ISSN 2278-3091
Volume 5, No.4, July – August 2016
International Journal of Advanced Trends in Computer Science and Engineering
Available Online at http://www.warse.org/IJATCSE/static/pdf/file/ijatcse04542016.pdf

Extraction of Original Text Document from a Set of Degraded Text Documents from the Same Source

Navya Prakash[1], L. Hamsaveni[2], Dr. Suresha[3]
[1]Student, former IEEE student member, India, navyaprakash040@gmail.com
[2]Asst. Prof., DoS in CS, UoM, MGM, India, hamsa1367@gmail.com
[3] Prof., DoS in CS, UoM, MGM, India sureshabm@yahoo.co.in

Abstract : Information extraction is the task of extracting structured data from a degraded document. It includes data extraction such as text, image or graphics from the sources such as an image, video or documents. Text detection and extraction from the degraded document finds application in wide range of study. In this paper, an Optical Character Recognition less (OCR-less) method of obtaining an original text document from a set of degraded text documents is proposed. It involves detection and extraction of text from a set of degraded text documents which belongs to the same source. The degraded text documents are treated as images and are input to our proposed algorithm. It involves Image Processing techniques such as image subtraction and image fusing. Deskewing of input degraded text document images are also proposed by detecting Speeded Up Robust Features (SURF) from the images. A methodology to convert Grayscale image format to RGB image format is also developed using YCbCr image format to obtain an original document in the RGB format same as the input degraded RGB text document images.

Key words : Image Fusing, SURF, YCbCr, RGB, Skew.

1. INTRODUCTION

The information extraction from any sources such as a document image or video is performed to bring all information into the digital world. Today's world is digitized, so that the information on paper is available easily to access through online means such as web media. Digitization of information on the papers is obtained by scanning those papers and preserving them as images. Then, images of the text documents are processed using OCR or other Image Processing techniques. The technique involves text detection, extraction, text matching and many other relevant procedures. The text detection from a degraded text document image is a confronting work. The detected text is to be extracted to study in detail about the degradation of that text document image. If an efficient method exists to obtain the text from a degraded document image then it would be a boon to the archaeologists for their enormous survey.

The work so far deals with two types of document images, they are: the handwritten document images as well as the scanned/printed document images. There are many methods that have been proposed to detect and extract text information from a document image. The methods are also designed to detect and obtain the graphics or images in the document image. But, there are not many works related to the extraction information from the degraded printed document images. Handwritten degraded document images are much more challenging to detect the missing text.

In our work, we concentrate on detection of the text missing from a set of degraded text documents of a 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 degraded document images. A method to deskew input images to the algorithm is also proposed. Conversion of Grayscale image format to RGB image format using YCbCr image format is also applied in this proposed system.

The paper is organised as Section II describing the literature survey that supports the proposed algorithm. The Section III explains the modules of our work. The Section IV discusses the experiment that was conducted to obtain the required results. The Section V concludes our work with the merits and demerits of the proposed algorithm.

2. RELATED WORK

The work of A.S.Kavitha, et al. [1] proposed a new method for segmenting text lines from degraded historical document images like Indus. This method introduces a new combination of Laplacian and Sobel operations for enhancing low contrast pixels in the images. The characteristics of the components in the image are studied to eliminate unwanted components, which results in text components pruning in the image. It also proposes a grouping process, which involves the nearest neighbor criterion for merging text components. The iterative clustering process is then used to separate text and non-text regions.

Bolan Su, et al. [2] proposed an adaptive contrast map constructed for an input degraded document image. The contrast map is then binarized and combined with Canny's edge map to identify the text stroke edge pixels. The document text is further segmented by a local threshold that is estimated based on the intensities of detected text stroke edge pixels within a local window. This method is simple, robust and involves minimum parameter tuning.

The method of Nimbalkar Amruta A, et al. [3] is a different binarization technique that explains to give clear output image. It involves contrast inversion, after which a gray scale conversion is used to help detect the text stroke edges of degraded document images. In this method, the evaluation parameter includes Signal-to-noise Ratio (SNR) and Peak Signal-to-noise Ration (PSNR) of image quality metrics.

A.Anandhi, et al. [4] proposed a new method called sub pixel mapping that addresses issues of text recognition in degraded documents by using adaptive image contrast and binarization technique. The adaptive image contrast is a mixture of the local image contrast and the local image gradient that is forbearing to text and background variation caused by dissimilar types of document degradations, binarization algorithm for each case proved to be a very difficult procedure itself. It uses the case of degraded historical documents, then applies the technique called sub pixel mapping to convert the document images to the text format.

