DESIGN OF MORPHOLOGICAL APPROACH TO DETECT AND ELIMINATE INK BLEED IN DOCUMENT IMAGES

Kusum Grewal *

Renu Malhan **

ABSTRACT

When we write some text on a paper either hand written or the printed it gives the impression of text on its back side. More the ink pressure more dark will be the impression on back side. Now if we write the text on this impression side it is not readable clearly. In case of scanned copy of such documents there is the problem to read the documents as well as to extract the actual text from this. 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. In this proposed work we are presenting the way to resolve the problem of ink bleed. Here the research is being performed using the morphological operators to detect and eliminate the ink bleed from a ink-bleeded document.

Keywords: *Ink Bleed, Manuscript, Morphological Operators, Elimination, Detection.*

***** Assistant Professor, ITM Gurgaon.

****** ITM Gurgaon.

I INTRODUCTION

The objective of document image processing is to recognize text and graphics components in images of documents, and to extract the intended information as a human would. Two categories of document image processing can be defined:

A) Textual processing deals with the text components of a document image. Some tasks here are: determining the skew (any tilt at which the document may have been scanned into the computer), finding columns, paragraphs, text lines, and words, and finally recognizing the text (and possibly its attributes such as size, font etc.) by optical character recognition (OCR). Graphics processing deals with the non-textual line and symbol components that make up line diagrams, delimiting straight lines between text sections, company logos etc. Pictures are a third major component of documents.

Document analysis systems will become increasingly more evident in the form of everyday document systems. For instance, OCR systems will be more widely used to store, search, and excerpt from paper-based documents. Page-layout analysis techniques will recognize a particular form, or page format and allow its duplication. Diagrams will be entered from pictures or by hand, and logically edited. Pen-based computers will translate handwritten entries into electronic documents. Archives of paper documents in libraries and engineering companies will be electronically converted for more efficient storage and instant delivery to a home or office computer. Though it will be increasingly the case that documents are produced and reside on a computer, the fact that there are many different systems and protocols, and also the fact that paper is a very comfortable medium for us to deal with, ensures that paper documents will be with us to some degree for many decades to come. The difference will be that they will finally be integrated into our computerized world.

B) Ink-bleed through Housed within the libraries of the world is a great collection of rare books and handwritten manuscripts. The information contained in these collective works was often unavailable to most people. Due to this sometimes high expense of time and money required to travel to the locations in which documents of interest were stored. With the recent advent of the Internet, or more specifically, the growth in computer and telecommunications technology, it is now possible to create online digital libraries enabling most people around the world to potentially access any document anywhere. These digital libraries help in:

- Preserving important information for decades
- Reproduction
- Distribution

- Retrieval
- Analysis

The first step in creating a digital library is to get the documents into digital format. The most popular method of digitizing any picture is to have it scanned. These scanned documents usually suffer from much degradation that can occur either during scanning or because of physical conditions.

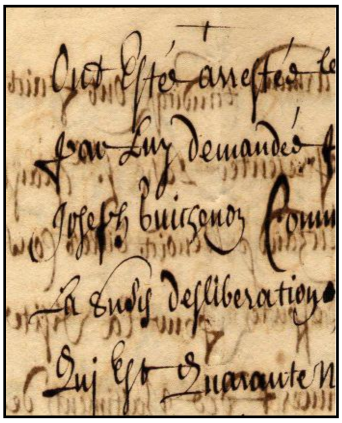

Figure 1 : Examples of ink-bleed through.

This thesis work concentrates on physical degradation that is “Ink-bleed through”. This degradation is due to ink’s seeping through the pages of documents after long periods of storage. The result is that characters from the reverse side appear as noise on the front side. Indeed, the content of the original side is combined with the content of the reverse side (Fig 1.2). This can deteriorate the legibility of the document if the interference acts in a significant way. The severity and characteristics of ink-bleed is related to a variety of factors including the ink’s chemical makeup, the paper’s physical and chemical construction, the amount of ink applied and the paper’s thickness (both spatially varying), the document’s age, and the amount of humidity in the environment housing the documents.

