Assessing Strategies to Remove Back-to-Front Interference in Color Documents

Rafael Dueire Lins
DES – CTG – UFPE
Recife, 50.670-000, PE
BRAZIL
+55 81 2126-8210 ext. 274
rdl@ufpe.br

Gabriel de França P. e Silva
DES – CTG – UFPE
Recife, 50.670-000, PE
BRAZIL
+55 81 2126-8213
gfps@cin.ufpe.br

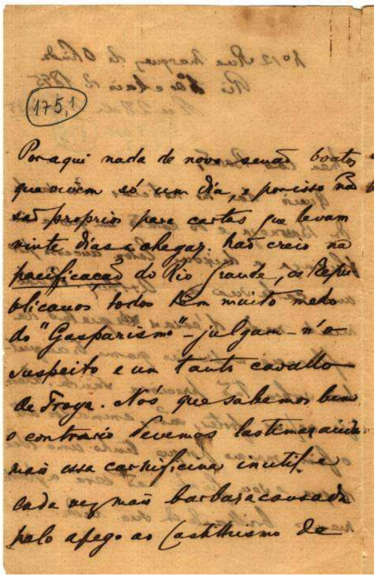
João Marcelo Monte da Silva
DES – CTG – UFPE
Recife, 50.670-000, PE
BRAZIL
+55 81 2126-8213
joao.mmsilva@ufpe.com

***Abstract*—Whenever a document is written on both sides of translucent paper there is a back-to-front interference, also known as bleeding or show-through. In the literature there are many algorithms to filter out the back-to-front interference in documents. This paper presents a new quantitative method to assess those algorithms. This method is based on the synthesis of images with interference.**

Keywords-* *document*; *back-to-front interference; bleeding; show-through; quantitative method

I. INTRODUCTION

If a document is written or printed on both sides of a translucent paper, the back printing will be visualized on the front face. This phenomenon, first addressed in the literature by Lins in 1994 [3], was called *back-to-front interference*. The Joaquin Nabuco`s Letter shown in Figure 1, belongs to the bequest of the Fundação Joaquim Nabuco – FUNDAJ [2], examples such interference.



175,1

Poraqui nada de novo senão boatos
que ocorrem só um dia, e por isso não
são proprios para cartas que levam
vinte dias a chegar. Não creio na
pacificação do Rio Grande; os Repu-
blicanos todos tem muito medo
do "Gasparismo" – julgam-n'o
suspeito e um tanto cavallo
de Troya. Nós que sabemos bem
o contrario devemos lastimar ainda
mais essa carnificina inutil e
cada vez mais barbara causada
pelo apego ao Castilhismo de

Figure 1. Image of historical document with *back-to-front interference*

In the literature there are many algorithms that were developed to filter the back-to-front interference [1], [5], [6], [7], [8], [9]. This paper describes a quantitative method to assess such algorithms when applied in color documents with back-to-front interference.

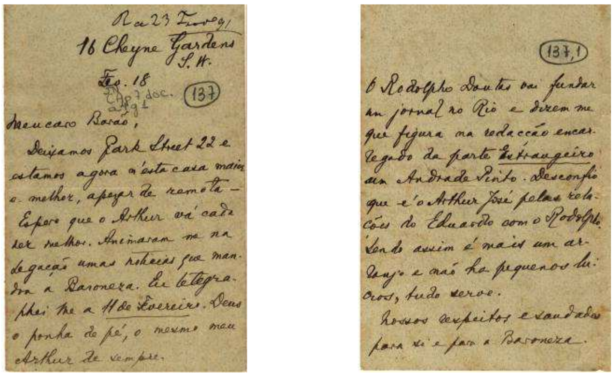
The process developed here is a "generalization" of the presented in [4] that works with gray-scale images.

