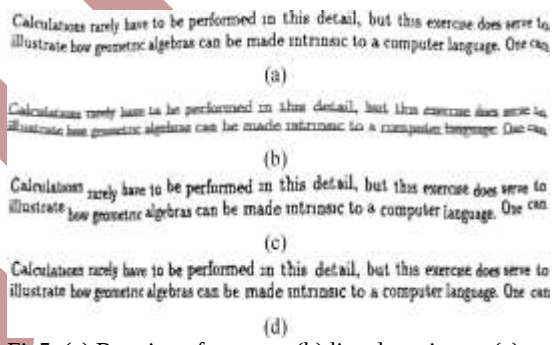


Fig5. (a) Rotation of text (b) line detection (c) Image after line detection (d) Recovered output.

II. Conclusion

In this paper we have presented a complete algorithm to remove undesirable distortions present in document image which does not requires any auxiliary hardware or information. Experimental results on several distorted document images show that the proposed method produces rectified images with high OCR.

III. Reference

[1] J. Liang, D. Doermann, and H. Li, "Camera-based analysis of text and documents: A survey," *Int. J. Document Analysis and Recognition*, vol.7, no. 2-3, pp. 84–104, 2005.

[2] L. Zhang, Y. Zhang, and C. L. Tan, "An improved physically-based method for geometric restoration of distorted document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 4, pp. 728–734, Apr.2008.

[3] A. Ulges, C. H. Lampert, and T. Breuel, "Document capture using stereo vision," in *Proc. ACM Symp. Document Eng.*, Milwaukee, WI, 2004, pp. 198–200.

[4] A. Yamashita, A. Kawarago, T. Kaneko, and K. T. Miura, "Shape reconstruction and image restoration for non-flat surfaces of document with a stereo vision system," in *Proc. 17th Int. Conf. Pattern Recognit.*, Cambridge, U.K., 2004, pp. 482–485.

[5] M. S. Brown and W. B. Seales, "Image restoration of arbitrarily warped documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 10, pp. 1295–1306, Oct. 2004.

CIR