Quantitative Measures

Ink-bleed through removal of a document image is carried out to enhance its readability and to get a clean image by using any of ink-bleed through removal approaches so that these images can be further used for reading or for OCR. It helps in performing segmentation in OCR as it is mainly used for pre-processing of an image. The quality of ink-bleed through removal algorithm is measured by using the following quantitative measures:

1. Precision: It shows how well the system can remove the interfering strokes. It can be evaluated as:

$$\Pr ecision = \frac{\text{No.of wordscorrectlydetected}}{\text{Totalno.of wordsdetected}}$$

2. Recall: It is a measure of the performance of the system in restoring the front page to its original state. It can be evaluated as:

$$\mathrm{Re}\,call = \frac{\text{No.of wordscorrectlydetected}}{\text{Totalno.of wordsin the document}}$$

3. F-Measure: A measure that combines precision and recall, it is the harmonic mean of precision and recall, the traditional F-measure or balanced F-score:

$$F = 2\,\frac{precision \times recall}{precision + recall}$$

This is also known as the F_1 measure, because recall and precision are evenly weighted. It is a special case of the general F_β measure (for non-negative real values of β):

$$F_\beta = (1 + \beta^2)\,\frac{Precision * Recall}{\beta^2 * Precision + Recall}$$

Two other commonly used F measures are the F_2 measure, which weights recall higher than precision, and the $F_{0.5}$ measure, which puts more emphasis on precision than recall.

The obvious drawback of ink-bleed is the reduction in the document's legibility. The motivation of our work is to provide a practical approach to reduce ink-bleed interference in imaged documents in order to improve legibility.

II LITERATURE SURVEY

Huang Y. et al. [1], presented a novel user-assisted approach for Ink-bleed through removal found in old manuscripts. The problem is addressed by first having the user provide simple examples of foreground ink, Ink-bleed, and the manuscript's background. From this user-labeled data, each pixel is classified as foreground, Ink-bleed, or background and used as the data costs of a dual-layer Markov random field (MRF) that simultaneously labels all pixels in both the front and back sides of the manuscript. This user-assisted approach produces better results than existing algorithms without the need for extensive parameter tuning or prior assumptions about the Ink-bleed intensity characteristics. Overall application framework was discussed along with details of the features used in the data costs, a comparison between K-nearest neighbour and support vector machine for likelihood estimation, the dual-layer MRF formulation with associated inter- and intra-layer costs, and a comparison of this approach against other ink-bleed reduction algorithms. Overall procedure that was adopted is as follows:

1) Image alignment with local refinement;
2) Training-data collection via minimal user-assistance;

3) Pixels labelling using the dual-layer MRF framework;

4) Output image generation.

The results demonstrate that DL-MRF-SVM and DL-MRF-KNN approaches can generate results superior to previous approaches.

Wolf C. [2], presented a method for blind document bleed through removal based on separate Markov Random Field (MRF) regularization for the recto and for the verso side, where separate priors were derived from the full graph. The segmentation algorithm is based on Bayesian Maximum a Posteriori (MAP) estimation. He concentrated on ink bleed through removal, i.e. the separation of a single scanned document image, into a recto side and a verso side. The novelty of the method is the separation of the MRF prior into two different label fields, each of which regularizes one of the two sides of the document. This separation allows to estimate the verso pixels of the document which are covered by the recto pixels, which, again through the MRF prior, improves the estimation of the verso pixels not covered by recto pixels, thus increasing the performance of the regularization. He showed that this formulation leads to an efficient algorithm based on graph cuts.

The performance of the method has been evaluated on scanned document images from the 18th century, showing that the restoration is able to improve the recognition performance of an OCR significantly, compared to non restored images but also compared to competing methods.

Moghaddam R. et al. [3], adapted the variational model by using an estimated background according to the availability of the verso side of the document image. In this approach, based on the availability of the verso side, different models have been developed which can be applied to document images degraded by bleed-through. The first model they present is variational model for restoration of double-sided document images. Then for restoration without the verso side image, a modified variational model is introduced which removes the interference patterns using the estimated background information. For document images containing very fine features and structures, the second model is modified to include a global classifier, the flow field, which helps preserving these very weak edges, while at the same time achieving a high degree of smoothing and enhancement. This is called global control or flow field. The solution of each resulting model was obtained using wavelet shrinkage or a time-stepping scheme, depending on the complexity and nonlinearity of the models. When both sides of the document are available, the proposed model uses the reverse diffusion process for the enhancement of double-sided document images. The results of experiments with real and synthesized samples are promising. The proposed model, which is robust with respect to noise and complex background, can also be applied to other fields of image processing. The performance of the models was evaluated against other methods, such as the ICA method, in both subjective and objective ways using several databases of different types of document script and degradation.

