Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2016, 3(1): 12-20

Research Article ISSN: 2394 - 658X

Hand-Drawn Annotation and Underline Detection and Removal in Scanned Documents Using Artificial Neural Network & Fuzzy C-Means Clustering

Tadbeer Kaur and Rinkesh Mittal

Department of Electronics and Communication Engineering, Chandigarh Group of Colleges,
College of Engineering, Mohali, India
tadbeerkaursweet@gmail.com

ABSTRACT

The OCR system is computerized scanning system that enables user to scan a text document into an electronic computer file that can be edited, usually the OCR system's performance gets badly affected due to the presence of hand drawn underlines (straight, curved, touched, untouched, bent, broken, elliptical etc) and annotations lines of various forms (such as straight lines, circular lines, elliptical, strokes or embossed lines etc). Such underlines and annotations are drawn by reader in free hand to memorize text, so this need to be removed from the scanned text, so as to make text legible thereby improving OCR efficiency. In this paper, we will discuss the merits and demerits of techniques used for detection and removal of underlines and annotations proposed earlier. Also an efficient technique to detect and remove different types of annotations and underlines is proposed in this paper which is based on Artificial Neural Network and Fuzzy C-means clustering.

Key words: OCR, ANN, FCM clustering, underline & annotation detection and removal

INTRODUCTION

A text document can be usually seen with various underlines (curve, bent, straight, touched, untouched or broken etc) and annotations(straight lines, circular, elliptical, strokes, embossed lines) made by the user to memorise text etc. In this paper we will deal with study of techniques which a have been used earlier in the detection and removal of these underlines and annotations. Also an effort has been made to design an efficient algorithm to detect and remove various kinds of annotations and underlines marks in the text document. In [1] Bai et al used the technique of common connected analysis along with bottom edge analysis to detect and remove the underlines in a document image. Arvind et al [2] proposed a method for line removal and restoration of erased areas of handwritten elements. Pratihar et al [3] developed 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. Pratihar et al [4] proposed yet another algorithm 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. Govindraju et al [5] explored a method for underline removal 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. Yu et al [6] used the method for line removal and character restoration using Block Adjacency Graph representation of binary image as input. In this paper we have proposed ANN based method to detect and remove the various kinds of underlines and annotation marks, so that we are able to improve the working of OCR system, to read characters from scanned images which are otherwise degraded by such annotations. The analysis of existing techniques and their merits and demerits are given in next section.

CONNECTED COMPONENT ANALYSIS AND BOTTOM EDGE ANALYSIS

This method was given by Bai et al [1], the steps followed in detection and removal of underlines is as follows:-First underline detection is applied secondly underline removal is applied lastly disambiguity module is practiced to reduce the risk of wrongly and doubtful underlines.

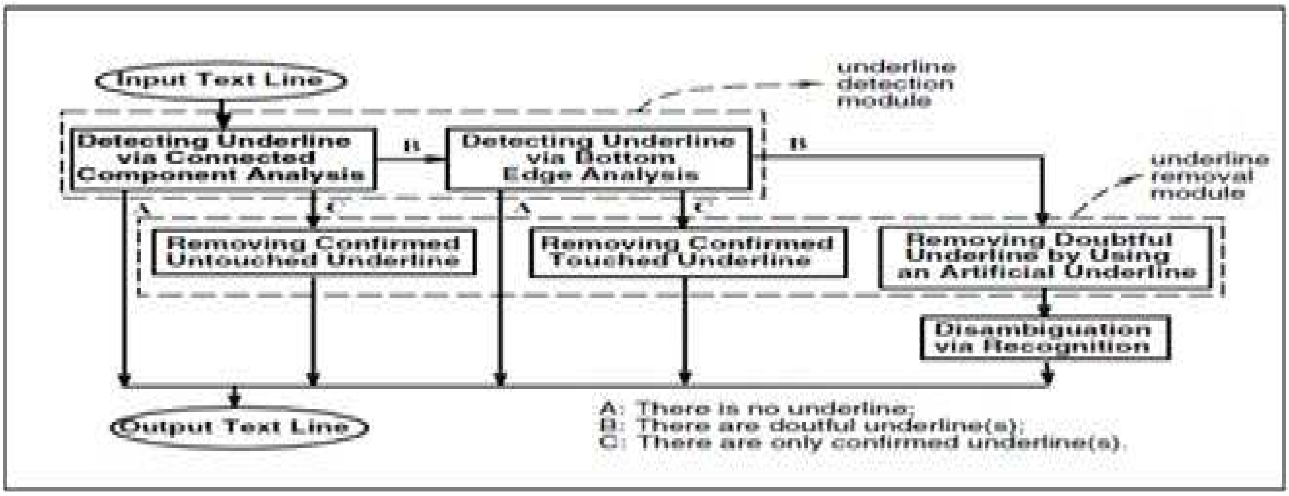

Input Text Line
underline detection module
Detecting Underline via Connected Component Analysis
B
Detecting Underline via Bottom Edge Analysis
B
underline removal module
A
C
A
C
Removing Confirmed Untouched Underline
Removing Confirmed Touched Underline
Removing Doubtful Underline by Using an Artificial Underline
Disambiguation via Recognition
Output Text Line
A: There is no underline;
B: There are doutful underline(s);
C: There are only confirmed underline(s).

