A REVIEW: DETECTION AND REMOVAL OF HAND-DRAWN ANNOTATION LINES

[1]ROVINA, [2]SEEMA BAGHLA, [3]SUNIL KUMAR

[1]M.Tech. Student (Computer Engg.), [2]Assistant Professor (Computer Engg.), [3]Assistant Professor
Yadavindra College of Engineering, Punjabi Univ. Guru Kashi Campus, Talwandi Sabo, Bathinda Punjab, India
E-mail: rimpy.16@gmail.com, garg_seema238@yahoo.co.in

Abstract- nowadays with the demand of computer vision, the importance of document images is increasing in various application areas and thus it is becoming important to improve the performance of optical character recognition. Performance of optical character recognition is badly affected due to hand-drawn annotation and underlines. Such annotation lines are normally drawn by reader in free hand to summarize some text. Thus the focus of this paper is on detection and removal techniques. Detection and Removal of various types of annotation lines such as untouched touched, straight, circular and other text surroundings lines.

Index Terms- Hand-drawn annotation lines, Text surrounding lines, Optical character recognition, untouched and touched lines.

I. INTRODUCTION

With the improvement of computer vision and pattern recognition the importance of document image processing is increasing day by day. Optical character recognition allows to automatically identifying characters through an optical mechanism. OCR can recognize both handwritten and printed text.

But the performance of OCR is depending on input documents. Optical character recognition system is poorly affected due to existence of hand-drawn annotation lines in various forms such as untouched underlines, circular lines, and other surrounding lines. To improve OCR performance, this annotation should be removed from the page. Hand-drawn annotation lines are normally used in many documents. OCR is a technology that gives authority to convert different types of documents such as scanned paper documents, PDF files by a digital camera into editable form.

II. PHASES OF OCR

It consist some steps to recognize text. These steps are: scanning, segmentation, preprocessing, feature extraction, recognition. The input image to OCR can be in any form hand written or printed text like books, generals, magazines, news papers, etc. Such input is given to OCR.

A. Scanning: Scanning process is the method of scan the document for input. Digital image of the document is captured with the help of scanning.

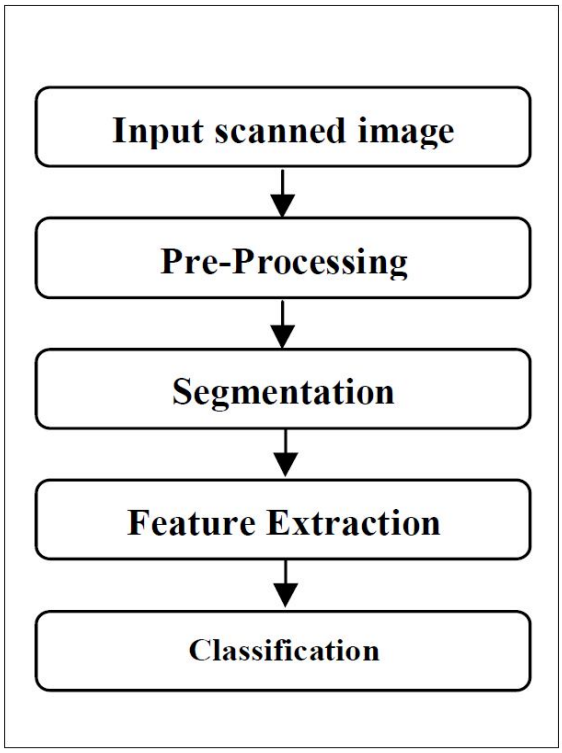

Fig.1 Phases of OCR

C. Pre-processing: After scanning process digital image obtained that may contain some amount of noise depending upon the quality of scanner. The lines might be skewed or broken. Thus, pre-processing is used for elimination of noise, Binarization of the image and segmentation.

D. Segmentation: In Character Recognition techniques, segmentation is done to make the separation between the individual characters of an image.

E. Feature Extraction: In this phase, the distinctive part of input characters is extracted. Feature extraction is the process to extract the most essential characteristics from the data. The most essential data means that's on the basis of that's the characters can be represented. Feature extraction can be one of the most difficult problems of pattern recognition.

