World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:11, No:1, 2017

Degraded Document Analysis and Extraction of Original Text Document: An Approach without Optical Character Recognition

L. Hamsaveni, Navya Prakash, Suresha

Open Science Index, Computer and Information Engineering Vol:11, No:1, 2017 publications.waset.org/10006489.pdf

***Abstract*—**Document Image Analysis recognizes text and graphics in documents acquired as images. An approach without Optical Character Recognition (OCR) for degraded document image analysis has been adopted in this paper. The technique involves document imaging methods such as Image Fusing and Speeded Up Robust Features (SURF) Detection to identify and extract the degraded regions from a set of document images to obtain an original document with complete information. In case, degraded document image captured is skewed, it has to be straightened (deskew) to perform further process. A special format of image storing known as YCbCr is used as a tool to convert the Grayscale image to RGB image format. The presented algorithm is tested on various types of degraded documents such as printed documents, handwritten documents, old script documents and handwritten image sketches in documents. The purpose of this research is to obtain an original document for a given set of degraded documents of the same source.

***Keywords*—**Grayscale image format, image fusing, SURF detection, YCbCr image format.

I. Introduction

THE information extraction from any source such as a document image or video, is performed to bring all information into the digital world [1]. Today's world is digitized, so that information on paper is available and easy to access through online means such as web media. The digitization of information from paper is obtained by scanning those pages and preserving them as images. Then, images of the text documents are processed using OCR or other Image Processing techniques [1]. The technique involves text detection, recognition, extraction and text matching methods. The text detection from a degraded text document image is a confronting task. The detected text is to be extracted to study in detail about the degradation region of that text document image.

Our work deals with two types of document images: The handwritten document images and the scanned/printed document images. There are many methods that have been proposed to detect and extract text information from a document image [5]. The methods are also designed to detect and obtain the graphics or images in the document image [5]. Handwritten degraded document images are much more challenging to detect the missing text without OCR technique. In our work, we concentrate on detection of the text missing from set of degraded text documents of same source. The detected text is processed to extract the actual text so as to obtain an original document image from its source of two or more input degraded document images. As a method to correct image orientation (deskew) input images to the algorithm is also presented. Conversion of Grayscale image format to RGB image format using YCbCr image format is also applied in this system.

The next section of the paper, Section II details about proposed system that has been enhanced using the existing system. In Section III, we discuss the results obtained on various data sets. Section IV concludes our work with the merits and demerits of the presented algorithm.

II. Proposed Work

An OCR-less approach is designed and developed to extract text from a set of degraded documents [3], [4]. The image acquisition is used to take the input set of degraded document images of the same source. Image processing techniques such as image subtraction, image fusing, detecting surf features, extraction and matching of surf features, conversion of gray scale image to RGB image using luminance comparison and normalization methods are used.

De-skewing of images is also programmed to detect and rectify the scale and angle of skew in the input document images. Image Quality Metrics such as Mean Square Error (MSE), Peak Signal to Noise Ratio (PSNR in dB), Normalized Cross-Correlation (NCC), Average Difference (AD), Structural Content (SC), Maximum Difference (MD) and Normalized Absolute Error (NAE) are used to evaluate the proposed algorithm [3], [4].

Fig. 1 represents the architectural diagram of the existing system. We enhance this system to obtain our presented system. The system takes two or more sets of degraded document images as input. The processing system consists (i) Reading the images. (ii) Degraded text extraction: This unit performs image subtraction finding absolute difference for all input images. This helps to detect the degraded text region in all the input document images. (iii) Fusing resultant images. The result of the absolute difference is fused with the respective input images. This helps in filling the degraded text region with the appropriate text in the document image to obtain an original document image. (iv) Deskewing image:

Navya Prakash , M. Tech in Computer Science and Technology from DoS in CS, UoM, MGM, Mysuru, Karnataka, India (phone: +91-9538704059; e-mail: navyaprakash040@gmail.com).

L. Hamsaveni, Professor, and Dr. Suresha, Professor, are with DoS in CS, UoM, MGM, Mysuru, Karnataka, India (e-mail: hamsa1367@gmail.com, sureshabm@yahoo.co.in).

