Enhancing the Filtering-out of the Back-to-Front Interference in Color Documents with a Neural Classifier

G.F. P e Silva(1), R. D. Lins(1), J.M. Silva(1), S. Banergee(2), A. Kuchibhotla(2) and M. Thielo(3)

(1) UFPE-Brazil, (2) HP Labs.-India, (3) HP Brazil R&D-Brazil
{ gabriel.psilva, rdl}@ufpe.br, {serene.banerjee, anji, marcelo.resende.thielo}@hp.com

Abstract

Back-to-front, show-through, or bleeding are the names given to the interference that appears whenever one writes or prints on both sides of translucent paper. Such interference degrades image binarization and document transcription via OCR. The technical literature presents several algorithms to remove the back-to-front noise, but no algorithm is good enough in all cases. This article presents a new technique to remove such noise in color documents which makes use of neural classifiers to evaluate the degree of intensity of the interference and besides that to indicate the existence of blur. Such classifier allows tuning the parameters of an algorithm for back-to-front interference and document enhancement.

1. Introduction

At the beginning of the 1990s, the important file of over 6,000 private letters of Joaquim Nabuco, a statesman, writer and diplomat, one of the leading figures of the freedom of black slaves in Brazil and the first Brazilian ambassador to the US, were digitized in a joint preservation effort between the Joaquim Nabuco Foundation [2] and the Universidade Federal de Pernambuco. About 10% of the images presented a noise which had not been previously described in the technical literature, which was called “back-to-front” interference [1]. Much later, other people called it “bleeding” [14] or “show-through” [15].

The back-to-front interference appears whenever the content of the verso side of a document is visible on the front side due to paper translucidity (Figure 1). Such artifact degrades the automatic document transcription via OCR and there is often the superposition of both sides whenever the image is binarized, yielding an unreadable document. In the case of historical documents paper aging is a complicating factor as it darkens the paper and causes an overlapping of the distributions of the RGB components of the ink on both sides of the paper.

Figure 1. Zoom in a document from Nabuco bequest with back-to-front interference filtered out using the algorithm described in reference [1] and the new strategy herein.

The technical literature presents several algorithms to remove the back-to-front noise, but no algorithm is good enough in all cases [12]. Depending on the degree of translucidity of the paper, the kind of the ink used in printing or writing, the porosity of the paper, etc. the interference may show itself stronger or weaker. Some algorithms perform better than others in different degrees of interference and even one chosen algorithm may perform better if its parameters are tuned to the intensity of the noise.

This article presents a new strategy to select and tune an algorithm to remove the back-to-front interference in color documents. It makes use of a set of neural classifiers to assess the intensity of the back-to-front interference and to automatically adjust the parameters of the algorithm described in reference [1] to filter the noise out of a given a document. The blanks yielded by removing the artifact are filled in with pixels that correspond to the paper area in the document in such a way to provide the reader with “a natural” look of the document as if it were written on one side only. Figure 1 presents a sample of the results obtained by the algorithm proposed herein, in which one may observe its efficacy.

This paper is organized as follows. Section 2 describes the new filtering strategy. The document...
features extracted are presented in Section 3. Section 4 details the noise detection mechanism. The results obtained are presented and analyzed in Section 5. The paper ends presenting its conclusions and draws lines for further work.

2. The Filtering System

This section presents a new strategy to remove the back-to-front interference in color documents. First, one needs to remove the framing borders in the document image. Such borders act as noise that interferes with the analysis performed by other algorithms [3]. PhotoDoc [4] was used to pre-process the whole file of documents. Then, there is the use of the classifiers to verify the existence and degree of back-to-front interference. In parallel to that analysis another classifier checks the presence of blur in blocks of the image. Once the former classifier detects the presence and intensity of back-to-front noise, the global threshold algorithm presented in reference [1] is tuned to remove the artifact. At the end the interfering pixels are painted with the colors of pixels that correspond to the sheet of paper, removing the interference in the resulting image.

2.1. Classification strategy

The architecture of the classifier is presented in Figure 2.