F. Classification: Classification is the process of allocate the sensed data to their corresponding class with respect to groups. Classification is executed on the basis of stored features.

III. PREPROCESSING

Preprocessing is compulsory to eliminate noise and unwanted variations in script. The preprocessing can be divided into individual tasks such as thresholding, slant, skew detection and smoothing. Reduce the large variability of handwriting and to make writing style as uniform as possible is the main goal of preprocessing. Some main problems in preprocessing are:

- Annotation line removal
- Skew removal
- Slant estimation and correction
- Scaling and noise elimination
- Contour smoothing
- Reference line detection

A. Annotation line detection and removal

Hand-drawn annotation lines create big problem in OCR system while doing preprocessing. It is essential to detect and remove the hand-drawn annotation lines from document image. First detection method is applied on document then after detection of hand drawn annotation lines removing of these lines in very important. Basically hand-drawn lines are usually drawn in free hand by reader while reading any newspaper, book, magazines etc. Detection module tells whether there exist any hand drawn annotation and the lines are touched and untouched. Then after detection of lines removal module is used for removing lines after detection .There are different methods for detection and removal of lines for different types of lines. Some main steps:

a) A paper document is converted to electronic document for detection and removal of hand- drawn annotation lines. Document conversion is used for image processing, displaying, and character recognition.

b) Hand- drawn annotation detection will be done. In this detection of lines is very important for accuracy of OCR.

c) After detection of lines removal will be done according to different lines found in detection.

Due to annotation, lines in a document when these documents are used as input of some OCR engines it creates lots of misrecognition, so accuracy of OCR engine become low. For that reason, character recognition method must remove the noise after reading binary image data; smooth the image for better recognition. In hand-drawn annotation lines different types of annotation are used for summarizing some text by reader. Different types of untouched underlines are drawn on documents. To extract the metadata correctly from paper documents by using OCR, annotation line removal on the document image should be carried out. For the improvement of these systems is the need to manage successfully with the very large flood of papers such as office documents, commercial forms, government records, postal mails, etc. From this point of view, documents can be of three types: a) Printed, b) Handwritten and c) Mixed document. Printed document can be generated by printing technology such as laser, inject, offset and screen Printing etc.

B. SOME HAND-DRAWN UNDERLINES:

1) Touched underlines: OCR software makes it possible to detect characters on underline while the words are not touched with underlines. But if the underlines are touched with character or word then it will create some damage while detecting .In OCR recognition then misrecognition will while detecting the underlines.

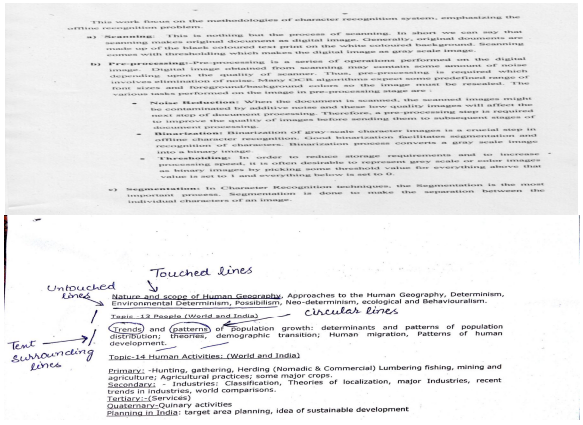

Fig. 2 Hand-drawn annotation lines example set.

2) Untouched underline: Untouched underlines are those which are not touched with the characters.
3) Various Scan Resolutions: Scanned images are not always fixed resolution. These methods are supposed to use fixed resolution.
4) Circular lines: Circular lines are drawn by reader for marking some text for further use in future.

Basically these types of lines are drawn while reading some document.

IV. PROPOSED METHODOLOGY

Different methods are used for detection and removal of hand drawn annotations lines. Basically this comes in preprocessing of OCR .In preprocessing of OCR some noises will occur with document image so for removing those noises preprocessing technique is used.

C. Gabor Filter Based

Gabor Filter Based: Gabor filter method was used for multi-channel processing of visual information in the human visual system. The Gabor filter was originally invented by Dennis Gabor in the year 1946. This is basically used for edge detection. Gabor filter was a linear filter used for image edge detection. When a Gabor filter is adjusted to an image, it gives the highest feedback at edges and at points where texture changes. On the other hand Gabor filter is a linear filter used for edge detection.