Fig.1 Architecture of Underline Removal and Detection Using Connected Component Analysis

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.

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.

USING GABOR FILTER ALONG WITH CONNECTED COMPONENT ANALYSIS

This technique was reportedly used by Das et al [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 neighbour approach.

Underline Detection Module

- A document image as input is taken then on it recursive Ostu Binarization algorithm is applied to get the binarized image.
- After that apply gabor filter so that with its help it can be identified which is the underline region.
- 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 red line region is low than underline region.
- Then particular underline region is chosen by using the Connected Component Analysis.
- After that non interested region is removed in red colour by using the Connected Component Analysis. As 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

- Apply thinning algorithm in the portion of the underline region and the output obtained is thinned image.
- Apply Connected Component Analysis.
- 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.
- 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 neighbour 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.

DIGITAL GEOMETRIC RULES AND INPAINTING TECHNIQUE

Pratihar et al [4], proposed an algorithm 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

- At first Boundary edges are extracted from the binarized image using structuring element of size 3x3.
- Then algorithm detects the annotation object boundary as a chain which is sequence of digital straight edges.
- 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.
- 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 neighbours. Let one neighbour 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.
- To find non singular direction of the connecting edges between two chain segments Bin Direction Code is followed.
- 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.

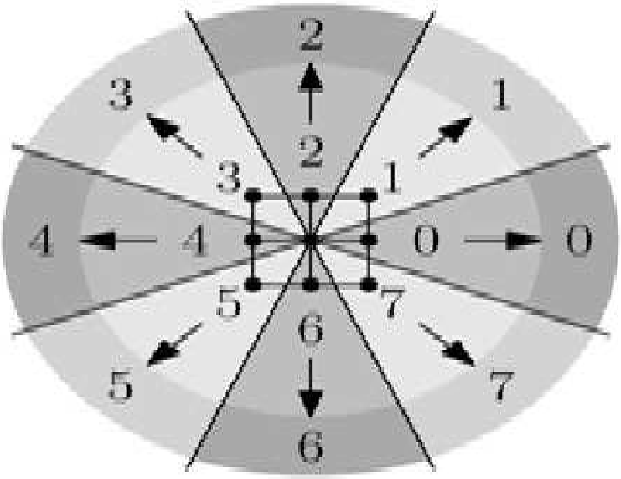


Fig. 2 Bin Direction Code

Document Cleaning by Inpainting

- Construct mask & source image for inpainting.
- Image smoothness estimators works on the weighted average of pixel gray values which is calculated over a known neighbourhood of image pixel to be inpainted.
- 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.

CONNECTED COMPONENT ANALYSIS AND BLOCK SEGMENTATION

This technique was applied by Arvind et al [2]
Steps

- **Noise removal:** It is carried out by connected component analysis, and ON no of the pixels is obtained.

$$Tp = np – minp/maxp\text{-}minp \quad (1)$$

$$Ta = na\text{-}min\ a/maxa\text{-}min\ a \quad (2)$$

np:- No of ON pixels, na:-aspect ratio of component

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

- **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.
- **Line detection and removal**: Where line exists there is a peak in the HPP (Horizontal Projection Profile).
- After potential line containing rows have been detected the rows are traversed and then the run length within them is obtained.
- Lines are removed using the connected component analysis.
- **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.

CONNECTED COMPONENT ANALYSIS, BOUNDARY EXTRACTION

Pratihar et al [3] gave this method; in it detection of almost straight lines from boundary edge map of underline parts has been performed.

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.

