A Survey On Improving Recto-Verso Degraded Documents

Jensiya Asirvatham
Department of Computer Science and Engineering, Karunya University
India

Abstract

The double sided documents many get degraded due to aging, or by ink, where the ink may spread from backside of document to front side i.e.(ink bleed through). It causes human readability problem, loss of historical information. These problems can be solved by the use of bleed-through ink removal methods. This paper is based on the survey of bleed through ink removal techniques used to restore recto-verso degraded documents. Each technique has got its own advantages as well as disadvantages.

Index Terms— Degraded documents-bleed through-ink removal

I. Introduction

Historical or manuscript documents are valuable resources which are worth to preserve in order to support our culture and knowledge. If we take a document it will have two sides, the front side of the document is recto (front) side whereas the backside of the same document is verso (back). When we write some text on a recto side paper either hand written or the printed it gives the impression of text on its verso (back) side. Impression of text is caused by various reasons. When we give more ink pressure while writing then more dark will be the impression on back side of the document. Now if we write the text on this impression (back) side it is not readable clearly by human as well as in scanned documents. This back side impression of the text is called ink bleed. Ink bleed is one of the major problems while reading the older documents or the manuscripts.

Bleed-through occurs when paper is too thin or the ink applied too heavily the color can bleed or seep through to the other side. This bleed-through mainly appears in old documents and low quality paper. If the bleed-through occurs high in level then reading and extracting the text will be a hectic process.

The severity and characteristics of ink-bleed through depends upon the following factors:
(1) Ink's chemical makeup
(2) Paper's physical and chemical construction
(3) Amount of ink applied and paper's thickness (both spatially varying)
(4) Document's age Amount of humidity in the environment housing the documents.

A degraded image will contain many problems to restore it successfully. The main problem is while restoring information loss may occur. Recovering original document from degraded document is difficult due to high interference of verso –side document. When the sources are largely correlated between each other then removing even residual interferences and separating the sources will be a risky task. Text that interfere from a side of the page to the other of the page are always blurred. Some initially tried to remove the ink-bleed in the document physically. In which the document is undergo a chemical wash that removes ink from it.
Limitations of this approach are:
(1) It will take large amount of time to reduce the bleed-through in a single bound volume, which require all pages to be unbound first ,remove the ink bleed by wash and dried, and then rebound.
(2) The source material of the document may get damage by handling it wrongly
(3) Lightly written foreground text may also get erased.

There are two methods to restoring degradation in a document they are blind and non-blind methods. Blind method is mainly based on intensity information. It takes single side degraded document and make distinction between foreground text and bleed through intensities. Many intensity based classification algorithm are available. They are hysteresis based thresholding, principal component analysis, independent component analysis.

Some of the limitations in blind based method is when bleed-through intensity is equivalent or darker than foreground text, then the intensity information is not enough to remove the bleed-through. Non-blind method considers both side of the document. A model based approach, kullback-leiber thresholding algorithm and binarisation algorithm of gatos, sauvola and pietikainen adaptive binarisation algorithm are some algorithm used in non-blind method. Disadvantage of non-blind method is it work only with registered recto-verso side document. In section 2 we discuss about various blind and non-blind methodology.

II. Methodology

Bleed- through removal techniques

A. Recto-verso registration, enhancement and segmentation of ancient documents

Bianco, Bruno, Tonazzini, Salerno and Console [1] proposed a method which consists of three steps: recto-verso registration, image enhancement, and image segmentation. It is necessary to restore the original aspect of the document as it appeared before the deterioration, without causing losses of important pieces of information. The procedure to improve the readability of recto-verso pairs of documents affected by bleed-through, starting from the raw grayscale scans.

A.1 Image Registration

Image registration is a process of aligning recto and verso side of the document. Misalignment of document may araise due to many reasons. To reduce the misalignment of the double-sided document image registration is a preliminary step.

A.2 Image Decorrelation

Image decorrelation is one of the image enhancement techniques. First, the foreground text and interference which are overlap in a registered document is view by blind source separation algorithm. Second, reduce the bleed through interference in the registered document

A.3 Image Segmentation

Image segmentation techniques are usually applied to grayscale images, and aim at individuating a threshold grey value, which allows to separate the textured background from the foreground text, i.e. to isolate the written words from the artifacts. The output of segmentation is then a binary image, where text pixels and background pixels are labeled. The segmentation techniques can be of two types: global and local.