Frequency and orientation image of Gabor filters are similar to those of the human visual system. In the spatial domain, a two-dimensional Gabor function g(x,y) consist of a sinusoidal plane wave of some frequency and orientation, modulated by a two dimensional translated Gaussian envelope. In Gabor filters all filters can be generated from one mother wavelet by using some dilation and rotation operation. Its impulse response is defined by a harmonic function multiplied by a Gaussian function.

D. Connected component analysis

First of all region boundaries will be detected; it is often useful to extract regions which are not separated by a boundary. Set of pixels which was not separated by a boundary is call connected. The output of the change detection method is the binary image that consist only two labels, i.e., '0' and '255', represented as 'background' and 'foreground' pixels with some noise. The motive of the connected component analysis was to detect the large sized connected foreground region or object.

E. Mathematical Morphology

Mathematical Morphology is a theory and technique for the analysis of geometrical structures depend on set theory, lattice theory, topology, and irregular functions. Mathematical morphology was adjusted on images. Mathematical morphology was also the authority of morphological image processing, which contain of a set of operators that transform images according to the above characterizations.

F. Artificial intelligence

An artificial underline is then constructed to facilitate the removal of underline. The height of the artificial underline was decided by the longest marked. Artificial underline removed either the touched or untouched underline.

V. RELATED WORK

Adak and Bidyut (2014) proposed a scheme to detect strike-through text/words. Graph based model was used to represent a textual connected component as a graph. This approach was deal with strike-through text in handwritten documents. This graph based algorithm was tested on english,bengali and devnagri scripts. This approach can be extended to some other scripts if basic structual features of scripts are known. This method was used for preprocessing stage of recognition in documents[1].

Bai et al. (2004) proposed three-module approach for underline detection and removal in Chinese/English OCR. IN this paper the detection module used two method for detection of lines connected component analysis and bottom edge analysis. Connected component analysis was for detecting touched underline and doubtfull underline. Removal module remove the underline on the basis of detection for removing doubtfull underline artificial underline was used. This approach can deal with untouched, touched, broken and curved underlines. They confirmed to work for dealing with untouched,touched, and slightly curved underlines[10].

Das and Banerjee (2014) represented a technique has been used for underline detection and removal in a Bengali and English document which have been confirmed to work for any kind of underline like touched, untouched using Gabor filter and connected component analysis. In this paper first document page was taken as input then binarization algorithm was used. After that Gabor filter was applied on binarized image. Now after using Gabor filter it was cleared which were headline region and which underline region was In the underline detection module first use Gabor filter in a specific direction to detect underline region and then used connected component analysis to detect particular underline. In another part i.e. underline removal module used nearest neighbor approach. This algorithm used for both touched and untouched underline. [6].

Oba et al. (2009) introduced an underline removal method specific to Japanese business document. Firstly it deal with multi resolution image and normalized the input image. Secondly to reduce processed time, thirdly to selecting various underlines. After that finally, it removed table ruled lines. Line template matching was used for detection in business documents and line template matching can also

remove both thin and thick underlines. This method can remove touched, untouched and table ruled lines. [3].

Partihar et al. (2012) introduced algorithm for detection and removal of hand drawn underlines present in a scanned document. The main feature of this paper was on hand drawn lines all lines are drawn by hand. In this paper detection of underlines and detection of egdes of their covers was done. This algorithm was worked for all types of underlines whether touched or untouched and whether lines are curved and bent, as commonly seen when drawn by hand. This method is insensitive with scripts. To manage scripts with headlines, a preprocessing step was needed [7].

Pinto et al. (2004) introduced the problem of handwritten underline removal. The used of these underlines are very much found in books and removals of these underlines are national goal of the national libraries in their process of building digital libraries. Firstly the binarization of image is done then after underline removal was used for character recognition. Mathematical morphology method was used and line detection based on small eigen values. Removing of these lines was important for good OCR performance and for visual appearance of degraded documents [2].