PROPOSED METHOD

The OCR algorithms may perform well for the ideal clean images. Recognition of objects and patterns that are corrupted by various noises has been a goal of recent research. Therefore underlining in the scanned documents is the major problem in OCR systems, which needs to be eliminated. Similarly other annotation lines i.e. skew lines, cross lines, curved or round annotation lines also need to be removed in order to get characters free from such annotations in the scanned documents. In existing work related to annotation removal, different techniques have been found but most of them work on a particular type of annotation line and fail to detect other lines. Therefore there is not a universal method available which can remove all types of annotation lines in document images. In this work, we have explored this problem and tried to find out solution for removing various types of annotation lines in a document in single running of the algorithm. The different explained below:-

Pre-processing

- **Collection of datasets:** All the images were acquired by scanning the text documents annotated with different colour pen and different types of annotation lines i.e. skew, elliptical, curved, crossed etc. are annotated to evaluate the performance of the algorithm.
- **Pre-processing the images:** The images were converted to PNG format while scanning and then converted to Lab format.

Gabor Filtering

In this we used Gabor filters to enhance the image character region as well as annotated region from the given text. Gabor filters are bandpass filters which are used in image processing for feature extraction, texture analysis, and stereo disparity estimation. The impulse response of these filters is created by multiplying a Gaussian envelope function with a complex oscillation. In this way, Gabor filters help in enhancing the edges in the image when applied in various directions on an image Gabor filters, which have been shown to fit well the receptive fields of the majority of simple cell in the primary visual cortex [11], are modulation products of Gaussian and complex sinusoidal signals.

A 2D Gabor filter oriented at angle is given by:

$$\begin{aligned} &g_{\theta,\Theta,\varphi,\sigma,\gamma}(x,y)=\exp(-\frac{x^2+\gamma^2 y^2}{2\sigma^2})\cos(2\pi\frac{x^2}{\theta}+\varphi) \\ &\textit{where} \\ &x'=x\cos\theta+y\sin\theta \\ &y'=x\sin\theta+y\cos\theta \end{aligned} \tag{3}$$

λ: This is the wavelength of the cosine factor of the Gabor filter kernel and here with the preferred wavelength of this filter
Θ: This parameter specifies the orientation of the normal to the parallel stripes of a Gabor function.
Φ: The phase offset φ in the argument of the cosine factor of the Gabor function is specified in degrees. Valid values are real numbers between -180 and 180.
γ: This parameter, called more precisely the spatial aspect ratio, specifies the ellipticity of the support of the Gabor function.
***b*:** The half-response spatial frequency bandwidth *b* (in octaves) of a Gabor filter is related to the ratio σ / λ, where σ and λ are the standard deviation of the Gaussian factor of the Gabor function and the preferred wavelength, respectively. The Gabor filter has been shown to be an efficient and robust edge detector which offers distinct advantages over traditional edge detectors, such as Roberts, Sobel, etc., and can be comparable even superior to Canny edge detector generally thought as an optimal edge detector.

FCM Clustering

FCM is a method of clustering which allows one piece of data to belong to two or more clusters. The main difference between the traditional hard clustering and fuzzy clustering can be stated as follows. While in hard clustering an entity resides only to single cluster, in fuzzy clustering entities are allowed to reside too many clusters with different degrees of membership. The most known method of fuzzy clustering is the Fuzzy c-Means method which is being most widely used in image processing applications. The steps involved in FCM are briefed as below. The following is description of the FCM algorithm, which is implemented Fuzzy Logic.
1) Select the number of clusters $c(2 \leq c \leq n)$, exponential weight $\mu(1 < \mu < \infty)$, initial partition matrix U^0, and the termination criterion. Also, set the iteration index l to 0.
2) Calculate the fuzzy cluster centers $\{V_i^1 \mid i = 1, 2, 3.....c\}$ by using U^1
3) Calculate the new partition matrix U^{1+1} by using $\{V_i^1 \mid i = 1, 2, 3.....c\}$
4) Calculate the new partition matrix $\Delta = \| U^{i+1} - U^1 \| = MAX_{i,j} \mid u_{i,j}^{i+1} - u_{i,j}^1 \mid .if \Delta > \varepsilon$ then set i= i + 1 and go to step 2, If not, then stop.
So fuzzy-c-means is applied to do clustering which results in segmentation of whole text image into number of clusters. FCM uses Euclidian distance for making clusters and hence clusters the annotation lines in separate clusters from the characters. After that two classes are obtained by using region merging in which annotation line pixels are set as first class and rest of the image as second class.

Artificial Neural Network (ANN)

ANN is used to separate the unwanted region that is the annotated area from the scanned text which is required as final output, the working of the ANN as well as the three layer model back propagation is as explained below :-
The multi-layer back-propagation neural network is best suited for the engineering applications.[12]. Many researchers proved that the multi-layer back propagation with three layers can perform arbitrarily complex classification. [13-14].