![Classifier architecture](image)

The classifier used is Random Forest [6], implemented in Weka [5], an open source tool developed at Waikato University, New Zeland, which offers a wide variety of classifiers implemented. A set of features is extracted from each image to allow its classification. The details about the training and test sets are provided in Section 3.

The classifier developed works in parallel for the detection in image blocks of two different kinds of noise: the back-to-front interference and blur. In the case of back-to-front interference the classification is performed by three cascaded classifiers that split the bleeding noise into three categories: strong, medium and weak.

2.2. Discriminating the interfering pixels

The entropy-based segmentation algorithm by Silva-Lins-Rocha [1] is used twice to find the back-to-front interference area. The first time, to split the text from the rest of the document. The second time to separate the interference from the paper. The algorithm uses the grayscale converted image as an intermediate to split the histogram into three different areas of interest (see Figure 3).

The loss factor \( \alpha \) is a parameter of the algorithm that allows a better statistical tuning between the distributions of the original and binary histograms and it is based on Shannon entropy [8]. For the second image filtering using the algorithm by Silva-Lins-Rocha, a new \( \alpha \) is defined taking into account the intensity of the back-to-front interference and the presence of blur in the image, Table 1 indicates the suggested values \( \alpha \), such as to allow a better separation between the interference and the paper distribution. The values for \( \alpha \) were experimentally found.

<table>
<thead>
<tr>
<th>Interference</th>
<th>Blur</th>
<th>No Blur</th>
</tr>
</thead>
<tbody>
<tr>
<td>Weak</td>
<td>0.90</td>
<td>1.00</td>
</tr>
<tr>
<td>Medium</td>
<td>0.78</td>
<td>0.90</td>
</tr>
<tr>
<td>Strong</td>
<td>0.60</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 1. Values for alpha.

In summary, to detect the interference area:
1. Apply the segmentation algorithm by Silva-Lins-Rocha to sieve the foreground ink from the rest of the document (see Figures 4a/4b);
2. For each image block classified as having back-to-front interference a new loss factor \( \alpha \) (see Table 1) is chosen. Filter it using the algorithm by Silva-Lins-Rocha to separate the interference ink from the paper (see Figures 4c
and 4d), yielding a blank sheet of paper with white holes where there was ink and ink interference in the original document image.

To illustrate the process, in Figure 4 the first threshold, $T_L$, is obtained by the first application of the Silva-Lins-Rocha algorithm and the blocks threshold, $TH$, by the second. The pixels for which their gray-levels are less than $T_L$ are classified as ink of the front face. The pixels with gray-level greater than $T_H$ are classified as belonging to the paper. Pixels with gray-levels between $T_L$ and $T_H$ are discriminated as interference. The difference here is the application on each image block of a fine tuning between the thresholds $T_L$ and $T_H$ taking into account local image information. It is also worth stressing that this new filtering procedure has the advantage of reducing the risk of “damaging” image areas in which there is no interference, once the segmentation algorithm is not applied on them.

![Figure 04](image)

Figure 04. Views of a document with back-to-front interference: (a) ink of the front face and (b) paper with interference. Image segments of Figure 3b: (c) interference and (d) paper.

### 2.3. Document Reconstruction

The process proposed here makes use of a “linear” interpolation to fill in the blank pixels that originally corresponded to the interference area. Two binary masks are defined: TEXT and INTERF. The first one identifies the pixels from the ink of the front text (see Figure 5a); the second one highlights the interference area (see Figure 5b). One could assume that only the INTERF mask would be sufficient to the fulfillment of the filtering procedure, because the pixels to be replaced “are known already”. Some difficulties appear, however. The key idea is to replace the colors of the noisy pixels with colors as close as possible to the paper in their neighborhood. This is achieved by interpolation, using the colors of the pixels that surround the area to be filled in. There is still the need to remove some of the vestigial shades surrounding the ink pixels in the resulting image; otherwise those pixels will “damage” the interpolation process, bringing in noisy dark colors to the interference area. To solve this problem, one should apply the dilate morphological expansion operation to both masks, with that, the text and interference contours will be properly classified as “text” and “interference”, respectively (see Figure 6a and Figure 6b). As mentioned earlier on, the pixels that are used in the interpolation process are surrounding the interference area and with the pixels belonging only to the paper. This mask, PAPER, is obtained by the complement of the logical OR operation between the TEXT and INTERF dilated masks (see Figure 6c).