This unit is used to detect and rectify input image orientation, if the orientation is skewed then it is straightened. It helps to deskew the input image with accurate skew scale value and skew angle with the direction of skewing either counter-clockwise or clockwise. (v) Image format conversion deals with conversion of obtained original document image to RGB format, only if the input images are in RGB format. During the fusing phase in the processing unit, the resultant image obtained is in grayscale image format; this has to be converted into RGB image format. (v) Image quality metrics deals with estimating the proposed method. The major seven metrics are considered and used to find the accuracy of the proposed system. The output will be an original document in an image format obtained by fusing all degraded regions [3], [4].

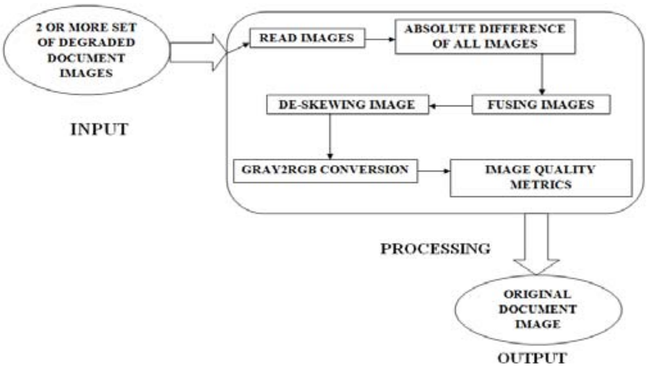


Fig. 1 Existing System Architecture

To detect degraded region in the set of degraded document image; input images are fed into the system. The algorithm finds the absolute difference of all input images to extract the degraded text region. Image fusing technique is used to consolidate images. The fused single original document image is obtained in the Grayscale image format [3], [4].

To deskew input set of degraded documents images; the algorithm detects SURF feature for input images. Feature extraction is performed [2]. Matching image features from the skewed and deskewed images are used to pair the pixel index of images to differentiate original points and distorted points in images. Estimation of geometric transformation in the skewed input image with respect to the deskewed input image is performed. SURF finds the distorted scaling value and the skew angle value with the direction of skew either counter-clockwise or clockwise with the help of estimated geometric transformation [3], [4].

To convert Grayscale image format to RGB image format, a special intermediate image format is used named YCbCr image format. The algorithm converts input images to YCbCr format. It performs normalization for both the images. It finally compares the luminance of both the images. The algorithm copies the luminance values of the input RGB image (initial degraded document) to the input fused grayscale original image. Conversion of resultant image from YCbCr to RGB format takes place to obtain the original document [3], [4].

To measure image quality metrics; algorithm reads the initial input image (degraded document) and original image (fused RGB resultant image) for comparing quality metrics between them. MSE is measured using (1). To find the error in the images, image subtraction technique is used. PSNR is measured using the MSE value using (2). NCC is measured using (3). AD is measured using (4). SC is measured using (5). MD is measured as in (6). NAE is measured as in (7).

MSE= sum (sum (error.* error)) / size (initial input image) (1)

If (MSE>0)

PSNR = 10*log (255*255/MSE) / log (10) (2)

Or else PSNR value is 99.

NCC= sum (sum (initial input image.* result image)) / sum (sum (initial input image.* initial input image)) (3)

AD= sum (sum (error)) / size (initial input image) (4)

SC=sum (sum (initial input image.* initial input image)) / sum (sum (result image.* result image)) (5)

MD=max (max (error)) (6)

NAE= Find sum (sum (abs (error))) / sum (sum (initial input image)) (7)

III. Results

Fig. 2 Input RGB printed degraded document images

Fig. 2 represents the three different printed degraded set of document images of the same source. Fig. 3 represents the absolute differences of the degraded set of document images. Fig. 4 represents the original document image obtained in grayscale image format after fusing all the intermediate results of absolute difference with the input degraded document images. Fig. 5 represents the skewed input degraded document image. Fig. 6 represents the SURF detection, extraction and matching to deskew the input degraded document image. Fig. 7 represents the skewed input image and the deskewed result image obtained using SURF of skewed image with a reference image. Fig. 8 represents grayscale to RGB converted image using YCbCr image format.