Vandana Gupta, et al. [5] proposed a new approach for detection and extraction of text data from both scanned document images and scene images. Text detection and extraction is performed in a four-step approach that consists of the pre-processing which includes binarization and noise removal of an image, image segmentation using connected component analysis, feature extraction using variance generation and finally classification by choosing a threshold value of variance property.

C.P. Sumathi, et al. [6] proposed a method based on morphological operators, wavelet transform, artificial neural network, skeletonization operation, edge detection algorithm, histogram technique. It discusses various schemes which were proposed earlier for extracting the text from an image. It also provides the performance comparison of several existing methods proposed by researchers in extracting the text from an image.

Chen Yan, et al. [7] proposed the multistage-approach and compared against some existing single-stage algorithms. The multistage approach recursively breaks down an images into sub-regions using quad-tree decomposition and extracts local features from each sub-region until an appropriate thresholding method can be applied to each sub-region. Quantitative analysis is performed using word recall and on 300 degraded historical images obtained from the Library of Congress.

Ergina Kavallieratou, et al. [8] has explained a hybrid binarization approach for improving the quality of old documents using a combination of global and local thresholding. First, a global thresholding technique specifically designed for old document images is applied to the entire image. Then, the image areas that still contain background noise are detected and the same technique is re-applied to each area separately. It achieves better adaptability of the algorithm in cases where various kinds of noise coexist in different areas of the same image while avoiding the computational and time cost of applying a local thresholding in the entire image.

Rachid Hedjam, et al. [9] has proposed a robust segmentation method for text extraction from the historical document images. This method is based on Markovian-Bayesian clustering on local graphs on both pixel and regional scales. It consists of three steps. In the first step, an over-segmented map of the input image is created. The resulting map provides a rich and accurate semi-mosaic fragments. The map is processed in the second step, similar and adjoining sub-regions are merged together to form accurate text shapes. The segmentation is obtained with clustering fixed number of classes. The method employs significantly the local and spatial correlation and coherence on both the image and between the stroke parts, and therefore is very robust with respect to the degradation. The resulting segmented text is smooth, and weak connections and loops are preserved. The output can be used in succeeding skeletonization processes which require preservation of the text topology for achieving high performance.

3. PROPOSED WORK

An OCR-less approach is designed and developed to extract text from a set of degraded documents. 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 to RGB images using luminance comparison and normalization are performed.

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.

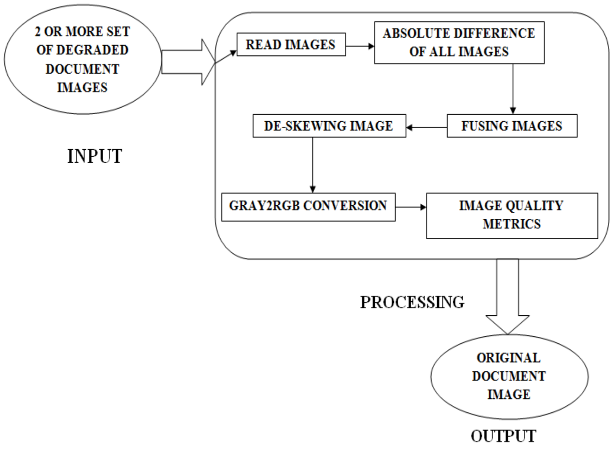


Figure 1: Proposed System Architecture

The Figure 1 represents the architectural diagram of our proposed system. The system takes 2 or more set of degraded document images as input. The processing unit consists of the following: Read Images – to read the input images. Absolute Difference of all Images – image subtraction is performed using absolute difference for all input images. This helps to detect the degraded text region in all the input document images. Fusing Images – images are fused. 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. De-skewing Image – detect skew in the input image. It helps to de-skew the input image with accurate skew scale value and skew angle with the direction of skewing either counter-clockwise or clockwise. Grayscale to RGB Conversion- is to obtain the original document image in RGB format, if the input image is in RGB format. If the input set of document images are RGB, the output original document image has to be RGB. During the fusing phase in the processing unit, the resultant image obtained is a Grayscale image; this has to be converted into RGB. Image Quality Metrics – to evaluate the proposed method. Major 7 metrics are considered and used to find the accuracy of the proposed system. The output will be an original document image with all the necessary text information that was obtained by extracting the text region from the input document images.

3.1 Algorithm 1 – Detect Degraded Region in Document Image

Step 1: Enter the number of input degraded text document images. Read all the input images. The images are in RGB format.

Step 2: Find the absolute difference of all input images. If A, B and C are the input degraded text document images. Then find the absolute difference of A, B; then A, C and finally B, C. Save all the resultant images for the next step.

Step 3: Fuse all the resultant images from step 2 with the respective input images from step 1. The fused single original document image is in Grayscale format.

3.2 Algorithm 2 – Image De-skewing

Step 1: Input one image without skewing as the reference image, another skewed image.

