Detailed Study of Detection and Removal Techniques of Underlines and Annotations in Scanned Documents

Tadbeer Kaur
Electronics & Communication
Cgc Landran,Mohali,Punjab,India
tadbeerkaursweet@gmail.com

Dr.Rinkesh Mittal
Hod ,Electronics & Communication
Cgc Landran,Mohali,Punjab,India

Abstract—

The ocr system's performance gets badly affected due to the presence of hand drawn underlines (straight, curved, touched, untouched, bent, broken) and annotations of various forms. Such underlines and annotations are drawn by reader in free hand .In this paper, we will discuss the merits and demerits of techniques used for detection and removal of underlines and annotations proposed in earlier

Keywords—*underrline detection and removal,annotation detection and removal,connected component analysis,bottom edge amalysis, annotation detection by digital geometric rules,ostu binarization module*

1. INTRODUCTION

A text document can be usually seen with various underlines (curve , bent , straight ,touched ,untouched our broken etc)and annotations(circular, elliptical) made by the user to memorise text etc. In this paper deals with study of techniques which a have been used earlier in the detection and removal of these underlines and annotations so as to improve the working of ocr system .In [1] Zhen –long Bai used the technique of common connected analysis along with bottom edge analysis to detect and remove the underlines in a document image . In [2] Arvind K.R. proposed a method for line removal and restoration of erased areas of handwritten elements .In[3] an algorithm for detection and removal of underlines from the scanned images by locating the underlines by detecting the edges of their covers as a sequence of approximately straight segments from the boundary edge map of underlined parts .after getting the exact cover of underline strategy is applied for underline removal. In [4] yet another algorithm was proposed in which a scheme for detection and removal of hand drawn annotations from scanned document page was applied. The cover of the annotated object was detected as sequence of straight edge segments after getting cover, method of inpainting was used where reconstruction was needed .In [5] underline removal is accomplished by separating text from overlapping strokes, the system first detects the smooth strokes and then identifies probable underlines ,by measuring the length of stroke. if it is greater than a certain length ,it is considered as non text and removed from the document . In [6] an algorithm was proposed for line removal and character restoration using Block Adjancy Graph representation of binary image as input.

2. Existing Techniques (merits and demerits)

2.1) Connected Component Analysis and Bottom Edge Analysis:-

This technique reportedly was used in [1] by Zhen –Bai. the steps followed in detection and removal of underlines are as follows :-First an underline detection is applied secondly underline removal is applied lastly disambiguity module is practiced to reduce the risk of wrongly and doubtful underlines

Underline detection module:-using connected component analysis for untouched underlines and bottom edge analysis for confirmed untouched lines

Underline removal module:- for untouched underlines ,the detected connected components are deleted directly and hence removed and for untouched underlines the disambiguity analysis carried first and hence the underlines are removed hence

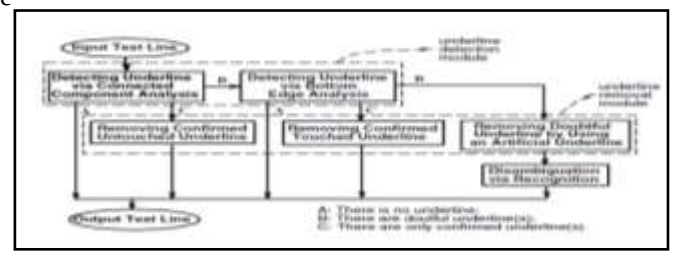


Figure1 :-Architecture of Underline Removal and Detection Using Connected Component Analysis

Merits :- it removes touched, untouched, slightly curved underlines

Demerits :- better strategies for dealing with broken and doubtful lines need to be developed .secondly the disambiguity module needs improvement

2.2) Using Gabor filter along with Connected Component Analysis:-

This technique was reportedly used in [7] for underline detection and removal in Bengali &English document

For underline detection first Gabor filter in a specific direction to detect the underline region and then connected component analysis is applied to detect the particular underline and then underline removal is carried out by nearest neighbor approach