Fig. 3 Intermediate stage of proposed algorithm

Fig. 4 Grayscale original document image

Fig. 5 Skewed input degraded document image

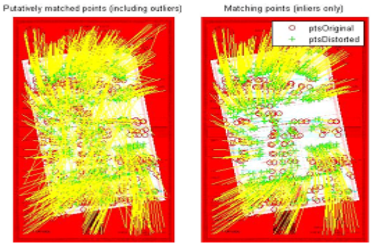

Putatively matched points (including outliers)
Matching points (inliers only)
ptsOriginal
ptsDistorted

Fig. 6 SURF detection, extraction and matching

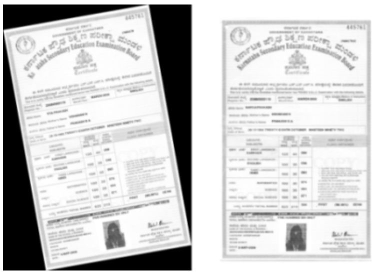


Fig. 7 Skewed and Deskewed image

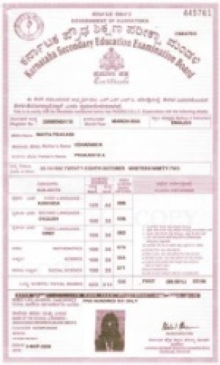
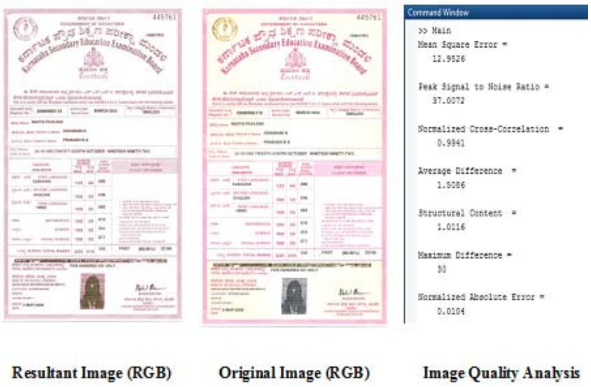


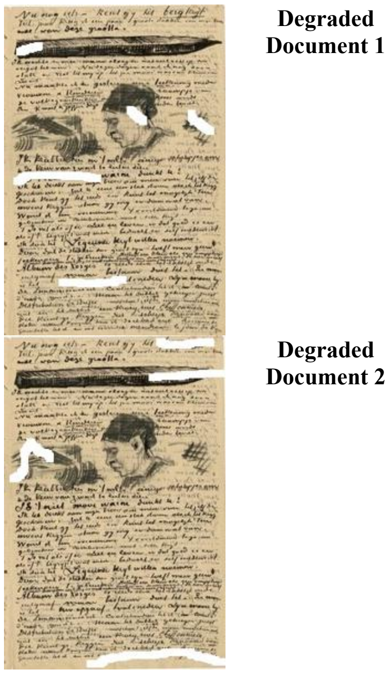
Fig. 8 Grayscale to RGB converted image



```
>> Main
Mean Square Error =
   12.9526

Peak Signal to Noise Ratio =
   37.0072

Normalized Cross-Correlation  =
    0.9941

Average Difference  =
    1.5086

Structural Content  =
    1.0116

Maximum Difference =
    30

Normalized Absolute Error =
    0.0104
```

Fig. 9 Image Quality Metric Values

Fig. 9 represents the values of various image quality metrics obtained. Fig. 10 represents the set of degraded handwritten document images of the same source. The handwritten document is degraded with text erosion and the handwritten sketch in the document is also degraded. The algorithm processes set of input degraded documents to extract original document. Fig. 11 represents the degraded region in the input documents. Finding these degraded regions helps us to recover the original document. The degraded region is obtained using the absolute difference of the input documents. Fig. 12 represents the original document in grayscale image format. The original document in this stage is obtained by image fusing method in the algorithm. Fig. 13 represents the deskewing of the input degraded documents using SURF detection [2].



Degraded Document 1

Degraded Document 2

Fig. 10 Input RGB handwritten degraded documents

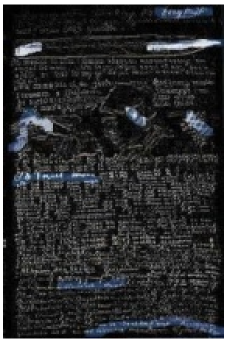


Fig. 11 Intermediate stage

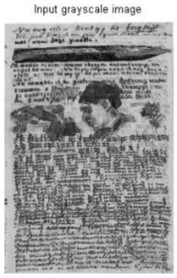

Input grayscale image

Fig. 12 Grayscale input image

Fig. 14 represents the original document in RGB image format. While finding the degraded region, the document is converted into grayscale image format for further processing. After deskewing phase, this grayscale image format is converted to RGB image format using YCbCr image format.

