Um Novo Método de Filtragem de Interferência Frente-Verso em Documentos Coloridos

João Marcelo Monte da Silva e Rafael Dueire Lins

***Resumo*— Um artefato conhecido como interferência frente-verso surge quando um documento é escrito em ambas as faces de papel translúcido. Tal interferência, além de apresentar dificuldades para a leitura do documento, impossibilita a sua transcrição automática e binarização. Este artigo apresenta um novo sistema de filtragem de interferência frente-verso em imagens de documentos coloridos, visando melhorar sua legibilidade. A idéia é discriminar a área interferente do documento e preenchê-la com pixels existentes no papel.**

***Palavras-Chave*— Interferência frente-verso, documentos históricos.**

***Abstract*— Back-to-front interference, bleeding or show-through is the name given to the artifact that arises whenever one writes on both sides of translucent paper. Such noise is not only a burden to the reading of documents, but makes non-viable their automatic transcription via OCRs and binarization. This paper presents a new filtering system to remove the back-to-front interference in color document images, aiming enhance theirs readabilities. The idea is find the document interference area and full it with existent pixels in the paper.**

***Keywords*— Back-to-front interference, bleeding, show-through, Historical Documents.**

I. INTRODUÇÃO

Com freqüência, seja em acervos históricos ou burocráticos, encontramos documentos escritos ou impressos em ambas as faces do papel. Caso o grau de translucidez do papel seja alto, pode-se ver de um lado a sombra do que foi escrito do outro. Tal sombra é tanto mais forte quanto mais transparente for o papel. O trabalho de digitalização do acervo de correspondências de Joaquim Nabuco, realizado entre a Fundação Joaquim Nabuco e a Universidade Federal de Pernambuco [1] mostrou que tais documentos apareciam freqüentemente. O envelhecimento do papel é um fator de dificuldade, pois o seu escurecimento diminui o "grau de separação" entre a tinta de cada um dos lados e o papel. A fluidez da tinta também representa um fator complicador, pois pode acarretar uma maior penetração na superfície aumentando a visibilidade da escrita de um lado no outro. O documento apresentado na Figura 1 exemplifica um documento com a interferência descrita, que foi relatada pela primeira vez na literatura técnica na referência [1] e chamada de interferência frente-verso (em inglês *back-to-front interference*, também chamada posteriormente por outros autores de *bleeding* e *show-through*).

João Marcelo Monte da Silva e Rafael Dueire Lins, Departamento de Eletrônica e Sistemas, Centro de Tecnologia e Geociências, Universidade Federal de Pernambuco, Recife, Brasil, E-mails: joaommsilva@gmail.com, rdl@ufpe.br.

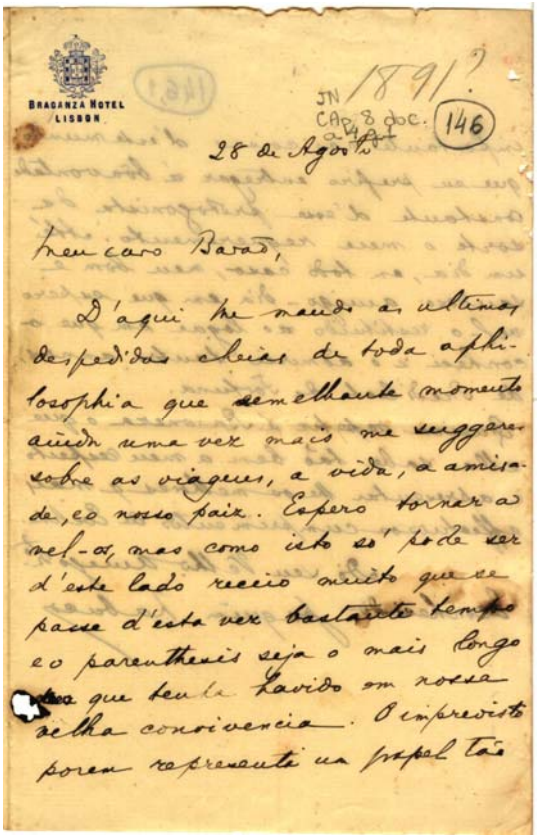


Fig. 1. Documento com interferência frente-verso.

A interferência frente-verso não só traz dificuldades para a leitura do documento, mas também impossibilita a sua transcrição automática e binarização. Este artigo apresenta um novo algoritmo de filtragem de interferência frente-verso em imagens de documentos coloridos, visando melhorar sua legibilidade. A idéia é discriminar a área interferente do documento e preenchê-la com pixels existentes no papel. Na seção II, faremos a descrição desse novo método de filtragem. Os resultados e as análises estão apresentados na seção III. Finalmente, na seção IV, apresentamos nossas conclusões e algumas linhas para trabalhos futuros.