II. DESCRIBING THE QUANTITATIVE ASSESSMENT METHOD

The method described here is divided in two steps:

1) *synthesis of images with back-to-front interference* – based on two images without interference; and
2) *comparison of the images filtered by an algorithm with a reference image* – calculating the PSNR (Peak Signal-to-Noise Ratio) for each color component in the RGB space.

The PSNR was chosen because it brings a low computational complexity compared with parameters that define the quality image by perceptual sense [10].



Ra 23 Fevº 91
16 Cheyne Gardens
S.W.
Fev. 18
137

Meu caro Barão,
Deixamos Park Street 22 e
estamos agora n'esta casa maior
e melhor, apezar de remota –
Espero que o Arthur vá cada
vez melhor. Animaram-nos na
legação umas noticias que man-
dou a Baroneza. Ele telegra-
phou-me a 11 de Fevereiro. Deus
o ponha de pé, o mesmo meu
Arthur de sempre.

137,1

O Rodolpho Dantas vai fundar
um jornal no Rio e dizem-me
que figura na redacção encar-
regado da parte Estrangeira
um Andrade Pinto. Desconfio
que é o Arthur José pelas rela-
ções do Eduardo com o Rodolpho.
Sendo assim é mais um ar-
ranjo e não ha pequenos lu-
cros, tudo serve.
Nossos respeitos e saudades
para si e para a Baroneza.

Front Image – F

Back Image – V

Figure 2. Original images of two documents without back-to-front interference.

I. *Synthesis of Images with Back-to-Front Interference*

This step starts taking two images of color documents without back-to-front interference, as one can see in Figure 2: F – role plays the document front; and V – role plays the document back.

By a composition of the images F and V, using the α channel, a third image FV_α is generated (see Figure 3).

The color components of this new image are given by:

$$FV_\alpha^{(C)} = \alpha F^{(C)} + (1-\alpha)V^{(C)} \quad (1)$$

where the index (*C*) represents the red (R), green (G) and blue (B) color components, and α is the opacity coefficient of the F image that assumes values from 0 to 100%. Furthermore, the factor $(1-\alpha)$ indicates the transparency coefficient of the F image.

To illustrate this operation, one can imagine F as a color glass. The color of each pixel is the color of the glass, that has opacity α, plus the color that is behind, reduced by the factor $(1-\alpha)$ that represents the glass transparency.

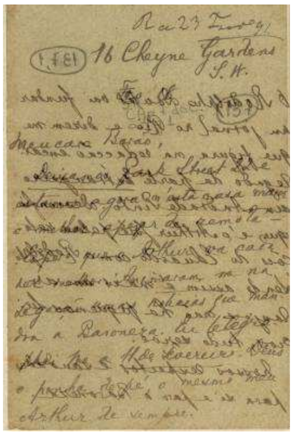

FV_α Image

Figure 3. Composition, using α channel, of the F and V.

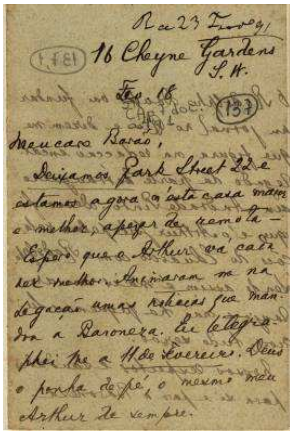

S_α Image

Figure 4. Final result of the synthesis process.

Finally, one performs a "darker operation" between the F and FV_α images, generating the final image S_α (see Figure 4). This operation is performd pixel by pixel, comparing the luminance of the correspondent pixels on both images. Following, such operation is described.