Equation 1 calculates a weighed mean, where the intensity of the nearest pixel from the pixel $P$ has the greatest weight. This is reasonable, because in a neighborhood, generally, the closer a pixel is from another, the more alike they should look. Figure 7b shows the result of the application of the proposed filtering strategy applied to the image in Figure 7a.

![Figure 05](image)

Figure 05. Masks that identify (a) the text and (b) the interference.

![Figure 06](image)

Figure 06. Dilated masks: (a) text (T) and (b) interference (I) (c) T or I.

Now, the interpolation process is presented. Let the coordinates be as depicted in Figure 7b:

- $(x_0, y_0)$ of a pixel $P$ from the interval to be interpolated;
- $(x_0, y_1)$ of pixel $P_N$ – first pixel north $P$;
- $(x_0, y_2)$ of pixel $P_S$ – first pixel south $P$;
- $(x_0, y_0)$ of pixel $P_W$ – first pixel west $P$;
- $(x_0, y_0)$ of pixel $P_E$ – first pixel east $P$.

Where $i_C(x, y)$ is the value of the component $C$ (R, G or B) of the pixel $(x, y)$. The intensity of the interpolated pixel $(P)$ is given by

$$i_C(x_0, y_0) = \frac{d_1 \cdot i_1 + d_2 \cdot i_2 + d_3 \cdot i_3 + d_4 \cdot i_4}{d_1 + d_2 + d_3 + d_4}, \tag{1}$$

where the $i_k$ and $d_k \ (k = 1 \ldots , 4)$ represent the intensities and the distances from the pixels – $P_N$, $P_S$, $P_W$ and $P_E$ – to $P$, sorted by increasing distances. For example, the closest pixel to $P$ has distance $d_1$ and intensity $i_1$, the second closest one has distance $d_2$ and
intensity $i_2$, and so on. The distance between any two pixels A e B with coordinates $(x_a,y_a)$ and $(x_b,y_b)$, is the standard Euclidian distance:

$$d_{AB} = \sqrt{(x_b - x_a)^2 + (y_b - y_a)^2}. \quad (2)$$

Figure 7. Images: (a) original, (b) Interpolation process and (c) filtered by the new strategy proposed here.

3. Classification Features

The choice of the features to be extracted from each image is of paramount importance to the success of the classifier. The following set of features, based a combination on the classifiers described in references [9]. Image binarization is performed by using Otsu [10] algorithm. The height and width stand for the number of pixels in the image. RGB size stands for the true color size of the image (if it is a color image). 8-bits size is either the size of the original image if in grey scale or the size of the grey-scale converted from true-color. #Black_pixels stands for the number of black pixels in the monochromatic converted image. The combination of the features presented in [9] and the two new features (Local Power Spectrum Slope and Maximum Saturation) in [11] yielded a relative gain to the performance of the classifier. Each of these features was taken on nine blocks of the image (see Figure 5).

Figure 8. Areas of interest for extract features

4. Back-To-Front Noise Detection

Back-to-front noise, depending on its strength, may make document binarization unviable. As most OCRs take a binary image as input, thus documents with such noise may not be automatically transcribed. Researchers [12][13] have pointed out that no algorithm in the literature is good enough to remove bleeding noise in all sorts of documents. Depending on the strength of the noise, some algorithms may perform better than others. Unfortunately, the back-to-front noise appears more often in the digitalization of documents than one may assume to start with. The test set of documents we used with show-thought had 260 real-world documents (no synthetic ones) which were obtained either from historical files (as shown in Figure 1). The images were divided into nine blocks totaling 2,340 blocks and were hand labeled according to four levels of interference as: strong (773), medium (856), light (524) and none (187). The classifiers for this noise were cascaded, as shown in Figure 2. The strong-classifier was trained with the blocks tagged as strong in the training set, against all the remaining images (Medium-Light-None) from the training set (strong 150; medium 150; light 100; and 20 none). Similarly, the medium-classifier was trained with the blocks labeled as medium, against the others with a lighter or no interference. The classification results obtained are shown in Table 2.

