Enhanced Bleed Through Removal for Scanned Document Images

Avinash Sharma
Hewlett-Packard Research Labs
Bangalore, India-560029
sharma@hp.com

Sahil Mahaldar
Shell, Inc.
sahil_iit4@yahoo.co.in

Serene Banerjee
Hewlett-Packard Research Labs
Bangalore, India-560029
serene.banerjee@hp.com

Abstract

Back-to-front interference is a common problem in documents, printed on translucent pages with insufficient opacity and is referred to as bleed through. The present state-of-art algorithms address bleed through based on entropy [1-3], entropic correlation [4] and discriminator analysis [5, 10]. However, a common drawback of such algorithms is their inefficient processing of documents that are either sparse in terms of content or have a very dark background. Our proposed algorithm, based on Otsu's binarization method [5] and pixel level classification addresses these problems. Experiments indicate that our algorithm performs comparable to state-of-the-art for most of the images and better than state-of-the-art for the low contrast images.

1. Introduction

Very often, certain parts of printing on the reverse of a document are visible on the front because of insufficient opacity of the paper. This is referred to as bleed through. Although human readability is not affected by presence of bleed through the document aesthetics increase without the presence of it. Machine readability also improves without the presence of bleed through artifacts [11]. Many techniques have been proposed to remove bleed through. However, a common problem with the state-of-art techniques arises when the documents have little useful information sparsely printed on them or the background is very dark/ very bright [3]. In both the cases, existing algorithms do not produce satisfactory results. This paper addresses this problem and proposes an algorithm that works well on such documents without compromising quality. An optimum decision criterion is proposed to classify pixels as foreground, background or bleed through. The proposed method takes advantage of the grayscale intensity spectrum of scanned documents which helps in distinguishing foreground pixel from background and bleed through pixel.

Section 2 reviews prior art in bleed through removal. Section 3 proposes our bleed through removal algorithm which also covers low contrast or sparsely written documents. Section 4 presents the results and does a subjective and objective comparison with state-of-the-art algorithms. Section 5 concludes the paper.

2. Prior art in bleed through removal

Many techniques have been proposed to remove bleed through. Initial methods like valley sharpening technique [7] or the difference histogram method [8, 10], use the information of neighboring pixels (or edges) to modify the histogram of the original image for thresholding. Pun *et al*. introduced the concept of entropy for threshold calculation. Kapur *et al*. calculate entropies of the foreground and background for various thresholds, and choose a threshold that maximizes the sum of them. Johannsen *et al*. minimize sum of entropies of the foreground and the background, and use a log of the foreground entropy. This algorithm does not perform as well as other algorithms [11]. Yen *et al*. use entropic correlation [4] to calculate an optimal threshold. Two of the most efficient algorithms for removing bleed through use entropy of the histogram of the image [1-3]. However, one drawback of global thresholding techniques is that they fail to remove bleed through from low contrast document images. To overcome this shortcoming we propose to modify global thresholding and incorporate a method to determine optimal local thresholds that would remove bleed through based on pixel level classification. Section 4 does a subjective and objective evaluation of the proposed method in removing bleed through, compared to the above algorithms.

3. Problem Statement

This paper presents an algorithm for efficient bleed through removal in documents including those that have very dark background with strong back-to-front interference and documents with sparse content. We present a new approach for "*Trinarization*" in which each pixel in an image is labeled as a foreground pixel (FG) or background pixel (BG) or bleed through pixel (BL). A label is attached to each pixel depending upon its positioning in the intensity space with respect to its neighborhood as well as the whole image. The size of the neighborhood is selected in an adaptive manner to incorporate sufficient information. Local level analysis enables efficient classification of pixels independent of the contrast of the image and the level of bleed through as long it is in the range of being declared as a bleed through.