The global techniques are applied to the whole image without distinction, while the local ones apply different types of thresholding, according to the various regions of the image.

By analysing this technique, decorrelation is effective as a pre-processing technique to make segmentation more reliable, especially when there is a poor contrast between the writings and the interfering patterns but this can't implement in color images.

B. A Recursive technique

Drira and Emptoz [2] proposed recursive approach to remove bleed through depends upon two major analysis, they are: Principal component analysis and K-means clustering algorithm. Degraded document will normally have three class problem: background, original text, interfering text in a page. This last class should be removed and replaced by the background. For this just involving one step clustering is not enough to remove the bleed through. So the recursive segmentation two step clustering–k means clustering and Principal Component Analysis is involved.

Principal component analysis which is mainly used for the reduction of dimensionality and feature extraction. It is used to reduce the correlation between different components where coherent patterns can be detected. After the dimension reduction and decorrelation of data using PCA,K means clustering is applied to it.

K-means clustering is a two-step process. The mean vector for all prototype in each cluster is find out and reassign the cluster cluster according to the closest prototype. The method starts with single cluster and then partition into subsets. Each subdivision leads to binary tree. The last iteration gives the final leaf which contains the original text.

C. Wavlet technique

Chew Lim Tan, Ruini Cao, and Peiyi Shen [3] proposed a wavlet technique which is used to remove the bleed through degradation in a document. The first process is to take the recto and verso side of a image into two images. Both images will have a same dimension. The second process is to overlay the image. The reverse side image is flipped and overlay with front image and then subtract the low intensity values. The strong interfering strokes will pass through the reverse side. These interfering strokes are removed

When removing interfering strokes some of the foreground strokes may get afftected. Therefore impaired foreground strokes on front side is consider and enhance them using wavlets. The detected foreground stroke is then compare with binarized foreground overlay image form the enhancement feature image. The same way impaired foreground strokes on reverse side is detected and them compare it with binarized strokes from reverse side and form the smearing feature image.

Canny edge detector which is used to detect the edges of the binarized image. The smearing effect will be cancelled by the enhancement process.

D. Inpainting technique

Eric Dubois and Patrick Dano [4] proposed a algorithm to restore the interference and bleed through in recto and verso side. This algorithm includes four steps :registration ,segmentation, inpainting, compression. This method is different from other method due to inpainting. Registration of both recto and flip verso side is important so that there will be a perfect alignment of bleed through on one side with the original text on other side. Segmentation which is used to identify the bleed through area. Each side of the document is segmentated into four regions like foreground, background, bleed through, mixed bleed through and foreground. Once the bleed

through area is identified in segmentation, it is then remove and the missing region is replace with background this way is known as inpainting. Compression technique is used for efficient storage and transmission.

E. Correlated Component Analysis

Anna Tonazzini and Bedini [5] proposed correlated component analysis to separate recto and verso color document from pairs of overlapped recto–verso side degraded documents without blur and all residual interferences. This method is based on based on second order statistics which work in fourier domain. It is fast method when compared to all other existing methods. Separation of text can be achieved also when the individual sources are largely correlated. It is used to remove even residual interference and while working in fourier domain which is used to prevent loss of information.

E.1 Image Enhancement

The aim of image enhancement is to improve the interpretability or perception of information in images for human viewers, or to provide `better' input for other automated image processing techniques. It involves sharpening, edge detection, filtering, classification, remove image from blur. Degraded Recto-Verso image is feed as source input, in which it is in time domain. The first step is to convert the time domain image into fourier domain, so that no information will be get loss. In fourier domain all process will get execute easily. Using fast fourier transform convert image from time domain to fourier domain.

The degraded recto-verso image in fourier mode is taken into account and estimate the point spread function with the help of Gaussian which is account for blur. Blur is unsharp image area caused by camera or subject movement, inaccurate focusing, or the use of an aperture that gives shallow depth of field. The Blur effects are filters that smooth transitions and decrease contrast by averaging the pixels next to hard edges of defined lines and areas where there are significant color transition. Gaussian blur is mixed to the degraded recto and verso image separately, in which Gaussian blur is a low pass filter used Combine Gaussian blur to individual recto –verso source image.

