Character Recognition of Degraded Document Images Removing Strike

Liz Maria Mathew[1] Suma R[2]
[1]P.G. Scholar [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]St. Joseph's College of Engineering and Technology Palai, Kerala, India

Abstract— Optical character recognition (OCR) is the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text. Character recognition becomes difficult if the document images are degraded. So binarization is a technique that is used to remove the degradations. Libraries and archives around the world store large amount of old and historically important documents and manuscripts. But, due to many environmental factors, the poor quality of the materials used in their creation and improper handling cause them to suffer a high degree of degradation of various types. Today, there is a strong move towards digitization of old documents such as manuscripts so that their content can be preserved for future generations. Character recognition from noisy and degraded documents is still a challenging task. While considering historical document analysis, old printed documents have a high occurrence of degraded characters, especially broken characters due to ink fading. The objective of document image analysis is to recognize the text in degraded document images and extract the intended information. Here, a character recognition technique is proposed for the degraded document images. A strike removal method is also implemented to remove the strikes that are drawn over the text in the document images. Also, a contrast enhancement has been done to the adaptive contrast image using three methods which are linear contrast enhancement, piecewise linear stretch and homomorphic filter. The performance analysis is done based on the contrast enhancement that is done to the adaptive contrast image. The character recognition step includes horizontal scanning, vertical scanning and recognition using Discrete Wavelet Transform (DWT). The proposed system recognizes almost all characters of the input image.

Key words: Binarization, Character Recognition, Discrete Wavelet Transform (DWT), Adaptive image contrast

I. INTRODUCTION

Optical character recognition (OCR) is the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text. Character recognition becomes difficult if the document images are degraded. So binarization is a technique that is used to remove the degradations. Libraries and archives around the world store large amount of old and historically important documents and manuscripts. But, due to many environmental factors, the poor quality of the materials used in their creation and improper handling cause them to suffer a high degree of degradation of various types. Today, there is a strong move towards digitization of old documents such as manuscripts so that their content can be preserved for future generations.

Character recognition from noisy and degraded documents is still a challenging task. While considering historical document analysis, old printed documents have a high occurence of degraded characters, especially broken characters due to ink fading. The objective of document image analysis is to recognize the text in degraded document images and extract the intended information. Before character recognition the document images are binarized. Binarization plays a vital role in the document image processing. Binarization of document images is challenging in the case of old documents. The historical documents are often degraded by the bleed through where the ink of the other side seeps through to the front. In addition, historical documents are often degraded by different types of imaging artifacts. These different types of document degradations tend to induce the document thresholding error and make degraded document image binarization a big challenge.

Before recognizing the characters document, four steps are performed to obtain the bianarized document. In the first step, contrast image construction is performed in which the contrast of the text is increased with respect to the background. Here, an adaptive contrast image is generated which then undergoes different contrast enhancements. In the second step, the text stroke edge pixels of the document image are detected. In the third step, local threshold is estimated. In the fourth step, post-processing is done in order to filter out some single-pixel artifacts. Thus after these steps the binarized image will be obtained.

Here, a character recognition technique is proposed for the degraded document images. A strike removal method is also implemented to remove the strikes that are drawn over the text in the document images. Also, a contrast enhancement has been done to the adaptive contrast image using three methods which are linear contrast enhancement, piecewise linear contrast enhancement and homomorphic filter. The performance analysis is done based on the contrast enhancement that is done to the adaptive contrast image. The character recognition step includes horizontal scanning, vertical scanning and recognition using Discrete Wavelet Transform (DWT). The proposed system recognizes almost all characters of the input image.

II. RELATED WORK

Many methods have been developed for the character recognition of documents. The literature survey of various existing techniques has been done.

Analysis on the paper "Binarization of Historical Document Images Using the Local Maximum and Minimum," [1] introduced a method in which document image binarization aims to divide a document image into two classes, namely, the foreground text and the document background . It is usually performed in the document preprocessing stage and is very important for ensuing document image processing tasks such as optical character recognition (OCR).

Analysis on the paper "Iterative Multimodel Subimage Binarization for Handwritten Character Segmentation" [2] introduced a new category where the image is considered a collection of subimages. Each subimage provides a statistical model for the handwritten

characters that can be used to optimize the binarization of other subimages based on gray-level and stroke-run features. The proposed method uses these multimodels to iteratively arrive at the optimal threshold for each subimage. It can be applied to different types of documents where prior knowledge about the noisiness of the subimages is not available.

