Enhanced Bleed Through Removal Using Normalized Picture Information Based Measures

Sahil Mahaldar, Serene Banerjee

HP Laboratories
HPL-2009-159

Keyword(s):
bleed through, document cleanup

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 Normalized Picture Information Measure (NPIM) [6] and Otsu's binarization method [5], addresses these problems. Experiments on 111 test images indicates that our algorithm performs comparable to state-of-the-art on 106 images and for the remaining 5 low contrast images our algorithm is able to remove bleed through whereas the state-of-the-art algorithms fail.

External Posting Date: July 21, 2009 [Fulltext] Approved for External Publication
Internal Posting Date: July 21, 2009 [Fulltext]

Enhanced Bleed Through Removal Using Normalized Picture Information Based Measures

Sahil Mahaldar[1]
Shell, Inc.
Sahil.Mahaldar@shell.com

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 Normalized Picture Information Measure (NPIM) [6] and Otsu's binarization method [5], addresses these problems. Experiments on 111 test images indicates that our algorithm performs comparable to state-of-the-art on 106 images and for the remaining 5 low contrast images our algorithm is able to remove bleed through whereas the state-of-the-art algorithms fail.[1]

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 on the other types of documents.

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. Related Work 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 picture information measures. Section 4.1 and 4.2 do 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. Our approach is an adaptation of a global thresholding algorithm (Otsu's algorithm) to be able to select local thresholds that removes bleed through accurately, thereby addressing this problem. It takes into consideration the

[1] This work was done as a part of Sahil Mahaldar's internship in the Paper in the Digital Enterprise project at the Hewlett-Packard Research Labs, Bangalore, India.

Normalized Picture Information Measure (NPIM) values along with the Otsu's algorithm to compute localized threshold values, that help differentiate the document background, bleed through, and the foreground.

3.1. Proposed solution

In the case of uni-modal histogram images, that is, grey leval 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 leval histograms to predict a thresholding value.

Picture Information Measure (PIM) [9] is highly sensitive to the image size. A more efficient measure of the information content of the image can be calculated by normalizing the value of PIM by total number of pixels in the image. An image with all the pixels having constant grey leval values will have Normalized PIM (NPIM) = 0. As such, the higher the number of grey levels in a block the more will be its NPIM or entropy.

2.1.1 Otsu's thresholding algorithm

Otsu's algorithm [5] uses the zeroth and first order cumulative moments of the grey level histograms to predict a threshold 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

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

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

2.1.2 Normalized Picture Information Measure

For an image A, PIM is defined as follow,

$$PIM(A) = \sum_{i=0}^{L-1} n_i - \max n_i$$

where *L* represents the number of grey levels in the image and n_i the number of pixels of that pixel level.

NPIM is obtained by normalizing the value of PIM by total number of pixels in the image denoted by *n*, as

$$NPIM(A) = (1/n)\sum_{i=0}^{L-1} n_i - \max n_i$$

3.2. Proposed NPIM based adapted Otsu's binarization for bleed through removal

If a grey scale image with bleed through is divided into blocks, six types of blocks are possible, one having only foreground pixels, the second one having only background pixels, the third one having foreground and background pixels, the fourth one having bleed through and background, the fifth one having foreground and bleed through pixels, and the sixth one having bleed through background and foreground pixels. The goal will be to screen blocks that have lower grey level variations, and apply Otsu's binarization on blocks with higher grey level variations, as they are likely to have bleed through. We thus use NPIM values to get a measure of how many grey levels are there in a block, compare that with the surrounding, and adapt the block size as necessary to make the NPIM values higher, and then perform Otsu's binarization on the adapted block size to remove bleed through. The algorithm is described below.