3.1 Proposed Solution

In the case of uni-modal histogram images, that is, grey level images whose histograms do not have two obvious peaks, Otsu's method can provide a satisfactory result. Therefore, it is referred to as one of the most powerful unsupervised methods for bi-level thresholding in literature and has been used for localized thresholding in the proposed algorithm. Otsu's algorithm [5] uses the zeroth and first order cumulative moments of the grey level histograms to predict a thresholding value. Given an image A, with k grey levels, it predicts a threshold value, k_{thresh} which divides the whole image into two classes of pixels. Let the mean and the variance of the object and background with respect to any arbitrary threshold value't' be denoted by *(M1, V1)* and *(M2, V2)* respectively, and *PT* be the cumulative probability of the foreground. Then, Otsu's algorithm proposes an optimum threshold grey level k_{thresh} such that

$$\alpha(k_{thresh}) = \max(\alpha(t)) \text{ , and}$$

$$\alpha(t) = \frac{PT(1-PT)[M1-M2]^2}{(PT)V1+(1-PT)V2}$$

Every pixel in an image with bleed through can be characterized by its positioning on the intensity space i.e. grayscale values, as being a FG or BG or BL pixel. Fig1 shows the intensity spectrum which is most likely to be followed in case of an image with bleed through. Figure suggests,

$$I_{FG} < I_{BL} < I_{BG}$$

where, I_{pixel_type} denotes the grayscale level of a pixel. This relation is independent of the contrast of the image. Even for an image with a darker background, the foreground will always be darker followed by the bleed through and the background lies on the brighter side of the spectrum. This criterion solely may help a human eye to distinguish between a FG, BG and BL pixels.

Further, a threshold which distinguishes a FG from a BL will be less than a threshold dividing BG and BL pixels. A global threshold separating foreground and background obtained from a standard binarization technique may lie in a region as shown in fig1. An adaptive fraction λ of the global threshold can enable the decision to distinguish a FG from a BL pixel or a BL from a BG pixel for a small neighborhood.

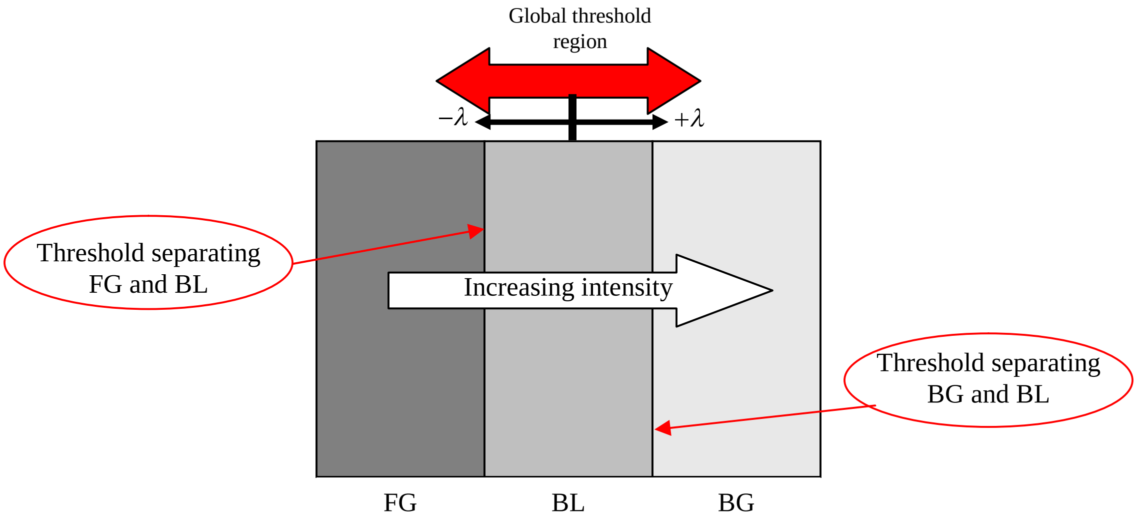


Fig.1: Intensity spectrum for positioning of foreground, background and bleed through pixels