Step 2: Convert the input images from RGB to Grayscale format. Resize the input images.

Step 3: Detect SURF features for both the input images for original points and distorted points.

Step 4: Extract features from both the input images for the detected SURF features in step 3.

Step 5: Match features from step 3 and step 4. Take the index pairs of points of original points and distorted points.

Step 6: Show the matched features from step 5 with the help of estimating the geometric transformation in the skewed input image with respect to the un-skewed input image. Plot all the inliers and outlier points and differentiate the original points from both the input images.

Step 7: Find the distorted scaling value and the skew angle value with the direction of skew either counter-clockwise or clockwise with the estimated geometric transformation from step 6.

Step 8: Reference the output image to the world coordinates from the 2D.

Step 9: Wrap the input skewed image by applying all the geometric transformations obtained in step 6 and step 7. Save (write to the file directory) the resultant image.

3.3 Algorithm 3 – Grayscale to RGB conversion of an Image

Step 1: Input the fused resultant grayscale image along with any one initial (degraded document image) input RGB image for reference to extract the RGB values from it.

Step 2: Convert both the input images to YCbCr format.

Step 3: Double the values of the images in YCbCr format. Find the maximum and minimum values of the images.

Step 4: Perform normalization for both the images.

Step 5: Compare the luminance of both the images. Copy the luminance value of the input RGB image to the input Grayscale image.

Step 6: Convert the resultant image from YCbCr to RGB format. Convert the image again to uint8 format and write it to the file directory.

3.4 Algorithm 4 – Image Quality Metrics

Step 1: Read the initial input image and the final image obtained in the Algorithm 3 for quality metrics.

Step 2: Find the number of dimensions of both the input images. Convert both the images from RGB to Grayscale format.

Step 3: Compare the size of both the input images, the input images have to be of same size to check the quality metrics.

Step 4: Mean Square Error (MSE) is measured. Double the value of pixels for input images. To find the error, subtract both the images. Find sum (sum (error.* error)) / size (initial input image).

Step 5: Peak Signal to Noise Ratio (PSNR) is measured. Double the value of pixels for input images. Find the Mean Square Error (MSE) value. If Mean Square Error value is greater than 0, then PSNR = 10*log (255*255/MSE) / log (10); else PSNR value is 99.

Step 6: Normalized Cross Correlation is measured. Double the value of pixels for input images. Find sum (sum (initial input image.* result image)) / sum (sum (initial input image.* initial input image)).

Step 7: Average Difference is measured. Double the value of pixels for input images. To find the error, subtract both the images. Find sum (sum (error)) / size (initial input image).

Step 8: Structural Content is measured. Double the value of pixels for input images. Find sum (sum (initial input image.* initial input image)) / sum (sum (result image.* result image)).

Step 9: Maximum Difference is measured. Double the value of pixels for input images. To find the error, subtract both the images. Find max (max (error)).

Step 10: Normalized Absolute Error is measured. Double the value of pixels for input images. To find the error, subtract both the images. Find sum (sum (abs (error))) / sum (sum (initial input image)). Write all the image quality metrics measured in a text file and save the file in the file directory.

4. RESULTS

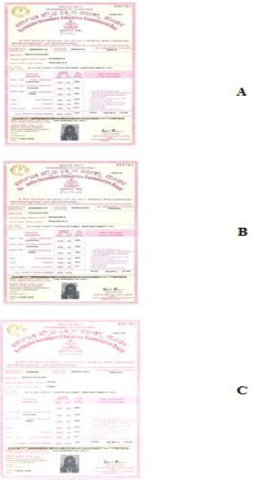


Figure 2: Input RGB degraded document images

The Figure 2 represents the 3 different degraded set of document images.

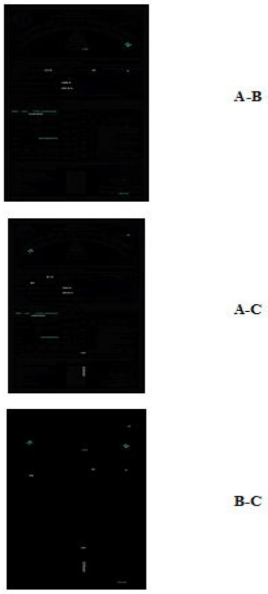


Figure 3: Intermediate stage of proposed algorithm

The Figure 3 represents the absolute differences of the degraded set of document images.

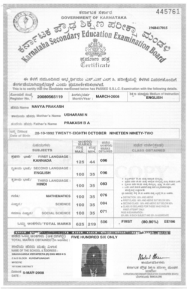


Figure 4: Grayscale original document image

The Figure 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.