The following steps are used in image enhancement using the point spread function which is used to restore the degraded blurred image.

Step 1: Read image
Step 2: Simulate a blur
Step 3: Restore the blurred image using PSFs of various sizes
Step 4: Improving the restoration
Step 5: Using additional constraints on the PSF restoration

The estimation of blur in show-through pattern is done through the reversing the principle of image deconvolution.

E.2 Estimation of interference

The enhanced image in fourier domain is further partition into circular spectra. The circular spectra is nothing but the fourier spectrum. In circular spectrum the image will be represented as circular bins. Using the circular spectrum of degraded image and source image, estimate the relationship between the source and data. From the relationship estimate the cost function which is used for smoothing the image. Minimize the cost function to estimate the unknown correlation. The amount of interference of blur is estimated with the help of cost function.

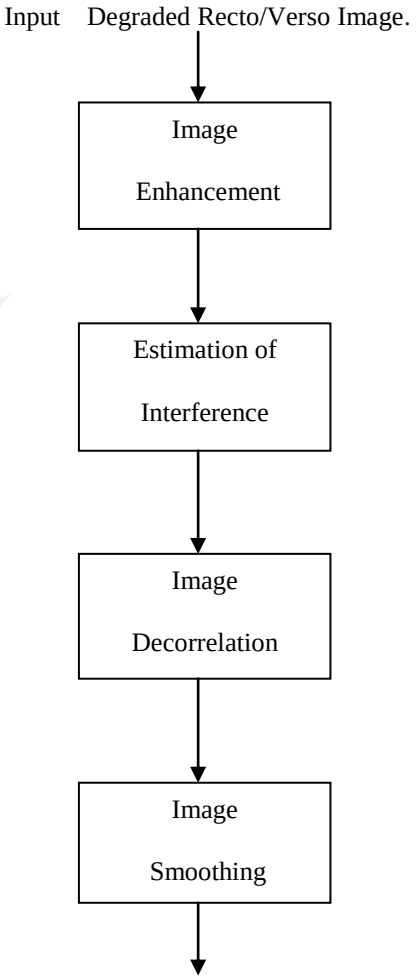

Fig.1 Architecture of CCA methodology

E.3 Image Decorrelation

The next level is to reduce the mixing level of interference in the image. For this reduction, alternate minimization is used. After the minimizer, global regularization is used to image decorrelation, when the high frequency of cross

spectra go quickly to zero then the recto and verso will uncorrelated beyond a certain frequency.

E.4 Image Smoothing

The separated image may be "dirty" (with dots, speckles,stains). Noise removal is used to remove speckles/dots on an image. There are many kinds of noise some are Gaussian noise,salt and peeper noise. Deconvolution is a process of removing blur and noise in the image, after the image decorrelation, the decorrelated image is then feed into inverse filter, which is used to separate the individual sources and blur .The inverse filter doesn't account for noise ,so wiener filter is used to remove remaining noise in the restored image

CCA is able to account for both blur on source and noise on data. Fig 1 shows about the architecture system of the correlated component analysis .The performance can be evaluated using the ground truth.

III. Conclusion

Bleed through is one of the major degradation problem in historical double sided documents. In this paper, the different bleed through removal techniques were analyzed. Other than bleed through removal blur are also considered. The major issue faced by most of the images is residual interference. Various techniques have been examined for the residual interference reduction.

References

[1]G. Bianco, F. Bruno ,A. Tonazzini, E. Salerno ,E. Console.(2004)"Recto-verso registration, enhancement and segmentation of ancient documents", ICDAR,pp.47-86.

[2]F.Drira, H.Emptoz.(2005)"A Recursive Approach For Bleed Through Removal", ICDAR, pp.105-109.

[3]Chew Lim Tan, Ruini Cao, and Peiyi Shen.(2002) "Restoration of Archival Documents Using a Wavelet Technique", IEEE Transactions On Pattern Analysis and Machine Intelligence,pp.1399-1404.

[4]Eric Dubois and Patrick Dano.(2001)"Joint Compression and Restoration of Documents with Bleed-through", IJDAR,pp.1084-1089.

[5]Anna Tonazzini, Bedini (2013) "Restoration of recto–verso colour documents using correlated component analysis", EURASIP Journal on Advances in Signal Processing, pp. 2013:58.