The proposed algorithm works around the above relational equations. For every pixel we try to identify it as a FG, BG or BL pixel with respect to a small neighborhood around it. The neighborhood selection is done on the basis of variational content around the pixel. Typically if a neighborhood is found more or less uniform then its extent is increased to bring in more variation into it. The variation is parameterized by the variance of the intensity values. Generally, a uniform region will have a low variance as compared to a region which contains a mix of FG, BG and BL pixels. The details of the algorithm are explained below:

Given a grayscale image $I(X\ x\ Y)$ with bleed through, normalize the intensity value to lie between 0 and 1 where 0 represents black and 1 represents white. Choose the initial value of radius for the neighborhood as $r = 4$ and maximum radius to be $r\_\max = 4r$. Apply Otsu's algorithm on the whole image to obtain a global threshold ($global\_th$). For every pixel (x, y) on the image compute the variance of the intensity values in a neighborhood of radius r around the pixel to generate a variance image (Var_img). On the variance image larger values corresponds to pixels having more information content in their neighborhood. Analyze Var_img to get a threshold (Var_th) which will distinguish pixels with more information in their locality from the pixels with more or less uniform content in their neighborhood. We again use Otsu's binarization on Var_img for obtaining Var_th . Following steps are performed from here:

```
Initialize $r\_current = r$
For all (x,y) on the image
do
{
    if ($Var\_img(x, y) > Var\_th$)                                --------step 1
    {
        Compute $thresh\_local$ using Otsu's binarization for the local neighborhood around (x,y)
        if ($thresh\_local < (1+\lambda) global\_th$)
            if ($I(x, y) < thresh\_local$)
                $label(x, y) = FG$
            else
                $label(x, y) = BL$
            end
        else
            if ($I(x, y) < thresh\_local$)
                $label(x, y) = BL$
            else
                $label(x, y) = BG$
            end
        end
    }
    else
    {
        $r\_current = r\_current + r$
        Compute $Var\_img(x, y)$ with $r\_current$
        if ($r\_current \sim= r\_\max$)
            GOTO step 1
        else
            Compute $thresh\_local$ using Otsu's binarization for the local neighborhood around (x,y)
            if ($thresh\_local < (1+\lambda) global\_th$)
                if ($I(x, y) < thresh\_local$)
                    $label(x, y) = FG$
                else
                    $label(x, y) = BL$
                end
            else
                if ($I(x, y) < thresh\_local$)
```

```
                        $label(x,y) = BL$
                    $else$
                        $label(x,y) = BG$
                    $end$
                $end$
            $end$
        $\}$
        $end$
$\}$
```

For all the pixels with label BL and BG replace the intensity with 1 and for pixels with label FG replace the intensity with 0. The resulting image is the cleaned up image with 0 representing foreground and 1 representing background with no bleed through. $\lambda \in [-1,1]$ is a variable parameter which decides the appropriate region on the intensity spectrum for classifying a region as containing foreground, bleed through and background or simply bleed through and background. λ is an adaptive parameter which depends on the document type.

4. Experimental Results

The proposed algorithm was tested on a database of document images with varied contents. These included images with low contrasts, images with dark foreground, images with very sparse foreground content etc. The algorithm performed compared to state-of-the-art in most of the cases and better in some cases. We use Mello-Lins algorithm [2] for comparison.

As mentioned, given an image, the proposed method tries to label each pixel as FG, BG or BL. Fig2 shows results after labeling of pixels on few example images showing the original image, the pixels marked as BG and BL and the final result with FG marked in black.

Fig.3 shows the effect of λ on the final image. It shows various outputs for different values of λ . As it can be seen from the figure choosing a value of $\lambda = 0.0$ or $\lambda = 0.1$ gives the best result.