Tonazzini A. et al. [4], proposed a system to process multispectral scans of double-sided documents. It can co-register any number of recto and verso channel maps, and reduce the bleed-through/show-through distortions by exploiting blind source separation. From RGB scans, it is also able to recover the original colors, thus improving the readability of a document while maintaining its original appearance. The recto and verso patterns obtained can then be further analyzed. The aim is twofold: to produce enhanced versions of all the available scans at the different wavelengths, and a restored visible document that, while cleansed of the unwanted interferences, maintains its useful features as much as possible .this approach mainly consist of two steps:

1) Registration of multispectral scans of recto-verso pairs
2) Removing interference from registered data

Experimental results shows that for RGB reconstructions of two RGB recto-verso scans , a significant attenuation of the bleed-through has been obtained, and the original color has been pretty well recovered. The RGB recto and verso images thus obtained can further be analyzed to extract possible extra uncorrelated patterns. The overall procedure could constitute a fast, reliable and effective system to be routinely used in libraries and archives for the enhancement of multispectral scans of degraded documents. The method is flexible for use in various contexts of document analysis, such as the extraction of hidden or masked patterns.

III **PROPOSED WORK**

A morphological operator is therefore defined by its structuring element and the applied set operator. For the basic morphological operators the structuring element contains only foreground pixels (*i.e.* ones) and `don't care's'. These operators, which are all a combination of erosion and dilation, are often used to select or suppress features of a certain shape, *e.g.* removing noise from images or selecting objects with a particular direction.

In erosion, every object pixel that is touching a background pixel is changed into a background pixel. In dilation, every background pixel that is touching an object pixel is changed into an object pixel. Erosion makes the objects smaller, and can break a single object into multiple objects.

Dilation makes the objects larger, and can merge multiple objects into one. As shown in (d), opening is defined as an erosion followed by a dilation. Figure (e) shows the opposite operation of closing, defined as a dilation followed by an erosion. As illustrated by these examples, opening removes small islands and thin filaments of object pixels. Likewise, closing removes islands and thin filaments of background pixels.

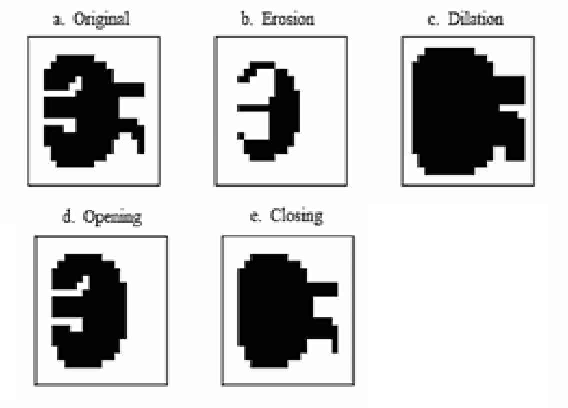

Figure. 2 Morphological operations. Four basic morphological operations are used in the processing of binary image: erosion, dilation, opening and closing.Fig (a) shows an example binary image. Fig (b) to (e) shows the result of applying these operations to the image in (a).

The proposed architecture is given as figure 3.

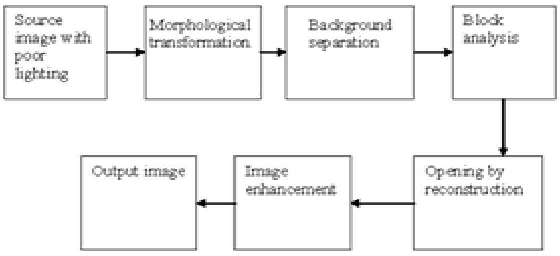
Source image with poor lighting
Morphological transformation
Background separation
Block analysis
Opening by reconstruction
Image enhancement
Output image

Figure 3 : Proposed Architecture

This method is introduced to overcome the limitation of thresholding in Blind segmentation approach. The proposed algorithm is as follows:

ALGORITHM

Step 1: Read the input image degraded from ink-bleed through.