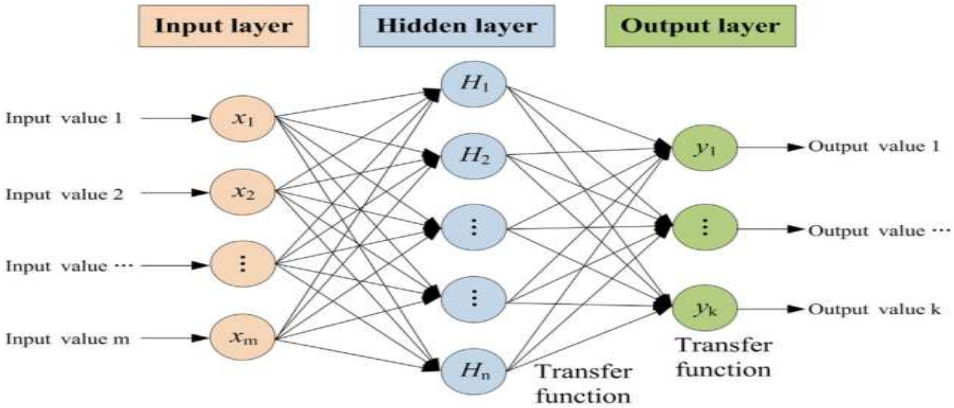


Fig. 3 Three Layers Model Back Propagation Neutral Network

Propagation of data takes place from input layer to the output layer. In supervised learning the network is presented with a series of matched input and output patterns and the connection strengths or weights of the connections automatically adjusted to decrease the difference between the actual and desired outputs. Patterns are presented to the network and a feedback signal which is equal to the difference between the desired and actual output is propagated backwards through the network for the adjustment of weights of the layers' connections according to the back propagation learning algorithm. Train lm is a network training function that updates weight and bias values according to Levenberg-Marquardt optimization. Train lm is often the fastest back propagation algorithm in the toolbox, and is highly recommended as a first-choice supervised algorithm, although it does require more memory than other algorithms.

Inpainting is performed finally; the parts of the characters which are lost are finally reconstructed by filling in the mean intensity values for the characters.

RESULTS AND DISCUSSIONS

Following are the scanned input images marked with underlines and annotations along with the resultant images after the detection and removal of these annotation lines obtained by following the proposed method.

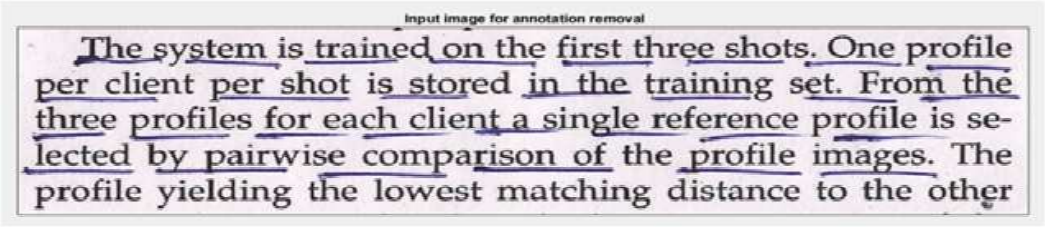
Text Image with Broken Underlines

a) In fig. 4, we have taken a scanned text image marked with broken underlines, so our aim is to remove these underlines from the text, such that the resultant output image obtained by applying the proposed algorithm, is free from the underlines

b) Next, this input image is then tested with the OCR system so as to see how well the OCR system performs with the input image marked with broken underlines; the image obtained is shown in Fig. 5.

c) The image is then tested with the proposed method ,in which FCM clustering is performed,along with applying gabor filter,finally ANN testing and training is performed on the input image so as to separate the annotated area that is the underlines from the characters, as a result the broken underlines present in the input image are removed, the image so obtained is free from underlines containing just characters as shown in Fig. 6.

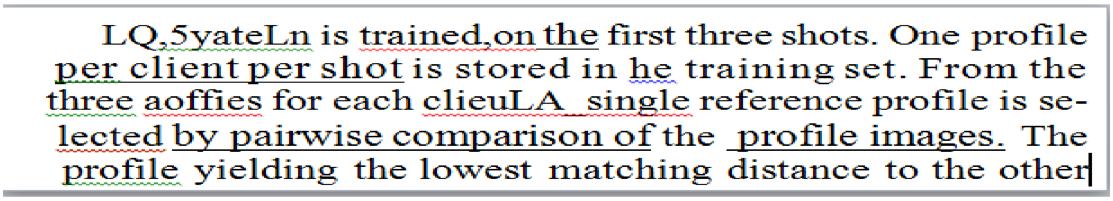
d) Finally, the output image obtained from the proposed method, which is free from the broken underlines is then tested with OCR system, so as to see if the OCR system's efficiency is improved, the result is as shown in Fig. 7. Thus it can be seen that the proposed method works well in the detection and removal of the underlines and also the output of the OCR system gives more accurate output after applying the proposed method to the underlined text image.