For a quantitative analysis, controlled bleed through was added by superimposing the faded background image on the foreground image on. The fade factor was varied from 0 to 255, where 0 indicates strongest back-to-front interference and 255 indicates the weakest. Three quality factors; text error, paper error and interference error, proposed in [11] were computed. Table1 shows the comparison with state-of-the-art methods. Fig4 shows the output for a fade factor of 120 and 160.

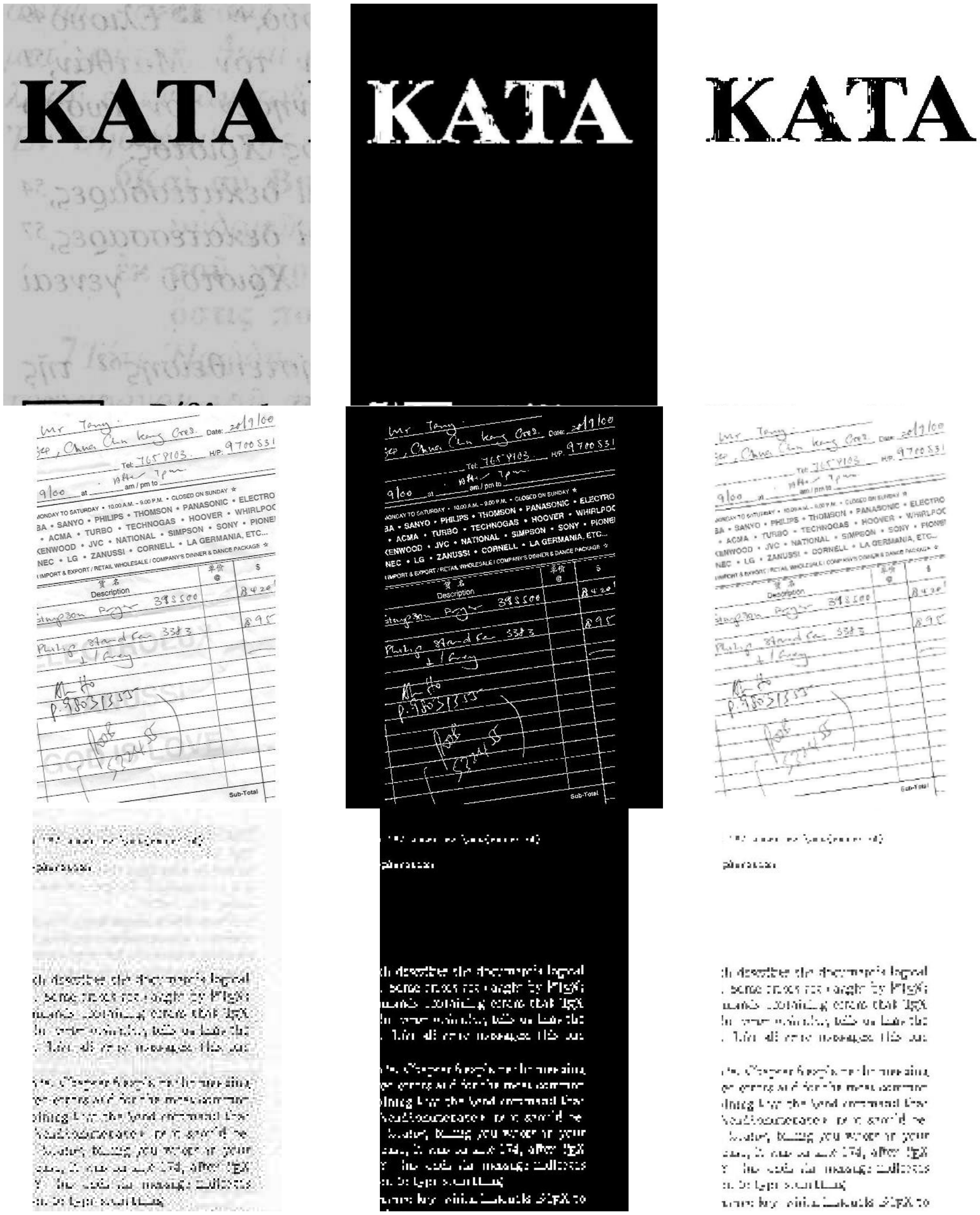
Fig.5 compares the proposed method with the algorithm by Mello-Lins [5] for few sample images. As it can be seen, the proposed algorithm outperforms the algorithm by Mello-Lins [5] significantly. The computational overload of the proposed algorithm is more than the state-of-the-art methods as the computation is performed for all the pixels but there is a significant scope of improvement in this regard.