- Let $f^{(Y)}(i,j)$ and $fv_\alpha^{(Y)}(i,j)$ be the luminance values of the pixels from F and FV_α images, respectively, in position (i,j) given by

$$\begin{cases} f^{(Y)}(i,j) = 0.299\, f^{(R)}(i,j) + 0.587\, f^{(G)}(i,j) + 0.114\, f^{(B)}(i,j) \\ fv_\alpha^{(Y)}(i,j) = 0.299\, fv_\alpha^{(R)}(i,j) + 0.587\, fv_\alpha^{(G)}(i,j) + 0.114\, fv_\alpha^{(B)}(i,j) \end{cases} \quad (2)$$

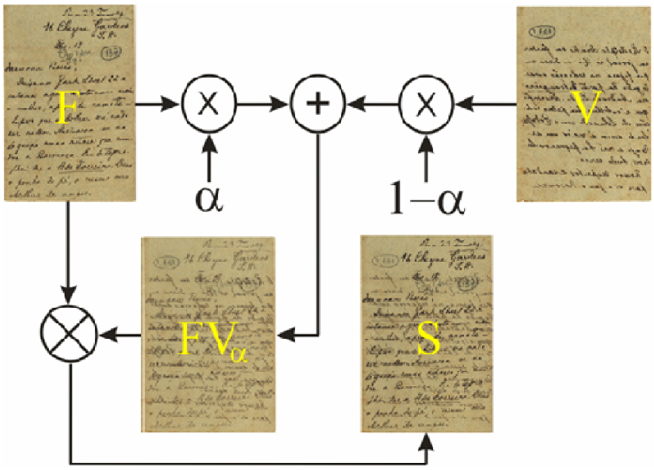
  where $(f^{(R)}(i,j), f^{(G)}(i,j), f^{(B)}(i,j))$ and $(fv_\alpha^{(R)}(i,j), fv_\alpha^{(G)}(i,j), fv_\alpha^{(B)}(i,j))$ are the color vector in the RGB space of the pixels from F and FV_α images, respectively, in position (i,j).

- If $f^{(Y)}(i,j) \le fv_\alpha^{(Y)}(i,j)$, i.e., if the pixel from F has luminance less than (darker) or equal to the luminance of the pixel from FV_α, then the pixel in the final image S_α will be the pixel from F. Else, it will be the pixel form FV_α. This is shown in Equation 3.

$$\begin{cases} \text{if } f^{(Y)}(i,j) \le fv_\alpha^{(Y)}(i,j) \text{ then} \rightarrow s_\alpha^{(C)}(i,j) = f^{(C)}(i,j) \\ \text{if } f^{(Y)}(i,j) > fv_\alpha^{(Y)}(i,j) \text{ then} \rightarrow s_\alpha^{(C)}(i,j) = fv_\alpha^{(C)}(i,j) \end{cases} \quad (3)$$

  where $s^{(C)}_\alpha(i,j)$ represents the red (R), green (G) and blue (B) components of the pixel from S_α image in position (i,j).

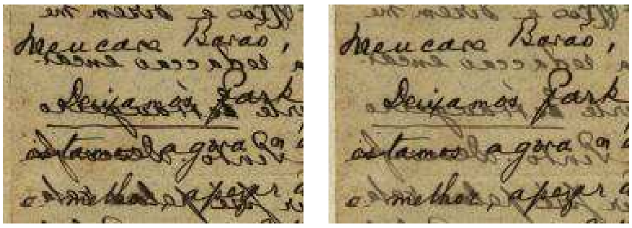
In Figure 5 is shown a block diagram of the synthesis process. The image indicated by V is the V image from Figure 2 mirror reflected.