Saba et al. (2014) introduced a comparison of various preprocessing techniques in offline script recognition. in offline script, the input was a paper image or a word. This method involved several preprocessing steps. some of them are hard such as line removal from text documents, skew removal, reference line detection, slant removal, noise elimination and skeleton. The goal for this method was to remove the handwriting variability that was inherited in scripts [8].

Sahu and kubde (2013) presented Off-line Handwritten Character Recognition. In this various methods are analyzed that have been proposed to realize the character recognition in an optical character recognition system. This material serves as a guide and update for readers working in the Character Recognition area.

The accurate recognition was directly depending on the nature of the material and quality. Current research was concerned with characters, but also words and phrases, and even the complete documents. This material served as a guide and update for readers working in the Character Recognition area[9].

Sharma and khandelwal (2013) presented a technique for improving the character recognition accuracy of Hindi OCR System. Optical Character Recognition was a process by which characters in text of printed document or scanned page are recognized and converted to ASCII character that a computer can read and edit.

Detection of each Hindi word requires some kind of smoothness in image which is done from preprocessing of an input image. In preprocessing technique two different step needed binarization and word segmentation. After preprocessing in underline detection of Hindi word lower zone of the word was selected as region of interest. If the region has maximum projection value it detects the word [4].

CONCLUDING REMARKS FROM LITERATURE

It has been observed that mis-recognition occurs in many document images due to various types of annotation lines. The ability to remove various hand drawn annotations lines from document various removal method has been defined. All these method work efficiently for underline whether it is touched or untouched. Connected component analysis work efficiently in case of single but this is not useful for whole paragraph. Thus to develop robust methods for detection and removal of hand-drawn annotation lines it is advisable to make hybrid techniques by combining different techniques on the basis of their advantages.

REFERENCES

[1] C. Adak and B. B. Chaudhuri, “An approach of strike-through text identification from handwritten documents”, IEEE International conference on frontiers in handwriting recognition, pp.643-648, 2014.

[2] J. R. C. Pinto, P. Pina, L. Bandeira, L. Pimentel and M. Ramalho, “Underline Removal on Old Documents”, Springer-Verlag Berlin Heidelberg, vol. 6, pp. 226-233, 2004.

[3] M. Oba, Y. Nozaki, T. Matsumoto and T. Onoyama, “Underline Removal Method by Utilizing Characteristics of Japanese Business Documents”, TENCON IEEE Region 10 Conference, pp. 1-6, 2009.

[4] N. Sharma and M. Khandelwal, “Detection of Bold Italic and Underline Fonts for Hindi OCR”, International Journal of Computer Trends and Technology (IJCTT), vol-4, pp. 2425-2428, 2013.

[5] P. Halgaonkar, “Connected Component Analysis and Change Detection for Images”, International Journal of Computer Trends and Technology, pp. 128-133, 2011.

[6] S. Das and P. Banerjee, “Gabor filter based hand-drawn underline removal in printed documents”, IEEE Automation, Control, Energy and Systems (ACES), First International Conference, pp. 1-4, 2014.

[7] S. Pratihar, P. Bhowmick, S. Suralt and J. Mukhopadhyay, “Detection and removal of hand- drawn underlines in a document image using approximate digital straightness“, Dar ’12 Methodology acm, pp. 124-130, 2012.

[8] T. Saba, A. Rehman, A. Altameem and M. Uddin, “Annotated comparisons of proposed preprocessing techniques for script recognition”, Springer verlag London, pp. 1337-1347, 2014.

[9] V. L. Sahu and B. Kubde, “Offline Handwritten Character Recognition Techniques using Neural Network”,International Journal of Science and Research, Vol-2, pp. 87-94, 2013.

[10] Z. L. Bai and Q. Huo, "Underline Detection and Removal in a Document Image using Multiple Strategies", IEEE Pattern Recognition, ICPR Proceedings of the 17th International Conference, vol. 2, pp. 578-581, 2004.
[11] Z. L. Bai and Q. Huo, "A Goal-Oriented Verification-based Approach for Target Text Line Extraction from a Document Image Captured by a Pen Scanner " , IEEE Pattern Recognition, ICPR Proceedings of the 17th International Conference, vol. 2, pp. 574-577,2004.

★★★