5. Conclusion

Intensity spectrum of pixels provides rich information about content in an image. We have proposed a new bleed through removal algorithm based on Otsu’s algorithm and intensity space positioning of pixels, to calculate optimum localized thresholds for appropriate labeling of pixels as foreground, background or bleed through pixel. The method tries to solve the bleed through removal problem as a local approach. For a small neighborhood around a pixel containing sufficient variation in content, an optimum decision rule is applied by comparing the local threshold with the fine tuned global threshold. The fine tuning of the global threshold is done through an adaptive parameter. The entire solution intuitively tries to imitate the human capability of distinguishing between a foreground pixel from a bleed through and a background pixel solely by relying on the intensity spectrum of these pixels. The proposed algorithm performs comparable to state of the art methods and better in special cases for documents with sparse foreground content.

Algorithm	**Text Error (%)**	**Paper Error (%)**	**Interference Error (%)**
Johanssen-Bille	0	1,215.66	86.51
Pun	0	492.09	83.90
Yen-Chang-Chang	0	41.38	66.78
Kapur-Sahoo-Wong	0	28.48	57.23
Mello-Lins	0	9.67	32.44
Oysu	0	6.40	26.31
Wu-S.-Hanqing	50.22	0	0
Silva-Lins-Rocha	6.13	0	6.07
Proposed	1.25	**0**	1.25