Analysis on the paper, "Optical Character Recognition Implementation Using Pattern Matching," [4] introduced a method in which a pattern matching based method for character recognition is proposed that would effectively reduce the image processing time while maintaining efficiency and versatility. The parallel computational capabilities of neural networks ensure a high speed of recognition which is critical to a commercial environment. The key factors involved in the implementation are: an optimal selection of features which categorically defines the details of the characters, the number of features and a low image processing time. The main functional modules in our OCR systems are: image acquisition module, pre-processing module, and feature extraction module and pattern generation. The main task of image acquisition module is to obtain text image from a scanner or a pre-stored image file. It is called 'image' because scanner inherently scans pixel of the text and not characters when patterns are scanned and digitised, the data may carry some unwanted noise. For example, a scanner with low resolution may produce touching line segments and smeared images.

Analysis on the paper, "Text Extraction and Document Image Segmentation Using Matched Wavelets and MRF Model,"[5] introduced a method in locating the textual data in an image. Further, it has extended text extraction scheme for the segmentation of document images. The text extraction scheme can identify and isolate textual regions in these kinds of images. The proposed document image segmentation algorithm separates the document image into three classes, namely text, picture (or graphics), and background. Such a system finds applications in image and text database retrieval, automated processing and reading of documents, and storing the documents in digitized form.

Analysis on the paper , "Top-Down and Bottom-up Cues for Scene Text Recognition"[7], introduced a method for scene text extraction. The method is inspired by the much advancement made in the object detection and recognition problems. A framework is presented that exploits both bottom-up and top-down cues. The bottom-up cues are derived from individual character detections from the image. Naturally, these windows contain true as well as false positive detections of characters. A Conditional Random Field (CRF) model is built on these detections to determine not only the true positive detections, but also what word they represent jointly. Top-down cues obtained from a lexicon-based prior are imposed, *i.e.* language statistics, on the model. In addition to disambiguating between characters, this prior also helps us in recognizing words.

Analysis on the paper "Robust Recognition of Degraded Documents Using Character N-Grams"[8] introduces a novel recognition approach is proposed that results in a 15% decrease in word error rate on heavily degraded Indian language document images. OCRs have considerably good performance on good quality documents, but fail easily in presence of degradations. Also, classical OCR approaches perform poorly over complex scripts such as those for Indian languages. These issues are addressed by proposing to recognize character n-gram images, which are basically groupings of consecutive character/component segments. Our approach is unique, since the character n-grams are used as a primitive for recognition rather than for post-processing. By exploiting the additional context present in the character n-gram images, better disambiguation is enabled between confusing characters in the recognition phase. The labels obtained from recognizing the constituent n-grams are then fused to obtain a label for the word that emitted them.

Analysis on the paper "Text Recognition of Low-resolution Document Images," [11] introduced a method to overcome the problem of using cheap and versatile cameras to capture a wide variety of documents. However, low resolution cameras present a challenge to OCR because it is virtually impossible to do character segmentation independently from recognition. In this paper these problems are solved simultaneously by applying methods borrowed from cursive handwriting recognition. To achieve maximum robustness, a machine learning approach is used based on a convolutional neural network.

III. PROPOSED METHOD

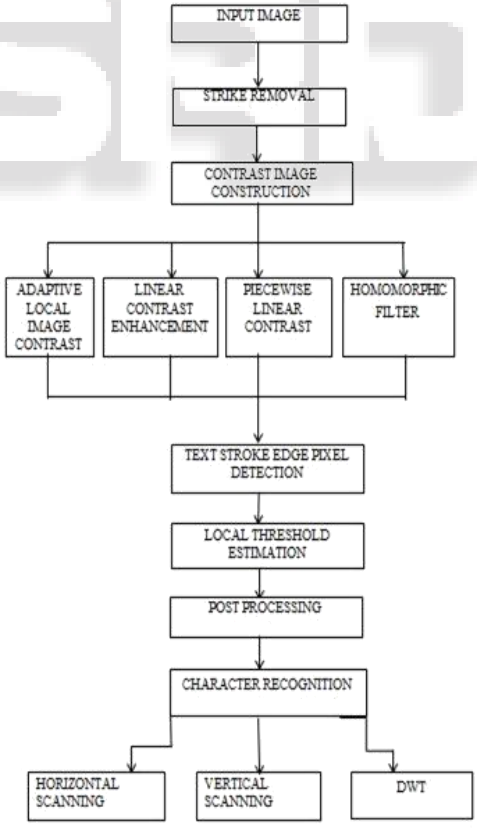


Fig. 1: Proposed System

Recovering of text from badly degraded document images is very difficult tasks due to the very high inter or intra-variation between the document background and the foreground text of different document images. Here, a character recognition technique is proposed for the degraded

document images. The document image binarization is done by using adaptive image contrast [13]. The following figure shows the block diagram of the proposed system.