The image of size MXN, is divided into mxn sized blocks, where 0<m< M and 0<n<N. The initial seed values of m and n can be set such that m x rows = M and n x columns = N, where rows and columns depict the number of rows and columns in the partitioned image. For each block, NPIM is calculated. Thus, a new 2-D array, with NPIM entries representing different blocks of the original image is formed. The transformation of the original image (12X18) to the characteristic array (3X6) is diagrammatically shown below. Each dot (.) in Fig. 1(a) denotes a pixel while in Fig. 1(b) each dot (●), represents NPIM value calculated from all the pixels in the respective block of the original image. Otsu's algorithm is applied on the entries of the 2-D array (Fig. 1(b)). Each block in the original image whose respective NPIM value is greater than the threshold value, is thresholded by Otsu's algorithm. In Fig. 1(c), the blocks having NPIM less than the threshold value have been crossed out. Except these blocks all the respective blocks of the original image (Fig. 1(d)) were operated upon individually by Otsu's algorithm. The size of all the blocks having NPIM values less than the threshold, are increased to m_1xn_1, where 0<m<m_1<M and 0<n<n_1<N and the original image is partitioned again. For all the blocks in Fig. 1(e), NPIM values are calculated again and a new 2-D array is obtained as shown in Fig. 1(f). Since, NPIM is a normalized statistical measure it provides a relatively better estimate of the entropies of blocks of unequal sizes w.r.t PIM. Again, Otsu's algorithm is applied on Fig. 1(f) and if the NPIM value of the blocks (m_1xn_1) is greater

than the threshold, they are binarized by Otsu's algorithm. The idea is to include a larger amount of information, by increasing the size of the block. The rest of the blocks of size mxn are not operated upon. The process is repeated until the entire image has been binarized.

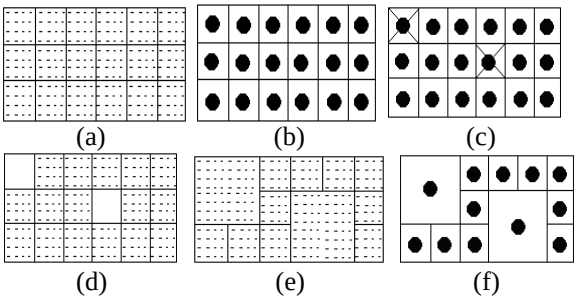

Figure 1: Proposed NPIM to be used with Otsu binarization based bleed through removal. (a) Original image, (b) NPIMs, (c) Thresholded NPIM, (d) Revisited image, (e) Adapted block size, (f) NPIM for new blocks.

So the algorithm can be summarized as:

1) Divide the grey leval image into mxn blocks
2) Find NPIM values of all the blocks
3) Apply Otsu's algorithm on the NPIM values
4) Binarize all the blocks, in the original image, having NPIM values greater than the threshold value, using Otsu's algorithm
5) For all the blocks with lower NPIM values, increase the block size to m1xn1
6) Find the NPIM values of the blocks of the re-partitioned original image
7) Apply Otsu's algorithm again on the original image and check if the threshold value for the bigger blocks (m1xn1) is higher than the new one
8) Binarize the bigger blocks (m1xn1) if they are above the threshold making use of Otsu's algorithm else go to step 5 and increase the size to m2xn2 , m1<m2 and n1<n2
9) Stop if all the image has been binarized

From the binary mask without bleed through artifacts, and the original image the grey scale image with mitigated bleed through is constructed by selective region filling.

3.3. Grey level output without bleed through

The grey scale output without the bleed through is generated by first generating a mask that shows the bleed through regions from the bleed through removed binary image as follows. The most frequent grey level of the original image is calculated, that indicates the document background color. Each of the foreground pixels in the bleed through removed binary image is filled with the most frequently occurring grey level. Each of the background pixels is replaced with the original image. Otsu's binarization on this image shows the bleed through regions. After that the bleed through regions are filled from the surrounding regions by Laplacian interpolation. The flowchart for mask extraction and bleed through region filling are elaborated in Figs. 3 and 4.