Table 1: Quantitative comparison on the basis of three quality factor

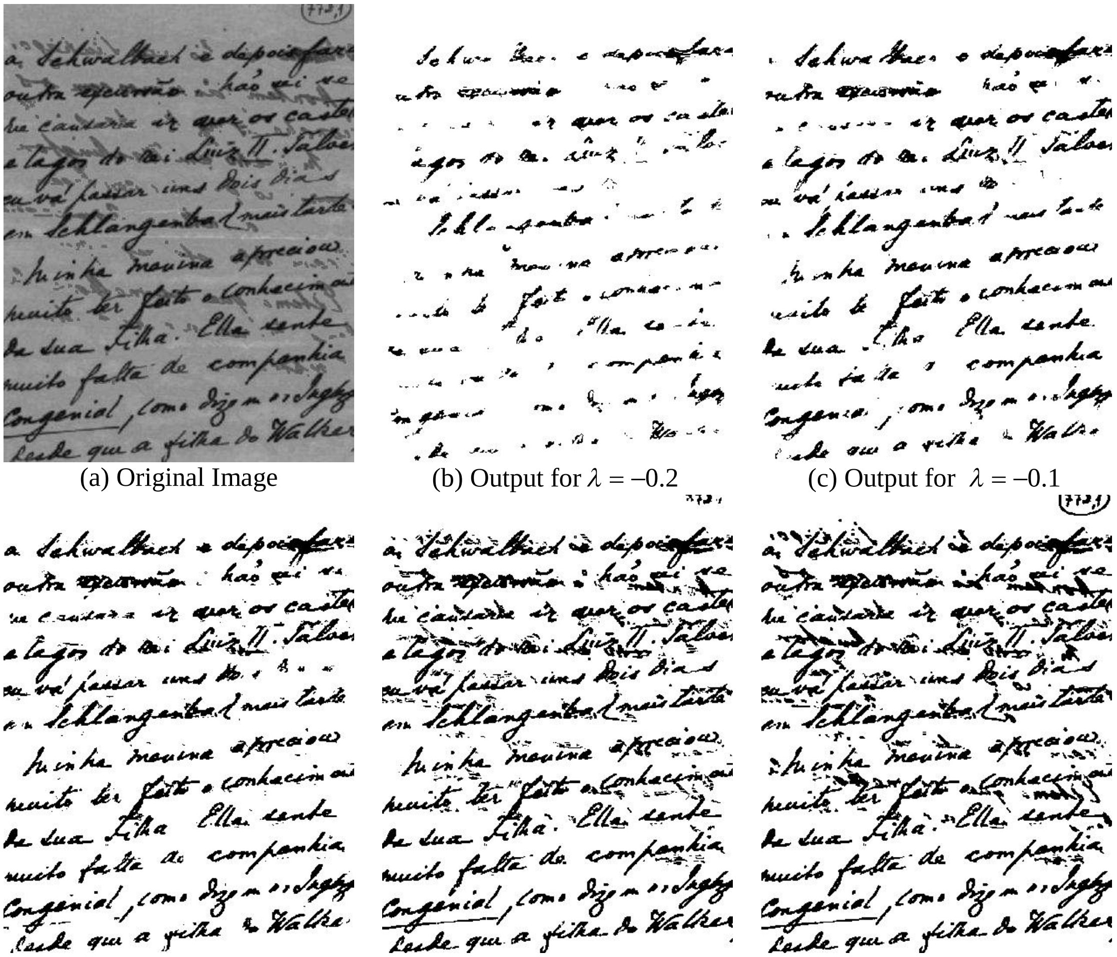


(a) Original Image

(b) Images with BG and BL pixels in black

(c) Images with FG pixels in black

Fig.2 : Images with FG, and (BG+BL) labeling



(a) Original Image

(b) Output for $\lambda = -0.2$

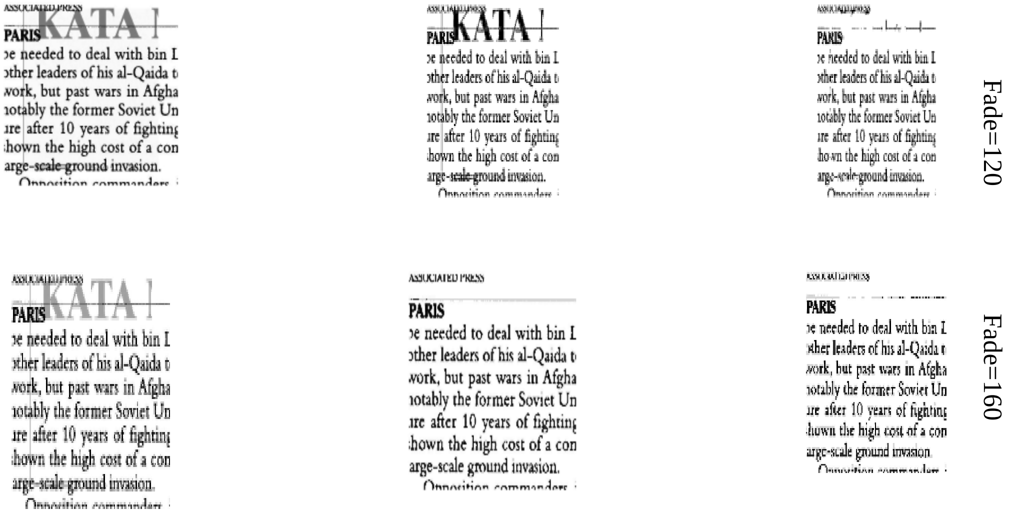
(c) Output for $\lambda = -0.1$

(d) Output for $\lambda = 0.0$

(e) Output for $\lambda = 0.1$

(f) Output for $\lambda = 0.2$

Fig.3 : Effect of λ on the output image

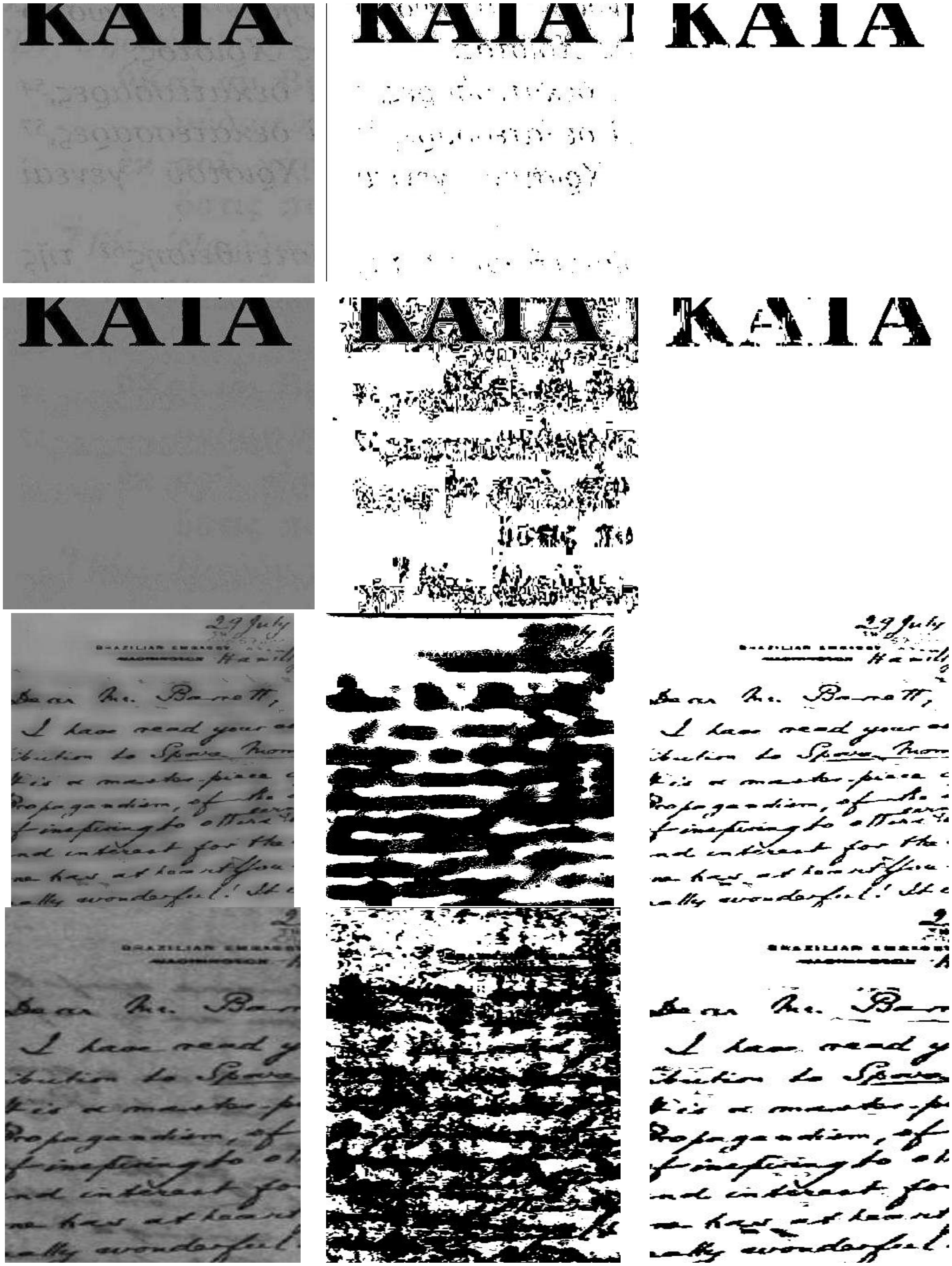

Fade=120
Fade=160

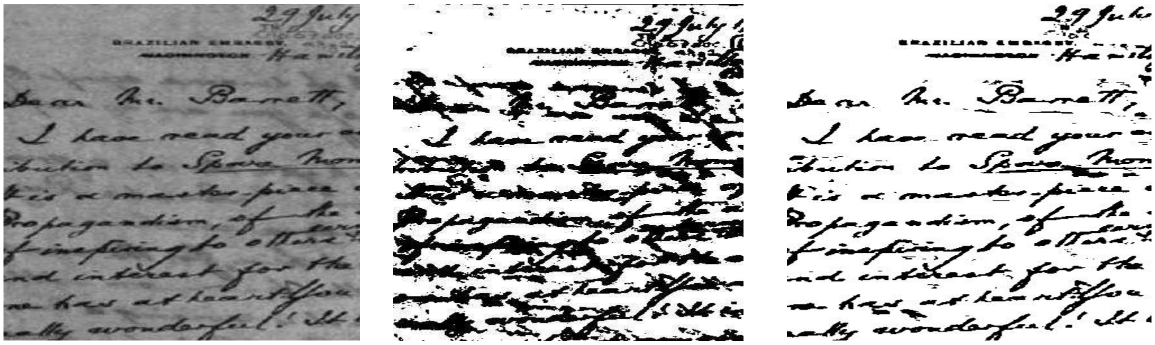
(a) Original image with controlled bleed through

(b) Output from Mello-Lins [5]

(a) Output from proposed method

Fig.4 : Results for an image with controlled bleed through


KAIA
29 July
BRAZILIAN EMBASSY
WASHINGTON
Dear Mr. Barrett,
I have read your