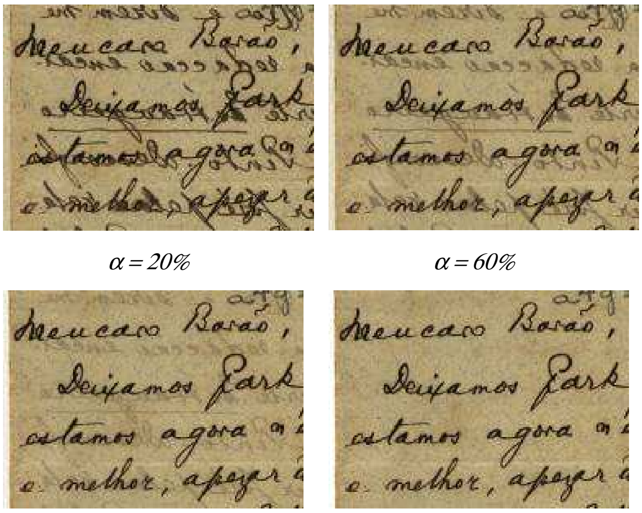
⊗ *– denote the darker operation*

Figure 5. Block diagram of the synthesis process.

To assess the filtering capability of the algorithms, the opacity factor α assumes values from 0 to 100% (α = 0, 1, ..., 100%).The effect of the α variation on the final synthesized document S_α generated from the documents presented in Figure 2 with V mirror-reflected is presented in Figure 6.


$\alpha = 0\%$

$\alpha = 40\%$



$\alpha = 20\%$ $\alpha = 60\%$

$\alpha = 80\%$ $\alpha = 100\%$

Figure 6. Synthesized images S_α with different values of α.

II. *Evaluating the PSNRs*

After the synthesis process, the synthesized images are filtered by the algorithms to be assessed. The filtering results are compared with the F image (see Figure 2). The PSNR (Peak Signal-to-Noise Ratio) will be used in this step to measure the quality of the resultant images of the filtering process. As it is working with true-color images the PSNR is evaluated for each color component in the RGB space as follows

$$PSNR_k^{(C)}(\alpha) = 20\log_{10}\frac{255}{\sqrt{\frac{\sum_{i=1}^{H}\sum_{j=1}^{W}[f^{(C)}(i,j) - filt_{\alpha,k}^{(C)}(i,j)]^2}{H \times W}}}, \qquad (4)$$

where H is the image height, W is the image width, $f^{(C)}(i,j)$ is the intensity of the C component in position (i,j) of the F image (reference) and $filt_{\alpha,k}^{(C)}(i,j)$ is the intensity of the C component in position (i,j) of the image $FILT_{\alpha,k}$ that is the filtering result of the S_α image by algorithm k.

In Figure 7, one shows a block diagram that illustrates the comparison process between the reference image F and the filtered image $FILT_{\alpha,k}$.

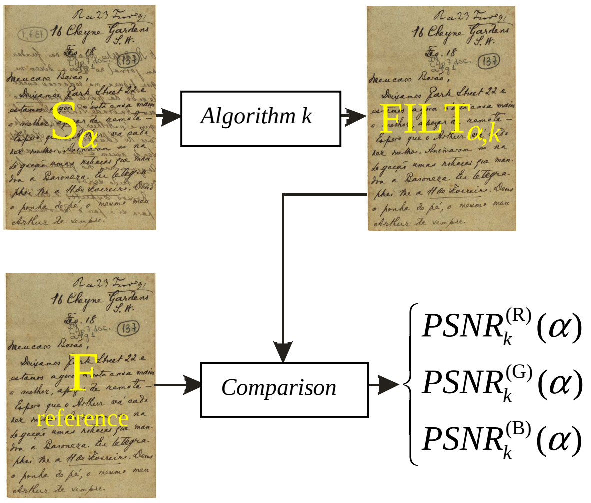


Figure 7. Block diagram of the comparison process between the reference image F and the filtered image $FILT_{\alpha,k}$.

III. RESULTS AND ANALYSIS

It was selected the algorithm proposed in [8] to demonstrate the proposed assessment method.

As mentioned in the method description it was synthesized 101 images with different back-to-front interference intensity (α = 0, 1, ..., 100%). These images were filtered by the selected algorithm and the PSNRs for each color component of each image were evaluated. The results were plotted in a graph (see Figure 8) that brings in the horizontal axis the opacity value α used to synthesize the S_α image, and in the vertical axis the PSNR values of the R, G and B color components.