Input image for annotation removal

The system is trained on the first three shots. One profile per client per shot is stored in the training set. From the three profiles for each client a single reference profile is selected by pairwise comparison of the profile images. The profile yielding the lowest matching distance to the other

Fig. 4 Text Input Image With Broken Underlines



LQ,5yateLn is trained,on the first three shots. One profile per client per shot is stored in he training set. From the three aoffies for each clieuLA_ single reference profile is selected by pairwise comparison of the profile images. The profile yielding the lowest matching distance to the other

Fig. 5 OCR Output of the Underlined Text Input Image

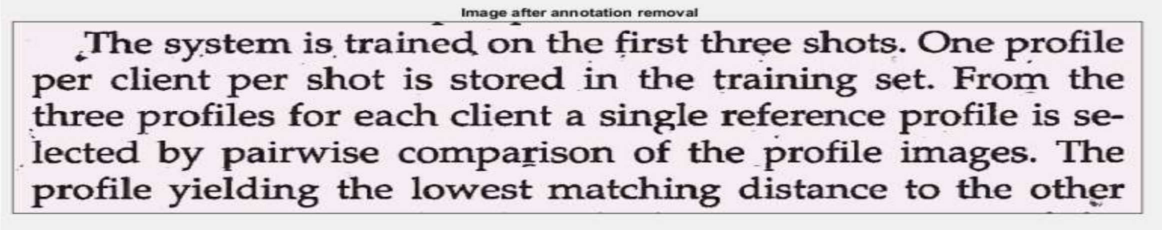
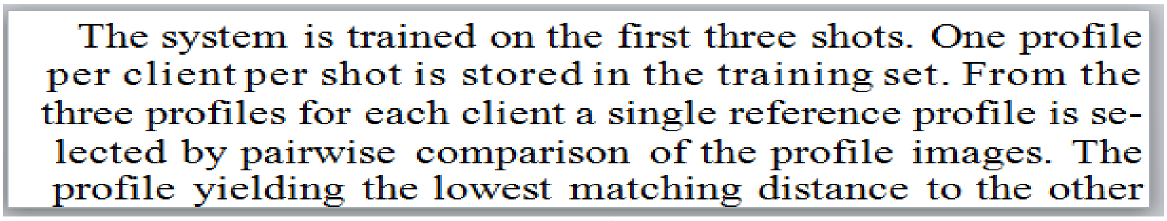


Image after annotation removal

The system is trained on the first three shots. One profile per client per shot is stored in the training set. From the three profiles for each client a single reference profile is selected by pairwise comparison of the profile images. The profile yielding the lowest matching distance to the other

Fig. 6 Output Image after Broken Underlines Removal



The system is trained on the first three shots. One profile per client per shot is stored in the training set. From the three profiles for each client a single reference profile is selected by pairwise comparison of the profile images. The profile yielding the lowest matching distance to the other

Fig. 7 OCR Output of Output Image

e) Following is the table which shows the comparison between the precision and recall ratio of the input image marked with broken underlines and the resultant output after underline removal.

Table -1 Precision Ratio and Recall Ratio of Input Image (Marked With Broken Underlines) and Output Image (After Underline Removal)

Type of image	No of annotation lines	True positive	False positive	False Negative	Total characters in the image	Recall Ratio	Precision ratio
a) i/p image with broken lines	29 detected	131	16	2	149	.9849	.8911
b) o/p image with broken line annotation removal	29 removed	149	0	0	149	1	1