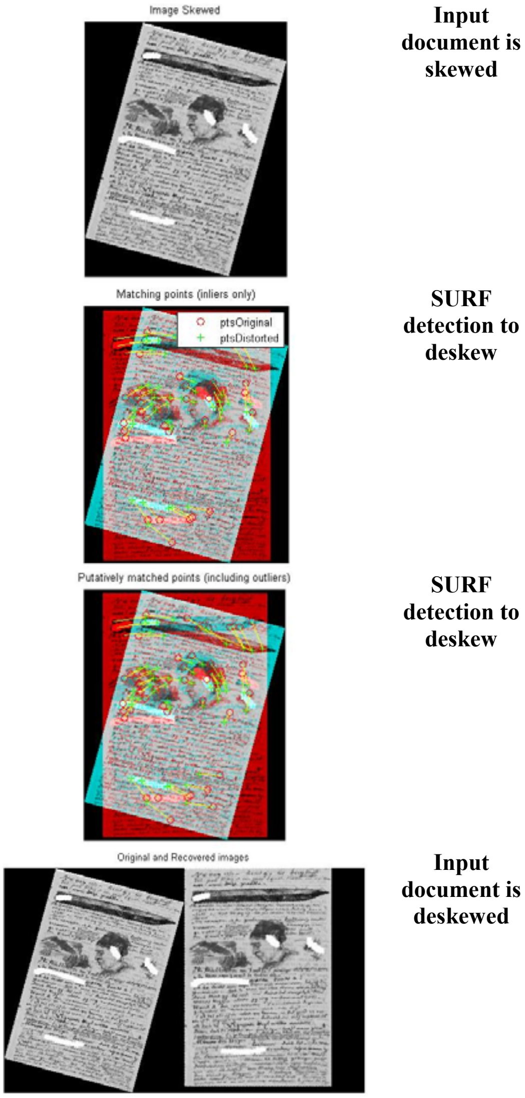


Image Skewed
Matching points (inliers only)
ptsOriginal
ptsDistorted
Putatively matched points (including outliers)
Original and Recovered images

Input document is skewed

SURF detection to deskew

SURF detection to deskew

Input document is deskewed

Fig. 13 Deskewing of degraded documents using SURF

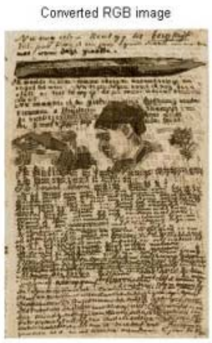

Converted RGB image

Fig. 14 Original Document in RGB image format

IV. Conclusion

An OCR-less approach is designed and developed to extract text from a set of degraded documents. The document images are subtracted from each other to find the degraded text region along with the text information that has been degraded in the image. It is further processed to find the composite of them by fusing the absolute difference results with the input degraded document images to obtain the original document image in Grayscale format. The deskewing method is also implemented to detect the skew of the input image, if any, and to find the skew scaling factor value with the skew angle value with its respective direction of skew, which helps to rectify/deskew the input image. Image Quality Metrics are used to evaluate our work of obtaining original text document image from the set of degraded document images from the same source [3], [4].

The proposed algorithm executes on printed degraded text documents, handwritten degraded text documents, graphics in the form of degraded handwritten sketches in document and degraded animated images in document. In future, the same can be designed to detect and rectify degraded video files.

REFERENCES

[1] Lawrence O'Gorman, Rangachar Kasturi, "Document Image Analysis", IEEE Computer Society Executive Briefings, ISBN 0-8186-7802-X, 2009.

[2] Herbert Bay, Tinne Tuytelaars, Luc Van Gool, "SURF: Speeded Up Robust Features", European Conference on Computer Vision, 2006.

[3] Navya Prakash, L. Hamsaveni, Prof. Dr. Suresha, "Extraction of Original Text Document from a Set of Degraded Text Documents from the Same Source", IJATSCE Vol. 5, No. 4, July-August 2016, ISSN 2278-3091.

[4] Navya Prakash, L. Hamsaveni, Prof. Dr. Suresha, "Extraction of Original Text Document from a Set of Degraded Text Documents from the Same Source", 4th International Conference on Computing, Engineering and Information Technology, Bangalore, 2016.

[5] A.S. Kavitha, P. Shivakumara, G.H. Kumar, Tong Lu, "Text Segmentation in Degraded Historical Document Images", Egyptian Informatics Journal, 2016.

Open Science Index, Computer and Information Engineering Vol:11, No:1, 2017 publications.waset.org/10006489.pdf