Observing the graph presented in Figure 8 one notes that the selected filtering algorithm increases its performance when the interference intensity decreases, i.e., when the opacity α grows. This behavior is kept while α is less than approximately 80%. For opacity values greater than 80% the algorithm performance has a slightly downfall, this occurs because, in documents with almost inexistent interference, such algorithm classifies as “interference” some pixels belonging to the paper.

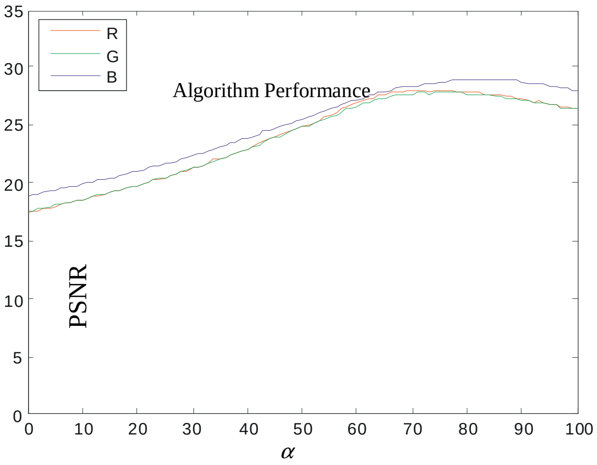


Figure 8. Graphs of the PSNR values of the filtered images FILT$_{\alpha,k}$ using the F image as reference.

From the set of synthesized images we selected three samples which represent existent interference in real documents from the Joaquim Nabuco`s bequest. The samples with α=40%, α=65% and α=90% (see Figure 9) represent images with strong, mid and weak interference, respectively. In the same figure, the results of the algorithm application are shown.

Observing the Filtered images in Figure 9 one can evidence that the left image ($\alpha = 40\%$) does not have a "so good" appearance and the others two ($\alpha = 65\%, 90\%$) have almost the same "good" appearance. If one takes the correspondent PSNR values in the graph from Figure 8, one will be evidenced that the assessment method proposed here is consistent with visual inspection, because the PSNR values are following the visual conclusion.

A study made using this quantitative assessment method aims to indicate which algorithm is more suitable to be applied in a specific α range. Thus, it is interesting to know "how to determine the α value in a specific document".

It is possible estimate the "opacity" of a real document:

- by the presented model, an interference pixel has luminous intensity *interf*(*i*,*j*) given by

$$interf(i,j) = \alpha \cdot paper(i,j) + (1-\alpha) \cdot ink_{\text{back}}(i,j) \qquad (5)$$

  were *paper*(*i*,*j*) and *interf*(*i*,*j*) are the pixel intensities of the paper and back ink, respectively;

- in most cases we can assume that the "back" ink intensity *ink*$_{\text{back}}$ has approximately the same value of the front ink *ink*$_{\text{front}}$, thus one collects interference, paper and front ink to represent the *interf, paper* and *ink*$_{\text{front}}$ intensities, respectively.

- finally, using the Equation 5 and assuming that *ink*$_{\text{back}}$≈*ink*$_{\text{front}}$ we can estimate the α value by

$$\alpha = \frac{interf - ink_{\text{front}}}{paper - ink_{\text{front}}} \qquad (6)$$

To exemplify, one takes the image form Figure 1. The collected values were

$$\begin{cases} ink_{\text{front}} = 23 \\ interf = 106 \\ paper = 201 \end{cases} \Rightarrow \alpha = \frac{106-23}{201-23} \approx 0.47 \qquad (7)$$

thus, we can say that such document has a opacity coefficient α≈47%. An automatic way to know the interference value could be obtained by the direct binarization of the color image. The experience shows that around no "grater" than 8%, of the total number of pixels in a document, are translated in black pixels. The direct binarization implies a greater amount of pixels, proportional to the opacity coefficient. The relationship between the α values and amount of black pixels is not established, and it is a future work.