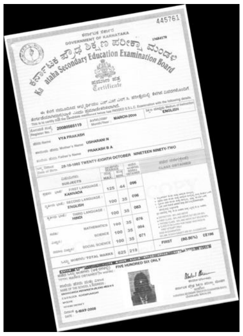


Figure 5: Skewed input degraded document image

The Figure 5 represents the skewed input degraded document image.

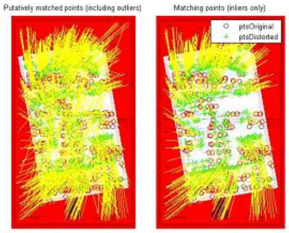

Putatively matched points (including outliers)
Matching points (inliers only)
ptsOriginal
ptsDistorted

Figure 6: SURF detection, extraction and matching

The Figure 6 represents the SURF detection, extraction and matching to deskew the input degraded document image.

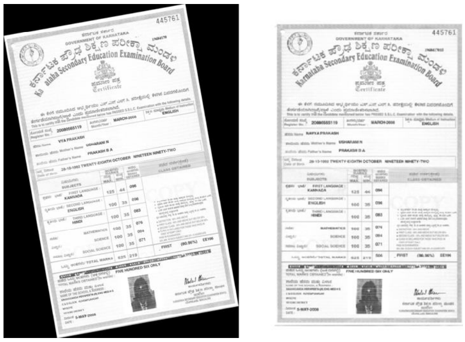


Figure 7: Skewed and Deskewed image

The Figure 7 represents the skewed input image and the deskewed result image obtained using SURF of skewed image with a reference image.

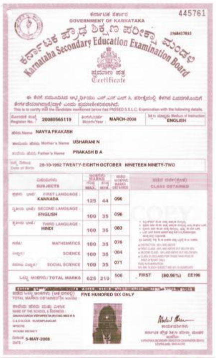
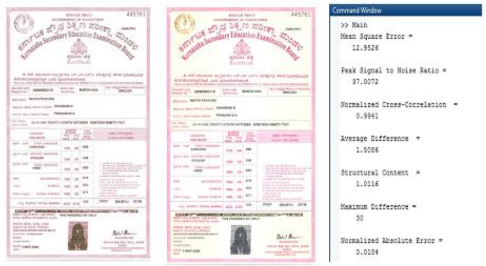


Figure 8: Grayscale to RGB converted image

The Figure 8 represents Grayscale to RGB converted image using YCbCr image format.



```
Command Window
>> 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
```

Resultant Image (RGB) **Original Image (RGB)** **Image Quality Analysis**

Figure 9: Image Quality Metric Values

The Figure 9 represents the values of various image quality metrics obtained.

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. Deskewing method is also implemented to detect skew of 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 same source.

Our algorithm is worked on printed text document images; in future it can be used for handwritten text document images as well as degraded images in the document. It can be further improvised to detect and rectify degraded video files. It can be converted into a mobile application (app) to capture real-time image of the degraded documents through the camera and use this as the input to the application, this has to be processed to result in an original document image. The application has to handle different orientations of the input images captured through the camera to produce an original document image in an accepted orientation form.

REFERENCES

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

[2] Bolan Su, Shijian Lu Member, IEEE, Chew Lim Tan Senior Member, IEEE, “A Robust Document Image Binarization Technique for Degraded Document Images”, IEEE Transaction on Image Processing, 2012.

[3] Nimbalkar Amruta A, Prof. Amrit Priyadarshi, “Recover degraded document images using binarization technique”, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 3 Issue: 6, June 2015.

[4] A. Anandhi, J. Renuka Jothy, A. Vijayalakshmi, D. Sathiyavani, “Sub Pixel Mapping in Degraded Document for Text Retrieval”, International Journal of Advanced Research in Computer and Communication Engineering, Vol. 3, Issue 2, February 2014.

[5] Vandana Gupta, Kanchan Singh, “A Novel Approach for Detection and Extraction of Textual Information from Scanned Document Images and Scene Images”, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

[6] C.P. Sumathi, T. Santhanam and G.Gayathri Devi.” A Survey on Various Approaches of Text Extraction in Images”, International Journal of Computer Science & Engineering Survey (IJCSES) Vol.3, No.4, August 2012.

[7] Chen Yan & Graham Leedham, “The Multistage Approach to Information Extraction in Degraded Document Images”**,** IEEE 2004.

[8] Ergina Kavallieratou and Efstathios Stamatatos, “Improving the Quality of Degraded Document Images”, Proceedings of the Second International Conference on Document Image Analysis for Libraries (DIAL’06), IEEE 2006.

[9] Rachid Hedjam, Reza Farrahi Moghaddam and Mohamed Cheriet, “Text Extraction from Degraded Document Images, EUVIP, IEEE 2010.