A. Training Phase

Training phase involves neural network training of the input images. This helps in recognizing the characters more accurately.

B. Strike Removal

Strike removal deals with removing the strike drawn over the text. The image is converted into binary. In this each row is scanned from left to right. A counter is also set with a threshold. While scanning whenever a zero value is encountered the counter is incremented and the corresponding row value is stored. If the counter value is greater than the threshold it is a strike or else it is the portion of a character. If it is strike then all the zeros are converted into ones.

C. Contrast Image Construction

The image gradient has been widely used for edge detection and it can be used to detect the text stroke edges of the document images effectively that have a uniform document background. On the other hand, it often detects many nonstroke edges from the background of degraded document that often contains certain image variations due to noise, uneven lighting, bleed-through, etc. To extract only the stroke edges properly, the image gradient needs to be normalized to compensate the image variation within the document background.

D. Contrast Enhancement

Three types of contrast enhancements have been done to the image with adaptive local contrast. By enhancing the contrast the character recognition can be done more correctly.

1) Linear Contrast Enhancement

This type referred as contrast stretching, linearly expands the original digital values of the data into a new distribution. Here, minimum-maximum linear contrast stretch is used in which the original minimum and maximum values of the data are assigned to a newly specified set of values that utilize the full range of available brightness values. Consider an image with a minimum brightness value of 45 and a maximum value of 205. When such an image is viewed without enhancements, the values of 0 to 44 and 206 to 255 are not displayed. Important spectral differences can be deselected by stretching the minimum value of 45 to 0 and the maximum value of 120.

2) Piecewise Linear Contrast Stretch

When the distribution of a histogram in an image is bi or remodel, an analyst may stretch certain values of the histogram for increased enhancement in selected areas. This method of contrast enhancement is called a piecewise linear contrast stretch. A piecewise linear contrast enhancement involves the identification of a number of linear enhancement steps that expands the brightness ranges in the modes of the histogram.

3) Homomorphic Filter

Homomorphic filter is the filter which controls high frequency and low-frequency components, homomorphic filtering aims at handling large of image intensity, it has a multiplicative model. Thus in homomorphic filter an excess contrast enhancement can be removed to make a better character recognition.

E. Text Stroke Edge Pixel Detection

The purpose of the contrast image construction is to detect the stroke edge pixels of the document text properly. The constructed contrast image has a clear bi-modal pattern, where the adaptive image contrast computed at text stroke edges is obviously larger than that computed within the document background. Therefore the text stroke edge pixel candidate is detected by using Otsu's global thresholding method. As the local image contrast and the local image gradient are evaluated by the difference between the maximum and minimum intensity in a local window, the pixels at both sides of the text stroke will be selected as the high contrast pixels. The binary map can be further improved through the combination with the edges by Canny's edge detector, because Canny's edge detector has a good localization property that it can mark the edges close to real edge locations in the detecting image. In addition, canny edge detector uses two adaptive thresholds and is more tolerant to different imaging artifacts such as shading. It should be noted that Canny's edge detector by itself often extracts a large amount of non-stroke edges without tuning the parameter manually. In the combined map, only the pixels that appear within both the high contrast image pixel map and canny edge map are kept. The combination helps to extract the text stroke edge pixels accurately.

F. Local Threshold Estimation

The text can then be extracted from the document background pixels once the high contrast stroke edge pixels are detected properly. Two characteristics can be observed from different kinds of document images: First, the text pixels are close to the detected text stroke edge pixels. Second, there is a distinct intensity difference between the high contrast stroke edge pixels and the surrounding background pixels. The document image text can thus be extracted based on the detected text stroke edge pixels.

G. Optical Character Recognition

Optical character recognition (OCR) is the mechanical or electronic conversion of images of typewritten or printed text into machine-encoded text. The character recognition involves three steps which are horizontal scanning, vertical scanning and recognition using Discrete Wavelet Transform (DWT).

1) Horizontal Scanning

After the binary image is obtained from the post-processing apply the horizontal scanning to find how many lines in the binary document image. Then traverse from left to right of the image. Through traversing find the white pixel's starting x position and starting y position. Save the positions. After completion of traversing count the number of starting and ending points. Then divide the total count value by 2 to find the total number of lines in the given input image. Finally crop the line from the image based on the starting and the ending point of the x and y position. After the horizontal scanning number of lines in the binary image is detected and then the line is cropped from the image.