IV. CONCLUSIONS

This work introduces a quantitative method to assess filtering algorithms to remove back-to-front interference from images of color documents. This method allows a study about such algorithms aiming to know what algorithm is more suitable to filter a specific document, taking to account its opacity α. To automate this choice it is necessary estimate sample values from the front ink, paper and interference. Another way is directly binarize the color image, and by a pre-established relationship determine the opacity α by the amount of black pixels of the binarized image. With the opacity value evaluated, one chooses the "best" algorithm to filter such document, taking into account a previous study.

The PSNR was used to measure the quality of the final image. Work on progress intends to define measures that inform the readability of the useful information, the preservation of the paper and the fulfilling quality of the interference area.

In some images the interference appears "blurred". To take into account this effect, one pass the image that will role play the document back by a low pass filter, before the synthesis process. The model could be more sophisticated, if one uses an adaptive filter that considers the data from the front image.

ACKNOWLEDGMENTS

The research reported herein was partly sponsored by CNPq – Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Brazilian Government. The authors also express their gratitude to the Fundação Joaquim Nabuco, for granting the permission to use the images from Nabuco bequest.

REFERENCES

[1] Cao, R., Tan, C. L. and Shen, P. *A wavelet approach to double-sided document image pair processing*. Proc. Int. Conf. Image Proc. Oct. 2001.

[2] FUNDAJ: www.fundaj.gov.br

[3] Lins, R. D., *et al*. *An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming*. pp. 111-121, North-Holland, 1995.

[4] Lins, R. D. and da Silva, J. M. M. *A Quantitative Method for Assessing Algorithms to Remove Back-to-Front Interference in Documents*. ACM SAC-DE, 2007, Seoul, Korea. New York, NY, USA : ACM Press, 2007. p. 610-616.

[5] Nishida, H. and Suzuki, T. *A Multiscale Approach to Restoring Scanned Color Document Images with Show-trough Effects*. Proc. of. ICDAR 2003, 2003.