Underline detection module:-

a) A document imaghe as input is taklen then on it recursive **Ostu Binarization algorithm** is applied to get the binarised image
b) After that apply gabor filter so that with its help it can be identified which is the underline region
c) next apply binarization algorithm on the gabor filter output image ,as an output one gets only the underline region of the document perfectly because the intensity of the read line region is low than underline region
d)then particular underline region is chosen by using the connected component analysis
e)after that non interested region is removed in red colour by using the connected component analysis.asa a result underline is detected separately

Underline Removal Module :-

For untouched line:-
an untouched line can be detected and removed by connected component analysis

For touched underline :-
a) Apply thinning algorithm in the portion of the underline region and the output obtained is thinned image

b) Apply connected component analysis

c) Next its decided whether the underlines are touched or untouched: - Move from left to right applying connected component analysis ,if the pixel is black then it is confirmed it is branch that is ,it is touched underline, if it is not black pixel then it is untouched underline

d) For removing the touched underline first remove the branch portion and the move from pixel in left to right of connected component region and when a black pixel is obtained then white value is put over that and 8 nearest neighbor pixel if its black, hence underline is removed

Merits:- It works efficiently for touched, untouched and broken underline

Demerits:-Broken underline removal needs improvement and also a method of how to utilize characters from business document needs to be developed

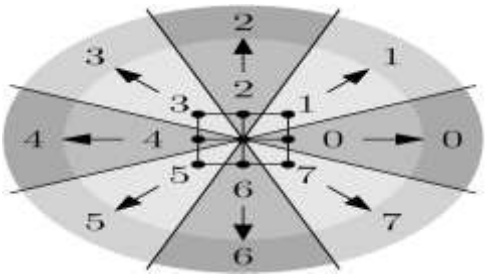
2.3) Digital geometric rules and inpainting technique

In [4] an algorithm was proposed for the detection and removal of hand drawn annotations by using the strategy of digital geometric analysis and inpainting technique using the fast marching method commonly called as FHH.

The system works in the following manner:

Detection of annotations by Digital Geometric Rules:-

a) At first Boundary edges are extracted from the binarized image using structuring element of size 3x3
b) Then algorithm detects the annotation object boundary as a chain which is sequence of digital straight edges
c) Finally a set ‘s’ of straight segments is extracted which covers only annotation object as much as possible but does not touch the characters
Every straight edges comprises at most two chains codes .for one of these its singular code ,the run length must be 1,for the non singular direction can have only two lengths which are consecutive integers
d) The set ‘s’ may vary on changing start point .if ‘p’ is the start point then procedure for tracing the straight edges from ‘p’ start in two directions as there will be two unvisited neighbors . Let one neighbor lie in direction d1 and the other indirection d2 .If d and d2 differ by more than 1 then point p is considered as a start point .The end point will be the point ‘q’ where the straight edge finding algorithm halts .consequently other start point (forming chin) is found
e) To find non singular direction of the connecting edges between two chain segments Bin Direction Code is followed
f) Following these steps the collection of boundary line segments that cover the annotation line area can be found out ,this covered area is used as mask for inpainting



I. EASE OF USE

Figure 2:- Bin direction code

2) **Document Cleaning by Inpainting**:-

a) Construct mask & source image for inpainting

b) Image smoothness estimators works on the weighted average of pixel gray values which is calculated over a known neighborhood of image pixel to be inpainted

c) Fast Marching Method is used to propagate image information)after detection of mask ,fix the source image by subtracting the mask that is the annotation mask from the input image

Merits:-Method can accurately quantify the area of annotation line whether they are touched ,untouched by text characters and whether the lines are curved or bent as commonly seen when drawn by hand

Demerits:-Final reconstruction of characters segments can be improved

2.4) Using Connected component analysis and block segmentation

This technique was applied in [2]