<table>
<thead>
<tr>
<th>Back-to-front</th>
<th>Strong</th>
<th>Medium</th>
<th>Light</th>
<th>None</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>Strong</td>
<td>703</td>
<td>58</td>
<td>11</td>
<td>1</td>
<td>90.94</td>
</tr>
<tr>
<td>Medium</td>
<td>27</td>
<td>816</td>
<td>4</td>
<td>9</td>
<td>95.32</td>
</tr>
<tr>
<td>Light</td>
<td>5</td>
<td>9</td>
<td>96</td>
<td>12</td>
<td>92.93</td>
</tr>
<tr>
<td>None</td>
<td>1</td>
<td>2</td>
<td>11</td>
<td>187</td>
<td>92.51</td>
</tr>
</tbody>
</table>

Table 2. Confusion matrix of the back-to-front noise classifier with sub-sampled block images

5. Blur Detection

The presence of blur may be an indicator of low quality digitalization, but can also be associated with other problems such as the scanning of hard-bound volumes. The blur noise is seldom global. In general, it affects some areas of a document. In the case of the documents studied here, the blur noise is originated from the spreading out of the ink in the verso face of the document. While the issues related to the analysis of image-blur have attracted much attention from researchers in recent years, the work reported in the literature focus mainly on solving the problem of deblurring. Blur detection is a complex task. The work reported in [7], points at blur as one of the greatest difficulties for the filtering out of the back-to-front noise in historical documents. Bluer detection is solved here by using the classifier presented in subsection 2.1 and the characteristics described in section 3. The image blocks were classified manually, 124 blocks with and 2216 without blur. The training set used 20 blurred blocks and 150 unblurred ones. Besides those, 500 blocks with synthetic blur were used to validate the
classifer. The result is shown by the myo classifer confusion matrix (see Table 3).

<table>
<thead>
<tr>
<th>Orientation</th>
<th>With</th>
<th>Without</th>
<th>Accuracy %</th>
</tr>
</thead>
<tbody>
<tr>
<td>With</td>
<td>615</td>
<td>9</td>
<td>98.55</td>
</tr>
<tr>
<td>Without</td>
<td>3</td>
<td>2,213</td>
<td>99.86</td>
</tr>
</tbody>
</table>

Table 3. Confusion matrix of the blur noise classifer with sub-sampled images in proposed architecture

6. Results and Analysis

The proposed algorithm was tested in a set of 260 images from the Joaquim Nabuco bequest of digitalized documents [2], yielding good results. Evidences of the efficiency of the new filtering technique are shown in Figure 9, as the back-to-front interference was removed yielding a more readable document with a “natural” look. Figure 9 provides the results of using different strategies, amongst them using as fulfillment for the blanks the result of the interpolation based on Laplace’s equation (the MATLAB function “roifill” was used). The third alternative is one of the strategies proposed by Castro and Pinto [16] that uses the algorithm by Salvola and Pietikainen [17] which defines a mask that identifies the pixels of the foreground and background objects. The final image is obtained through keeping the object pixels and replacing the background pixels with the average of the colors of the pixels in that class. The latter strategy yielded the best results. The two strategies proposed herein yielded very similar quality results. However, the one based on Laplace interpolation leaves the filled-in area look undesirably uniform with a “flat” color. On the other hand, the linear interpolation yields a residual pattern of vertical/horizontal stripes. The strategy proposed by Castro and Pinto [16] aims to yield a uniform paper surface with unchanged text, while the ones presented here try to remove only the interference, keeping the pixels from the paper and text unchanged. However, in the very few images in the Nabuco file that the back-to-front interference looks very "blurred" (see last segments of Nabuco file in Figure 9), the proposed algorithm did not perform well. The detection of the whole back-to-front interference area is far from being a trivial task.

References

Figure 9 - Parts of documents from the Nabuco file: original and filtered.