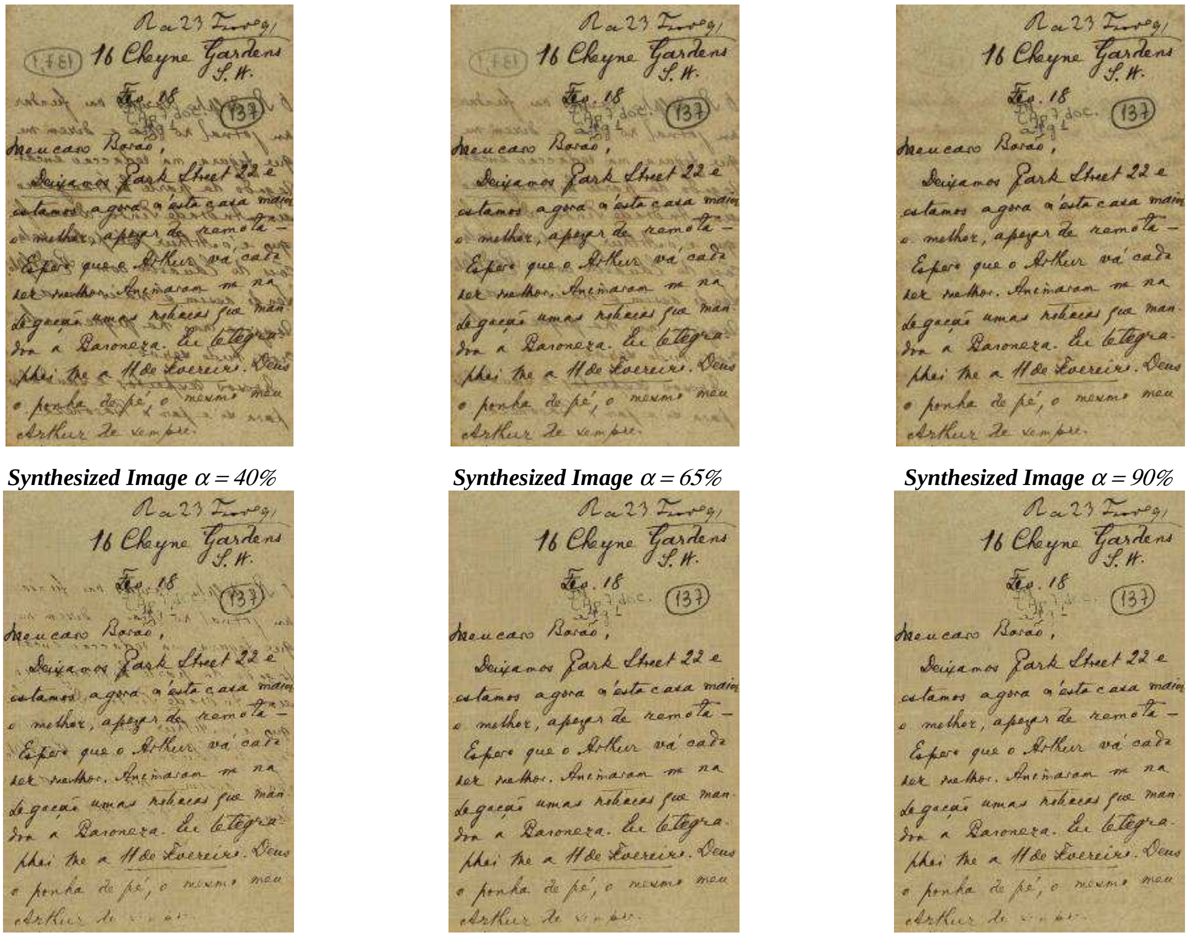
[6] Ophir, B. and Malah, D. *Improved cross-talk cancellation in scanned images by adaptive decorrelation*., 23rd IEEE Convention of Electrical and Electronics Engineers in Israel, 2004, pp. 388-391.

[7] Sharma G., *Show-trough cancellation in scans of duplex printed documents*. IEEE Trans. Image Processing, v10(5):736-754, 2001.

[8] Silva, João Marcelo Monte da, Lins, R. D., Silva, G. F. P., Enhancing the Quality of Color Documents with Back-to-Front Interference In: International Conference on Image Analysis and Recognition, 2009, Halifax. Proceedings of ICIAR 2009 - Lecture Notes in Computer Science. Heidelberg: Springer Verlag, 2009. v.5627. p.875 – 885.

[9] Su, F. and Mohammad-Djafari, A. *Bayesian Separation of Document Images with Hidden Markov Model*. 2nd International Conference on Computer Vision Theory and Applications, Barcelona, Spain, 2007.

[10] Zamplo, R. F. and Seara, R. *Estudo Comparativo entre Métricas para Avaliação da Qualidade de Imagens*. XXII SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES, pp. 237-241, Campinas-SP, Brazil, 2005.



R.a 23 Fev.o 91

16 Cheyne Gardens
S.W.

Fev. 18

137

Meu caro Barão,

Deixamos Park Street 22 e estamos agora n'esta casa maior e melhor, apezar de remota - Espero que o Arthur vá cada vez melhor. Animaram-me na Legação umas noticias que mandou a Baroneza. Eu telegraphei-lhe a 11 de Fevereiro. Deus o ponha de pé, o mesmo meu Arthur de sempre.

Synthesized Image $\alpha = 40\%$ *Synthesized Image* $\alpha = 65\%$ *Synthesized Image* $\alpha = 90\%$

Filtered Image $\alpha = 40\%$ *Filtered Image* $\alpha = 65\%$ *Filtered Image* $\alpha = 90\%$

Figure 9. Synthesized and filtered images.