2) Vertical Scanning

The cropped line is obtained from the horizontal scanning. The cropped line is used to find the number of characters in the image using the vertical scanning. First select the line image which is obtained through horizontal scanning process. Then traverse from top to bottom of the image. Through traversing find the white pixel's starting x position and starting y position. Save the x and y positions. After completion of traversing count the number of starting and ending points. Then divide the total count value by 2 to find the total no of character in the given input image. Finally crop the character from the image based on the starting and the ending point of the x and y position.

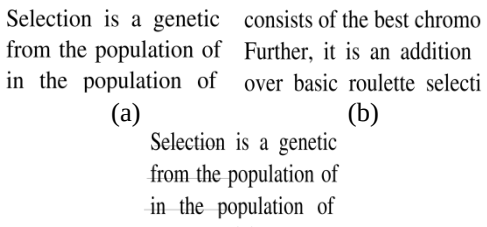
3) Character recognition using Discrete Wavelet Transform (DWT)

The discrete wavelet transform is a very useful tool for signal analysis and image processing. It can decompose signal into different components in the frequency domain. One-dimensional discrete wavelet transform (1-D DWT) decomposes an input sequence into two components (the average component and the detail component) by calculations with a low-pass filter and a high-pass filter. Two-dimensional discrete wavelet transform (2-D DWT) decomposes an input image into four sub-bands, one average component (LL) and three detail components (LH, HL, HH). In image processing, the multi-resolution of 2-D DWT has been employed to detect edges of an original image. The traditional edge detection filters can provide the similar result as well.

The cropped line of the image from horizontal scanning is selected. First get the training image folder which contains the image. Then apply the DWT on to the image, DWT split the image into four parts such LL,LH,HL and HH(The L means Low pass filtered image, H means High pass filtered image). Then find the average value of all the sub images and then store the average values. Then get the input image which is to be get from the horizontal scanning. Then apply DWT on to the image and then find the average value of these images. Finally find the difference value of this average value with the trained value images. Then find the minimum value to find the corresponding character and put these characters to the document.

IV. Experiments and Results

Experiments were conducted using MATLAB (R2014a) and on images with and without strike. The images can be of .png, .jpg, .bmp formats. The input images are fed into the system and the operations are performed. If the input image contains strike it is removed and then proceeded to further steps. Then adaptive contrast image is generated. The adaptive contrast image undergoes a contrast enhancement using three methods: linear contrast enhancement, piecewise linear contrast enhancement and homomorphic filter. This helps for better character recognition. The remaining steps include edge map generation, edge width estimation, local threshold generation, final binary result and the optical character recognition which involves horizontal scanning, vertical scanning and recognition using DWT. The final output is the characters that are recognized.



(a) (b)

(c)

Fig. 2: (a) A.png (b) B.png (c) strike.png

1) Performance Analysis

Here, a performance analysis is done by providing the adaptive contrast image into different contrast enhancements. The input image after undergoing adaptive contrast enhancement undergoes linear contrast enhancement, piecewise linear contrast enhancement and homomorphic filter. So the adaptive contrast image is measured between the images that are enhanced using linear contrast enhancement, piecewise linear contrast stretch and homomorphic filter. The performance analysis is done using a quality metrics, namely Peak Signal to Noise Ratio (PSNR). PSNR is one of the most widely used quality evaluation metrics.

Image Name	Methods		
	Linear Contrast Enhancement	Piecewise Linear Contrast Stretch	Homomorphic Filter
A.png	9.33891	9.01726	48.0389
B.png	8.16377	7.85977	48.1653
strike.png	8.91621	8.56987	48.1498

Figure 3: PSNR (peak signal-to-noise ratio) values of different methods

V. Conclusion

The focus of the work is on the character recognition of the degraded document images which are binarized. Here, a strike removal method is also implemented in order to remove the strikes that are drawn over the text in the document images. Also, a contrast enhancement has been done to the adaptive contrast image using three methods. The performance analysis is done based on the contrast enhancement that is done to the adaptive contrast image. The character recognition is done using Discrete Wavelet Transform (DWT). The proposed system recognizes almost all characters of the input image.

References

[1] Su, Bolan, Shijian Lu, and Chew Lim Tan. "Binarization of historical document images using the local maximum and minimum." In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems, pp. 159-166. ACM, 2010.

[2] Dawoud, Amer, and Mohamed S. Kamel. "Iterative multimodel subimage binarization for handwritten character segmentation." Image Processing, IEEE Transactions on 13, no. 9 (2004): 1223-1230.