Step 2: Apply the morphological operations like Erosion, Dilation, Opening, Closing on the source image.

Step 3: Perform the block analysis on the image to find the blocks of different intensity level and that requires the enhancement.

Step 4: Reconstruct the blocks separately.

Step 5: Perform Image Enhancement. The lighter part will be dissolve and the darker is left

Step 6: Increase the contrast of remaining image.

Step 7: Show the Restored output image.

IV CONCLUSION

Here a new algorithm is presented to remove the ink bleed impression from the document images. The morphological approach gives a segmented approach and it appear the results will be better and efficient then existing approaches.

REFERENCES

1. Y.Huang, M.S. Brown, and D. Xu, Member, IEEE,"User-Assisted Ink-Bleed Reduction", IEEE transaction on image processing, pp.2646-2658, vol.19, no. 10, October 2010.
2. C .Wolf. , "Document ink bleed-through removal with two hidden markov random fields and a single observation field", IEEE transaction on pattern analysis and machine intelligence, pp.431-447, vol.32, no.3, march 2010.
3. R. F. Moghaddam, M. Cheriet, Senior Member, IEEE, "A variational approach to degraded document enhancement", IEEE transaction on pattern analysis and machine intelligence, pp.1347-1361, vol. 32, no. 8, August 2010.
4. A. Tonazzini, G. Bianco, E. Salerno, "Registration and enhancement of double-sided degraded manuscripts acquired in multispectral modality", in Proceedings of 10th International Conference on Document Analysis and Recognition, 2009.
5. J. Wang, M. S. Brown, C. L. Tan, "Accurate Alignment of Double-Sided Manuscripts for Bleed-Through Removal," in Proceedings of the eighth International Association for Pattern Recognition(IAPR) International Workshop on Document Analysis Systems, pp.69-75, 2008.
6. C. Wolf, "Improving recto document side restoration with an estimation of the verso side from a single scanned page", in Proceedings of international Conference of Pattern recognition (ICPR), pp.1-4, 2008.
7. A. Tonazzini, E. Salerno, and L. Bedini, "Fast correction of bleed through distortion in grayscale documents by a Blind Source Separation technique", International

Journal on Document Analysis and Recognition (IJDAR), vol. 10, pp.17–25, June 2007.

8. R. D. Lins and J. M. M. da Silva, "Assessing Algorithms to Remove Back-to-Front Interference in Documents", in Proceedings of ITS-2006, Fortaleza, Brazil, IEEE Press, pp.868-873, 2006.
9. F. Drira, F. L. Bourgeois, and H. Emptoz, " Restoring ink bleed-through degraded document images using a recursive unsupervised classification technique", In Document Analysis Systems, pp.38–49, 2006.
10. F. Drira , "Towards Restoring Historic Documents Degraded Over Time," dial, in Proceedings of Second International Conference on Document Image Analysis for Libraries (DIAL'06), pp.350-357, 2006.
11. R.C. Pinto, L. Bandeira, M.C. Sousa, P. Pina, "Combining Fuzzy Clustering and Morphological Methods for Old Documents Recovery", in Proceedings of the Iberian Conference on Pattern Recognition and Image Analysis (IBPRIA), pp.387-394, 2005.
12. Q. Wang, T. Xia, L. Li, C. L. Tan, "Document Image Enhancement Using Directional Wavelet," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '03) , vol. 2, pp.534, 2003.
13. H. Nishida and T. Suzuki, "Correcting show-through effects on document images by multiscale analysis", in Proceedings of the International Conference on Pattern Recognition(ICPR) , vol. 3, pp.65–68, 2002.
14. E. Dubois and A. Pathak, "Reduction of bleed-through in scanned manuscript documents", in Proceedings of Image Capture Systems (PICS), pp.177–180, 2001.
15. Q.Wang, C.L. Tan, "Matching of Double-Sided Document Images to Remove Interference," IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'01) , vol. 1, pp.1084, 2001.
16. K. Knox, "Show-Through Correction for Two-Sided Documents", United States Patent 5,832,137, Nov. 1998.