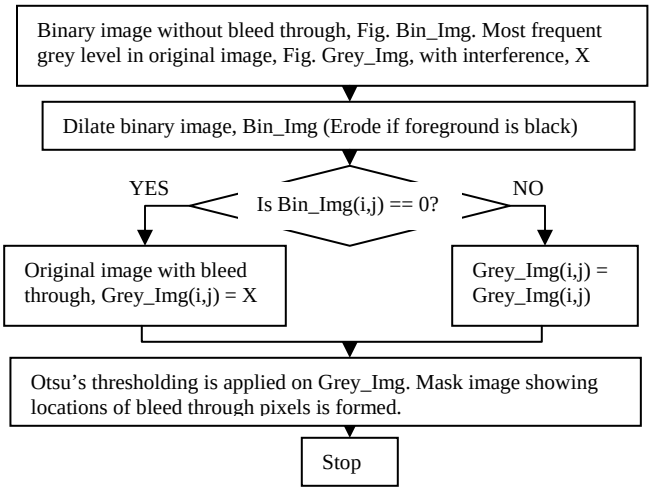

Figure 3: Mask generation to identify bleed through pixels

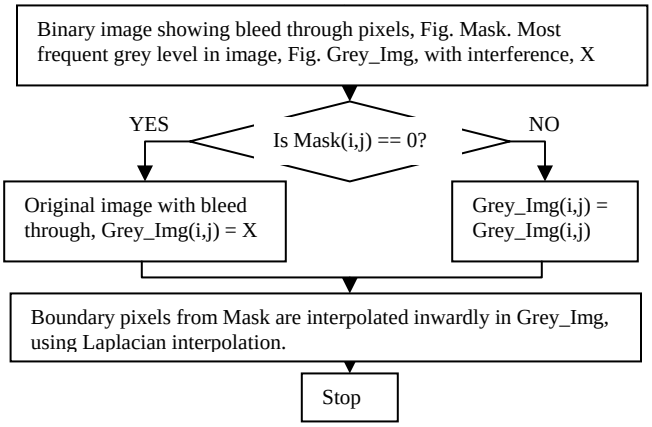

Figure 4: Bleed through removed grey scale image generated by selective region filling.

4. Experimental Results and Comparisons

The proposed algorithm was tested on a database of 111 images of varied content. 5 of them were low contrast images where state-of-art algorithms failed. The proposed method successfully removed bleed through on all the five images. The results were comparable on the remaining 106 test images. Fig. 21 shows efficient bleed through removal with the proposed algorithm and compares it with the state-of-art Mello-Lins algorithm [2] and Otsu's

binarization with global threshold [5].

4.1 Comparison with state-of-art

As explained, the proposed algorithm performs better with the images that have been written sparsely but have back-to-front interference throughout. Image Fig. 5 has very little useful information on the foreground and a very dark background with strong back-to-front interference. The proposed algorithm screens out the blocks having low value of NPIM (or entropy) and binarizes only those blocks which have high NPIM values above a particular threshold. The size of such blocks is increased till their entropies exceed the threshold, and then they are binarized with local threshold values.

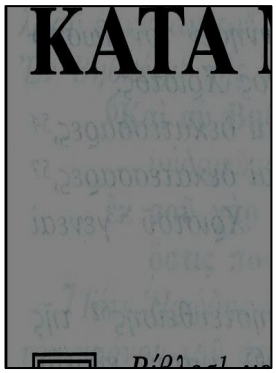

Figure 5: Original Image

Applying the proposed algorithm to Fig. 5, by partitioning the image with the block size of 128X128 gave the following result, Fig. 6 (after first pass). It can be seen that the portions of the image which were having low entropy (say constant grey level background), did not pass the threshold criteria of the NPIM and as such were not processed by Otsu's algorithm. However, in the second pass all such blocks increased in size. The final result is shown in Fig 7.

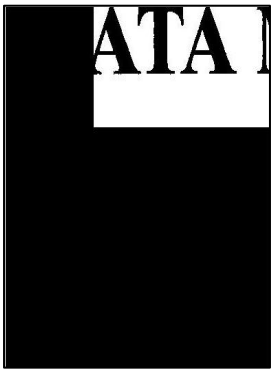

Figure 6: First Pass Image

Figure 7: Final result from proposed algorithm

Figure 8: Result from Mello and Lins [2]

The original image was also processed by the algorithm proposed by Mello and Lins [2]. The result obtained by processing the dark background image is presented in Fig 8. However, fading the background of the original image produced comparable results.

The faded original image, Fig. 9, along with the resulting images from the two algorithms, Fig. 10 and Fig. 11 have been shown respectively.

Figure 9: Faded image

For obtaining the results in grey scale, Fig. 10 was further processed. The mask (Fig. 13), showing the bleed through pixels, was obtained by dilating Fig. 10 and assigning the corresponding pixels of Fig. 5, with the most frequent grey level in the original grey scale image thus forming Fig. 12, after which Otsu's algorithm was applied.

Figure 10: Proposed Algo.