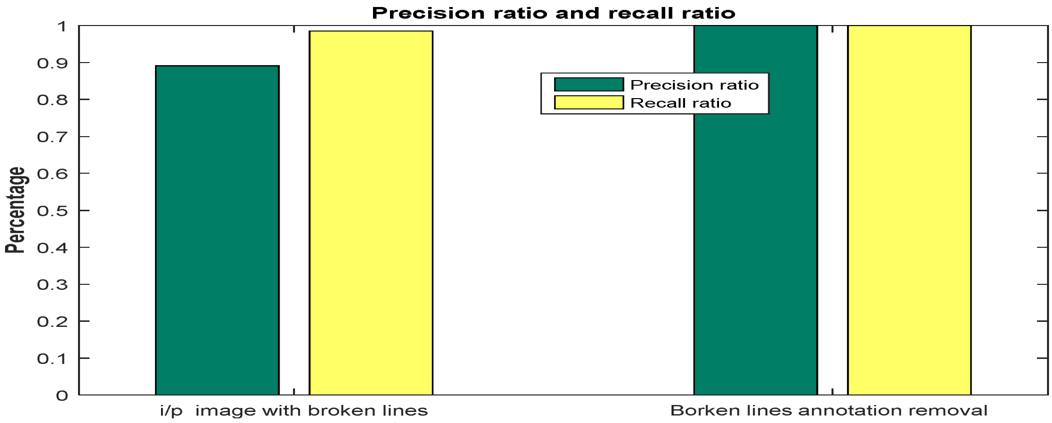


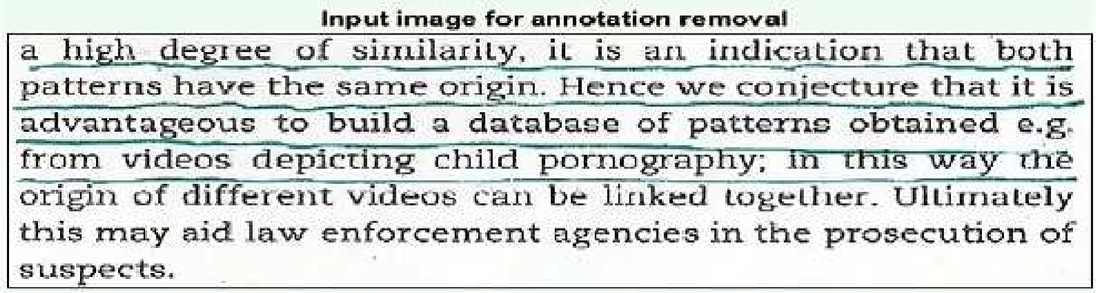
Fig. 8 Graphs Depicting Precision and Recall Ratio of the Input Underlined Text Image and the Output Image with Underline Removal

f) From the Fig. 8, it is observed that the proposed method improves the precision ratio and recall ratio of the input image thereby improving the working of the OCR system. Similarly the annotation removal technique was applied to the text images marked with different types of annotations (such as straight lines, circular lines, elliptical, strokes or embossed lines etc) and underlines (straight, curved, touched, untouched, bent, broken, elliptical) and the resultant output obtained was free from the marks made by user. Also the images were tested by marking the text with different coloured pen like green, red etc and the output obtained for the following are as below.

Input Image with Green Coloured Underlines

a) Following is an input text image marked with green coloured pen. In similar manner to the removal of broken underlines discussed above, algorithm is applied to the image so as to remove the underlines as shown in Fig.9.

b) After applying the ANN based algorithm to the input image for detection and removal of the underlines from the input text image the image so obtained is as shown in Fig. 10.


Input image for annotation removal

a high degree of similarity, it is an indication that both patterns have the same origin. Hence we conjecture that it is advantageous to build a database of patterns obtained e.g. from videos depicting child pornography; in this way the origin of different videos can be linked together. Ultimately this may aid law enforcement agencies in the prosecution of suspects.