Steps

a**) Noise removal**:-it is carried out by connected component analysis,and ON no of the pixels are obtained

$$Tp = np - minp/maxp-minp$$
$$Ta = na-min\ a/maxa-min\ a$$

np:- No of ON pixels

na:-aspect ratio of component

Tp&Ta :-less than 0.002 computed empirically

b) **Block Segmentation**

Run length smoothening of the image with the parameter selected so that inter and intra gap characters uptil the paragraph are filled

c**) Skew Detection and correction**:-

Assuming that the maximum skew would not be greater than 10 degree the image is rotated & HPP is obtained along with entropy values

d) **Line detection and removal**:-

1) Where line exists there is a peak in the HPP (Horizontal Projection Profile)

2) After potential line containing rows have been detected the rows are traversed and then the run length within them is obtained

3) lines are removed using the connected component analysis

e) **Restoration of Handwritten elements**:-

it involves two steps the detection of the strokes and filling up of the erased area

Merits: - Restoration of hand written elements (in a fast manner) with multiple lines passing over them with varying thickness and secondly the document is divided into blocks and skew correction was done

Future scope :- Restoration of printed characters

2.5) Using Connected Component Analysis, Boundary Extraction

This technique was used in [3] .In it Detection of almost straight lines from boundary edge map of underline parts

Method: - Height and weight is found by the connected component analysis and boundary edge extraction is used to detect the underline covers

Merits: - it efficiently removes the touched , untouched ,curved or slightly bent ,this method works even in the presence of headlines

Demerits: -to find the broken, small length and doubtful underlines a few more thresholds can be set

3) CONCLUSION

Connected Component Analysis works efficiently in underline removal of single line at a time , but this method can not be applied to whole paragraph so it is time consuming, the best method so far is the detection by digital geometric technique used in [4]since it efficiently detects and removes annotations and underlines both ,reconstruction of the characters is carried out by the technique of inpainting

4) FUTURE SCOPE

An efficient hybrid method can be developed which removes all types of underlines (straight, curved, bent touched, untouched, broken) and annotations (curved or elliptical bounded regions ,notes etc written on the document image)can be made. Removal of broken underlines needs to be stressed . Lastly more efficient reconstruction techniques can be applied. Inpainting methods can be improved

5)REFERENCES

[1] Z.-L. Bai and Q. Huo, "Underline detection and removal in a document image using multiple strategies," Proceedings of the 17th InternationalConference on Pattern Recognition, pp. 578-581, 2004.

[2] K. R. Arvind, J. Kumar, and A. G. Ramakrishnan, "Line removal and restoration of handwritten strokes," Proc. Con!

Computational Intelligence and Multimedia Applications, vol. 3, pp. 208-214, 2007.

[3] sanjoy pratihar , Partha Bhowmick,Shamik Sural ,Jayanta Mukhopadhyay "detection and removal of had drawn underlines in a document image using approximate Digital Straightness"

[4] sanjoy pratihar , Partha Bhowmick,Shamik Sural ,Jayanta Mukhopadhyay "Removal of Hand drawn Annotations Lines from Document Images by Digital Geometric Analysis"

[5] Y. Govindaraju and S. H. Srihari, "Separating handwritten text from interfering strokes," From Pixels to Featuures III - Frontiers in hand- writing recognition, S. Impedovo, 1.e. Simon (eds.), vol. North-HollandPublication, pp. l7-28, 1992.

[6] B. Yu and A. K. Jain, "A generic system for form dropout," *IEEE Trans. on PAMI*, Vol. 18, No. 11, pp.1127-1134, 1996.

[8] R. Klette and A. Rosenfeld, Digital Geometry: Geometric Methods for Digital Picture Analysis. San Francisco: Morgan Kaufmann, 2004.

[9] A. Telea, "An image inpainting technique based on the fast marching method," 1. Graphics, GPU, and Game Tools, vol. 9, no. I, pp. 25-36, 2004.

[10] R. C. Gonzalez and R. E. Woods, Digital Image Processing, 3rd ed. Pearson Education, 2009.

[7] Supriya Das ,Purnendu Banerje, " Gabor Filter Hand Drawn Underline Removal in Printed Documents