(a) Original Image (b) Mello-Lins (c) Proposed Method

Fig.5: Comparison between Mello-Lins [5] and proposed method

6. References

1. C. A. B. Mello and R. D. Lins. Image segmentation of historical documents, Visual 2000, Mexico City, 2000.
2. C. A. B. Mello and R. D. Lins. Generation of images of historical documents by composition. ACM Doc. Eng., VA, USA, 2002.
3. J. M. M. da Silva, R.D.Lins and V.C.da Rocha Jr. “Binarizing and Filtering Historical Documents with Back-to-Front Interference,” *ACM-SAC, Dijon, France, ACM Press, 2006*.
4. J. C. Yen, F. J. Chang, and S. Chang, “ A new criterion for automatic multilevel Thresholding”, *IEEE Trans. Image Process. IP-4, 1995, pp. 370–378.*
5. N. Otsu, “A threshold selection method from grey level histograms”, *IEEE Trans. on Sys. Man & Cybernetics, 1979, 62–66.*
6. Juan Cheng; Xijian Ping, “Text image retrieval based on generalized normalized picture information measure”, *IEEE International Conf. on Natural Language Processing and Knowledge Engineering, vol. 30, Nov 2005, pp.503-505.*
7. J. S. Weszka, R. N. Nagel, and A. Rosenfeld, "A threshold selection technique,"*IEEE Trans. On Computer, vol. C-23, 1974, pp. 1322 -1326.*
8. S. Watanabe and CYBEST Group. "An automated apparatus for cancer prescreening: CYBEST," *Comp. Graph. Image Process. vol. 3, 1974, pp. 350—358.*
9. Shi Kuo Chang, “Principles of Pictorial Systems Design”, *Prentice Hall, 1989, pp. 61- 81.*
10. R. Maurer, “A Low Complexity Method for Background Smoothing and Bleed-Through Reduction in Two-Sided Scanned Document Images”, *Disclosure, HP Labs, Israel.*
11. R. D. Lins, J. M. M. d. Silva, “A Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents”, *ACM Symp. On Applied Computing, on Mar. 11-15, Seoul, Korea, 2007.*