Fig. 9 Input Image with Green Coloured Underlines

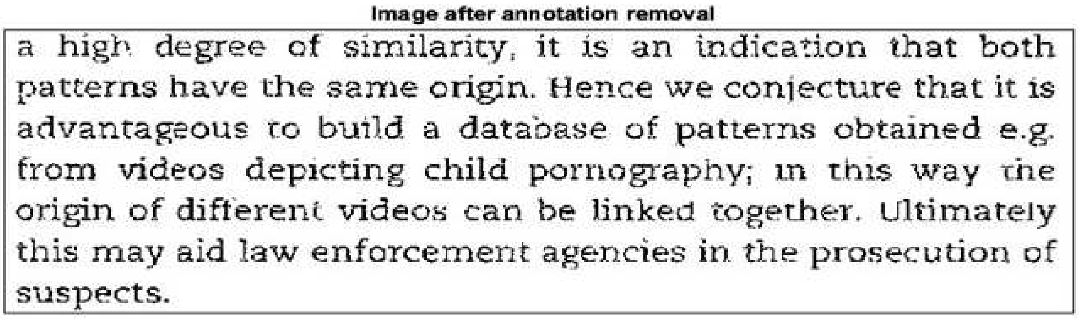


Image after annotation removal

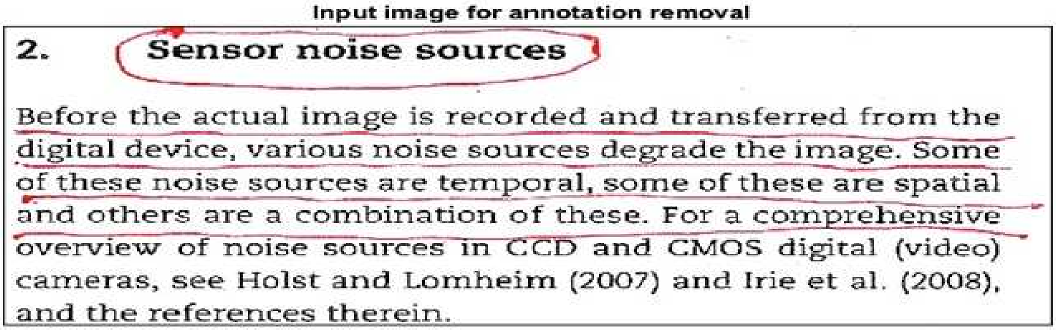
a high degree of similarity, it is an indication that both patterns have the same origin. Hence we conjecture that it is advantageous to build a database of patterns obtained e.g. from videos depicting child pornography; in this way the origin of different videos can be linked together. Ultimately this may aid law enforcement agencies in the prosecution of suspects.

Fig. 10 Output Image with Removal of Green Coloured Underlines

Input Image with Red Coloured Elliptical Annotation and Underlines

a) The proposed methodology was also applied to input text image marked with red coloured elliptical annotations and underlines, following are an input text image with red coloured underlines as shown in Fig. 11.

b) After applying the proposed method, the underlines and elliptical annotations are removed successfully and hence it in turn improved the efficiency of the OCR system, as the output image can be read clearly by the OCR as compared to the input image corrupted by noise, following is the output image so obtained as shown in Fig. 12.



Input image for annotation removal

2. Sensor noise sources

Before the actual image is recorded and transferred from the digital device, various noise sources degrade the image. Some of these noise sources are temporal, some of these are spatial and others are a combination of these. For a comprehensive overview of noise sources in CCD and CMOS digital (video) cameras, see Holst and Lomheim (2007) and Irie et al. (2008), and the references therein.

Fig. 11 Input image with red coloured annotation marks and underlines

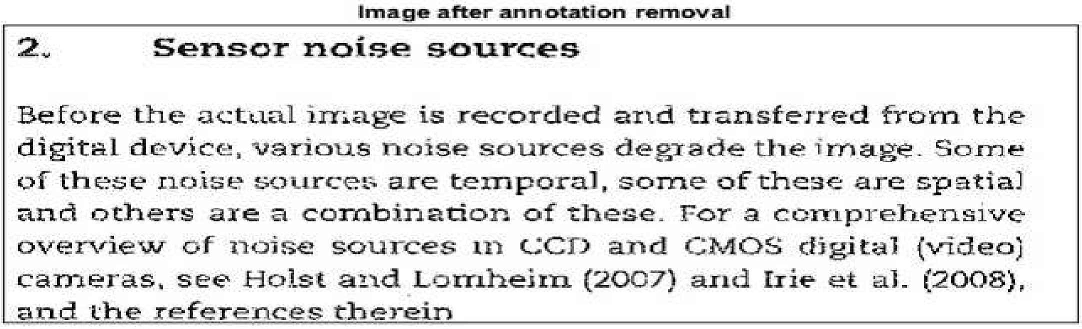


Image after annotation removal

2. Sensor noise sources

Before the actual image is recorded and transferred from the digital device, various noise sources degrade the image. Some of these noise sources are temporal, some of these are spatial and others are a combination of these. For a comprehensive overview of noise sources in CCD and CMOS digital (video) cameras, see Holst and Lomheim (2007) and Irie et al. (2008), and the references therein

Fig. 12 Output Image with Removal of Red Coloured Annotation Marks and Underlines

CONCLUSION

In this work, we have explored the problem and tried to find out solution for removing various types of annotation lines in a document in single running of the algorithm. In this we used Gabor filters to enhance the image character region as well as annotated region from the given text. In this way, Gabor filters help in enhancing the edges in the image when applied in various directions on an image. After this Fuzzy-C-Means is applied to do clustering which results in segmentation of whole text image into number of clusters. FCM uses Euclidian distance for making clusters and hence clusters the annotation lines in separate clusters from the characters. After that two classes are obtained by using region merging in which annotation line pixels are set as first class and rest of the image as second class. These two classes are then trained with artificial neural networks to obtain the ANN object. After these different images are tested using the trained object and annotation line region is inpainted with the desired alphabet or background colour. Experimental results have been carried out on images marked different types of annotation lines and underlines using different coloured pen, good precision ratio and recall ratio is obtained for output image as compared to the annotated input image. The proposed algorithm gives good results in almost all annotations except the embossed cross lines

This algorithm uses the intensity of the image to remove the annotation lines and underlines. In this we have used different colour annotation lines and algorithm works well for almost all colors. But it fails when characters and annotation lines are of almost same intensity. Therefore in future, the algorithm can be modified to get output for varied closely bound intensities and better in-painting of the characters as well as the background region can be worked upon.