Figure 11: Mello & Lins

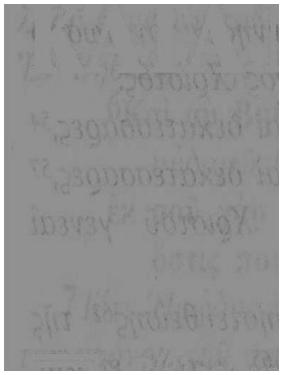
Figure 12: With foreground text suppressed

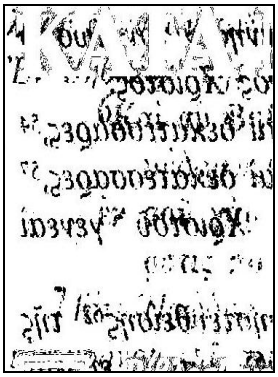
Figure 13: Mask showing bleed through pixels

The pixels locations in Fig. 5, corresponding to the mask were assigned the most frequent grey level value occurring in the original grey scale image and boundary interpolation with Laplace equation is applied at the boundaries of the bleed through features. The final grey scale image without bleed through is shown below (Fig. 14).

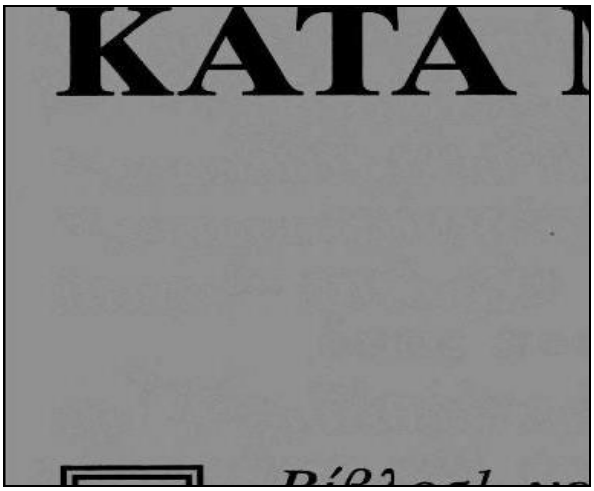

Figure 14: Final grey scale result

The results obtained by applying the algorithm on another sample image (Fig. 15) have been shown. The sample image was partitioned with the blocks of size 64X64. The final binarized image (Fig. 17) has been compared with results from Otsu's algorithm [5] (Fig. 18). The results show that the adapted local threshold selection on different sized blocks of the original image removes bleed through more efficiently compared to the image generated by selecting a global threshold using Otsu's algorithm. The result obtained by the application of algorithm proposed by Mello and Lins [2] has been shown in Fig. 19. The proposed method outperforms it as well. The final result in grey scale has been shown in Fig. 20.

4.2 Quantitative analysis

For a quantitative analysis, controlled bleed through was added by superimposing the faded background image on the foreground image on one of the low contrast images [11]. The fade factor was varied from 0 to 255, where 0 indicates strongest back-to-front interference and 255 indicates the weakest. For the low contrast image, bleed through was successfully removed for fades greater than 80, where state-of-the-art failed for fades up to 180.

A similar analysis on high contrast images showed that the results of the proposed algorithm were comparable to state-of-the-art. For one of the high contrast images at fade equaling 80, the percentages of text, paper and interference errors [11] are compared to the state-of-the-art in Table 1. The text-error measures how much text is erased, the paper error measures how much dirt in the paper is left over, and the interference error measures how much interference is left over. Although the proposed algorithm retains some paper dirt, it is better in retaining text and removing bleed through pixels compared to the state-of–the-art algorithms.