[3] Seethalakshmi, R., T. R. Sreeranjani, T. Balachandar, Abnikant Singh, Markandey Singh, Ritwaj Ratan, and Sarvesh Kumar. "Optical character recognition for printed Tamil text using Unicode." Journal of Zhejiang University Science A 6, no. 11 (2005): 1297-1305.

[4] Mohammad, Faisal, Jyoti Anarase, Milan Shingote, and Pratik Ghanwat. "Optical Character Recognition Implementation Using Pattern Matching." IJCSIT) International Journal of Computer Science and Information Technologies 5, no. 2 (2014): 2088-2090.
[5] Kumar, Sunil, Rajat Gupta, Nitin Khanna, Santanu Chaudhury, and Shiv Dutt Joshi. "Text extraction and document image segmentation using matched wavelets and mrf model." Image Processing, IEEE Transactions on 16, no. 8 (2007): 2117-2128.
[6] Huang, Yi, Michael S. Brown, and Dong Xu. "User-assisted ink-bleed reduction." Image Processing, IEEE Transactions on 19, no. 10 (2010): 2646-2658.
[7] Mishra, Anand, Karteek Alahari, and C. V. Jawahar. "Top-down and bottom-up cues for scene text recognition." In Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2687-2694. IEEE, 2012.
[8] Dutta, Shrey, Naveen Sankaran, K. Pramod Sankar, and C. V. Jawahar. "Robust recognition of degraded documents using character n-grams." In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, pp. 130-134. IEEE, 2012.
[9] Pan, Jiyan, Ye Chen, Bo Anderson, Pavel Berkhin, and Takeo Kanade. "Effectively leveraging visual context to detect texts in natural scenes." In Proc. Asian Conf. Comp. Vis, vol. 12, p. 39. 2012.
[10] Zhang, Hongwei, Changsong Liu, Cheng Yang, Xiaoqing Ding, and KongQiao Wang. "An improved scene text extraction method using conditional random field and optical character recognition." In Document Analysis and Recognition (ICDAR), 2011 International Conference on, pp. 708-712. IEEE, 2011.
[11] Jacobs, Charles, Patrice Y. Simard, Paul Viola, and James Rinker. "Text recognition of low-resolution document images." In Document Analysis and Recognition, 2005. Proceedings. Eighth International Conference on, pp. 695-699. IEEE, 2005.
[12] Al-amri, Salem Saleh, N. V. Kalyankar, and S. D. Khamitkar. "Linear and non-linear contrast enhancement image." International Journal of Computer Science and Network Security 10, no. 2 (2010): 139-143.
[13] Su, Bolan, Shijian Lu, and Chew Lim Tan. "Robust document image binarization technique for degraded document images." Image Processing, IEEE Transactions on 22, no. 4 (2013): 1408-1417.
[14] Namane, Abderrahmane, and Patrick Meyrueis. "Multiple classifier for degraded machine printed character recognition." In Colloque International Francophone sur l'Ecrit et le Document, pp. 187-192. Groupe de Recherche en Communication Ecrite, 2008.
[15] Bai, Jinfeng, Zhineng Chen, Bailan Feng, and Bo Xu. "Image character recognition using deep convolutional neural network learned from different languages." In Image Processing (ICIP), 2014 IEEE International Conference on, pp. 2560-2564. IEEE, 2014.
[16] Brakensiek, Anja, Daniel Willett, and Gerhard Rigoll. "Improved degraded document recognition with hybrid modeling techniques and character n-grams." In Pattern Recognition, 2000. Proceedings. 15th International Conference on, vol. 4, pp. 438-441. IEEE, 2000.
[17] Leedham, Graham, Chen Yan, Kalyan Takru, Joie Hadi Nata Tan, and Li Mian. "Comparison of some thresholding algorithms for text/background segmentation in difficult document images." In null, p. 859. IEEE, 2003.
[18] Lu, Shijian, Bolan Su, and Chew Lim Tan. "Document image binarization using background estimation and stroke edges." International Journal on Document Analysis and Recognition (IJDAR) 13, no. 4 (2010): 303-314.
[19] Kleber, Florian, Markus Diem, and Robert Sablatnig. "Robust Skew Estimation of Handwritten and Printed Documents Based on Grayvalue Images." In Pattern Recognition (ICPR), 2014 22nd International Conference on, pp. 3020-3025. IEEE, 2014.
[20] Chen, Datong, Hervé Bourlard, and Jean-Philippe Thiran. "Text identification in complex background using SVM." In Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on, vol. 2, pp. II-621. IEEE, 2001.