REFERENCES

[1] ZL Bai and Q Huo, Underline Detection and Removal in a Document Image using Multiple Strategies, *Proceedings of the 17th International Conference on Pattern Recognition*, Cambridge ,UK, **2004,** 578-581.

[2] KR Arvind, J Kumar and AG Ramakrishnan, Line Removal and Restoration of Handwritten Strokes, *Proceedings on Computational Intelligence and Multimedia Applications*, Tamil Nadu, India, **2007**, 208-214.

[3] Sanjoy Pratihar, Partha Bhowmick, Shamik Sural and Jayanta Mukhopadhyay, Detection and Removal of Hand-Drawn Underlines in a Document Image using Approximate Digital Straightness, *DAR'12*, India, **2012,** 124-131.

[4] Sanjoy Pratihar, Partha Bhowmick, Shamik Sural and Jayanta Mukhopadhyay, Removal of Hand-Drawn Annotation Lines from Document Images by Digital Geometric Analysis and Inpainting, *2013 Fourth National Conference on Computer Vision, Pattern Recognition, Image Processing and Graphics*, Indian Institute Of Technology, Jodhpur, India, **2013,** 1-4.

[5] Y Govindaraju and S.H Srihari, Separating Handwritten Text From Interfering Strokes, published in the book by Sebastiano Impedovo named *From Pixels to Features III - Frontiers in Hand- Writing Recognition*, North Holland Publication University of Michigan ,**1992**, 1728.

[6] B Yu and AK Jain, A Generic System for Form Dropout, *IEEE Trans. on PAMI*, **1996**, 18 (11), 1127-1134.

[7] Supriya Das and Purnendu Banerje, Gabor Filter Based Hand-Drawn Underline Removal in Printed Documents, *IEEE First International Conference Automation, Control, Energy and System,* Hooghly, India, **2014**, 1-4.

[8] R Klette and A Rosenfeld, *Digital Geometry: Geometric Methods for Digital Picture Analysis*, 1st Edition, Morgan Kaufmann, San Francisco, **2004.**

[9] A Telea, An Image Inpainting Technique Based on the Fast Marching Method, *Graphics, GPU, and Game Tools*, **2004** , 9 (1), 25-36.

[10] RC Gonzalez and RE Woods, *Digital Image Processing*, 3rd edition Pearson Education, New Jersey, **2009**.

[11] JG Daugman, Complete Discrete 2-D Gabor Transforms by Neural Networks for Image Analysis and Compression, *IEEE Transaction on Acoustic Speech Signal Processing*, **1988**, 36 (7), 1169–1179.

[12] MS Obaidat, MA Suhail and B Sadoun, An Intelligent Simulation Methodology to Characterize Defects in Materials, *Information Sciences,* **2001,** 137 (1-4), 33-41.

[13] Amitava Roy, P Barat and Swapan Kumar De, Material Classification through Neural Networks, *Ultrasonics*, **1995,** 33 (3), 175-180.

[14] Karray and De Silva, *Soft Computing and Intelligent Systems Design*, Addison Wesley, Pearson, New Delhi, **2004.**

[15] Tadbeer Kaur and Rinkesh Mittal, Detailed Study of Detection and Removal Techniques of Underlines and Annotations in Scanned Documents, *National Conference on RTICCN-2015,* CGC-COE, Punjab, India, **2015.**

[16] Online OCR System, *http://www.onlineocr.net/*

[17] Tanzila Saba, Amjad Rehman, Ayman Altameem and Mueen Uddin, Annotated Comparisons of Proposed Pre-processing Techniques for Script Recognition, *Neural Computing and Applications*, **2014,** 25(6), 1337-1347.

[18] P Nagabhushan, Rachida Hannane, Abdessamad Elboushaki and Mohammad Javed, Automatic Removal of Handwritten Annotations from between Text-Lines and Inside Text-Line Regions of a Printed Text Document, *International Conference on Advanced Computing Technologies and Applications,Mumbai,***2015,**205-214.

[19] Abdessamad Elboushaki, Rachida Hannane, Mohammad Javed and P Nagabhushan, Automatic Removal of Marginal Annotations in Printed Text Document, *Proceedings of Second International Conference on Emerging Research on Computing, Information, Communication Application*, **2014,**1, 123-131.