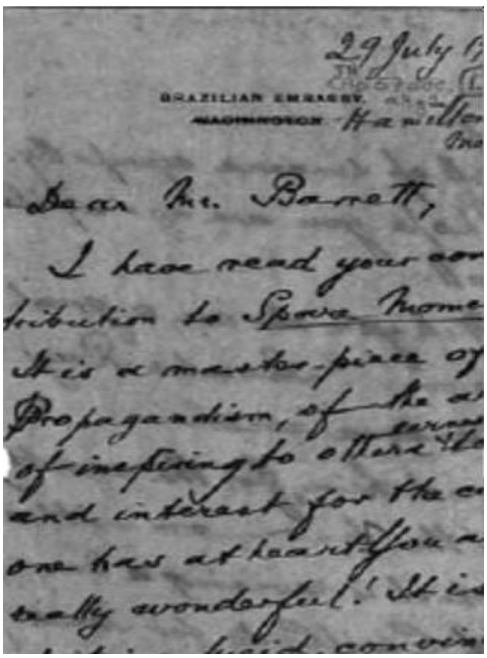

Figure 15: Orig. sample image

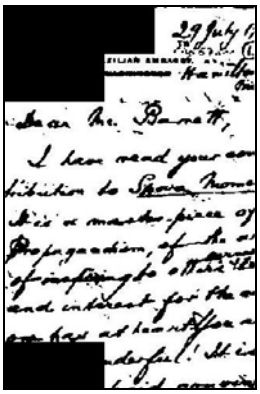

Figure 16: After First Pass

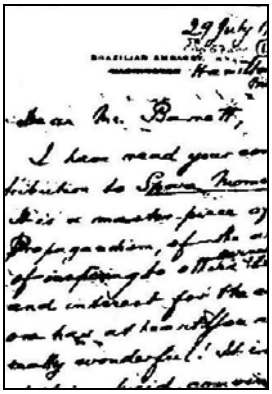

Figure 17: Proposed Algo.

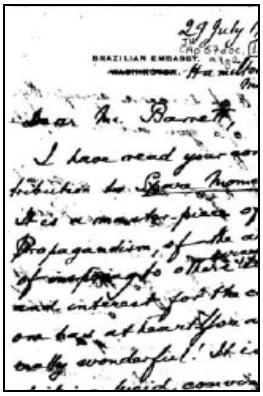

Figure 18: Otsu's Algo.

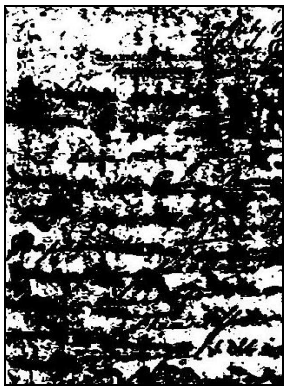

Figure 19: Mello & Lins

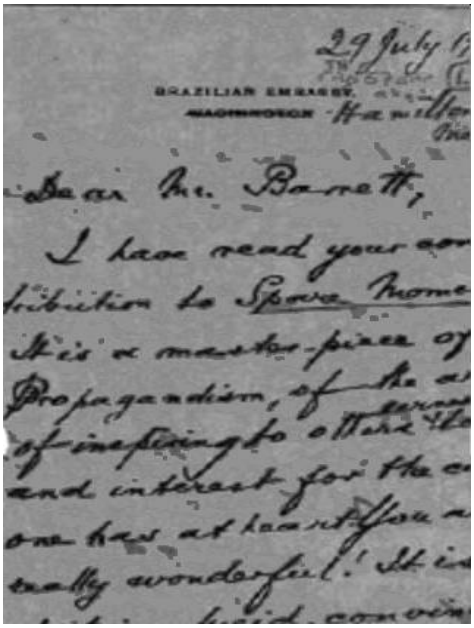

Figure 20: Final grey scale image

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	**2.89**	**1.98**	**1.07**

Table 1: Value of three quality factors at fade=80

The memory requirements for our algorithm is O(MXN) where MXN is the document image size. The computational complexity of the algorithm is O((MXN)/(mxn)X L2) where, mxn is the partitioned blocks' size, and L is the total number of grey levels (0, …, 256).

5. Conclusions

We have proposed a new bleed through removal algorithm based on Otsu's algorithm and NPIM, to calculate optimum localized thresholds of different sized blocks in the image for binarization of documents degraded due to back-to-front interference. The algorithm deals with problems that have been stated in the present state-of-art algorithms [3]. It has been tested on 111 images and the results were of better visual quality based on visual inspection and quantitative analysis in comparison to the other algorithms, when the images had a very dark background with strong back-to-front interference or had sparse content. The proposed algorithm had comparable performance to state-of-the-art for the other images in the database.

Figure 21: Comparison of the proposed, Mello-Lins [2] and Otsu's [5] algorithms for bleed through removal.

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.