II. SISTEMA DE FILTRAGEM POR LIMIAR

Nesta seção, apresentaremos um sistema para filtrar a interferência frente-verso em imagens de documentos coloridos de uma forma automática. Este novo algoritmo mostra uma das possíveis formas de automatizar o processo descrito em [2].

A idéia aqui é discriminar os *pixels* provenientes da interferência frente-verso e substituí-los por *pixels* existentes

no papel. Assim, dividiremos o problema em duas partes:

1. discriminação dos *pixels* provenientes da interferência; e
2. preenchimento da área interferente.

A. Discriminação dos Pixels da Interferência

Este passo do processo de filtragem utiliza o algoritmo de segmentação Silva-Lins-Rocha [3] fazendo, contudo, algumas otimizações. Este algoritmo é utilizado duas vezes: uma para separar o texto do resto do documento e outra para separa a interferência do papel. Dessa forma, é essencial uma breve descrição do funcionamento desse algoritmo e as otimizações feitas para esta aplicação.

1) Algoritmo Silva-Lins-Rocha

O algoritmo Silva-Lins-Rocha faz uso da entropia de Shannon [4] para fazer um ajuste estatístico entre a distribuição dos níveis de cinza da imagem original e a distribuição da imagem binária que identifica o objeto e o fundo.

O processo é realizado da seguinte forma:

1. Converte-se a imagem original (vide Figura 1) para sua versão em níveis de cinza (vide Figura 2a) através da seguinte equação:

$$gray(i,j) = 0{,}299r(i,j) + 0{,}587g(i,j) + 0{,}114b(i,j) \qquad (1)$$

   onde $r(i,j)$, $g(i,j)$ e $b(i,j)$ são, respectivamente, as componentes de vermelho, verde e azul do *pixel* na posição (i,j) da imagem no espaço RGB; e $gray(i,j)$ é o nível de cinza correspondente na imagem em níveis de cinza.

2. Levanta-se o histograma da imagem em níveis de cinza (vide Figura 2b) e calcula-se sua entropia através da seguinte forma:

$$H = -\sum_{i=0}^{255} p_i \log(p_i), \qquad (2)$$

   onde $\{p_0, p_1, \ldots, p_{255}\}$ é a *distribuição de probabilidade a priori* dada por:

$$p_i = \frac{\#\text{ de } pixels \text{ com nível de cinza i (0 a 255)}}{\#\text{ total de } pixels \text{ da imagem}}, \qquad (3)$$

   e o $\log(\cdot)$ é usado na base 2.

3. Varrem-se os níveis t, calculando para cada um a *distribuição de probabilidade a posteriori* $\left\{P_t = \sum_{i=0}^{t} p_i,\ 1-P_t\right\}$, enquanto $P_t \leq 0{,}5$, e a entropia associada a essa distribuição:

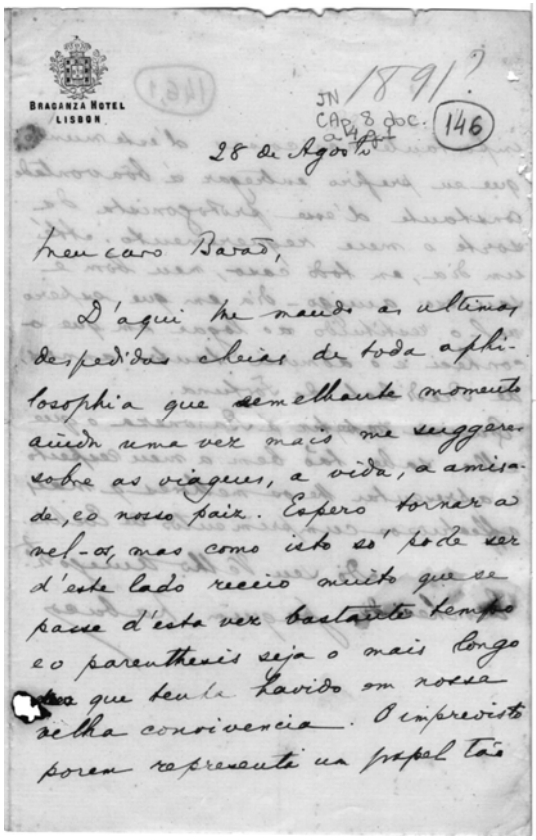
$$H'(t) = -P_t \log(P_t) - (1-P_t)\log(1-P_t). \qquad (4)$$

4. Finalmente, determina-se o limiar ótimo $t=T$ que minimiza a função $|e(t)|$ dada por:

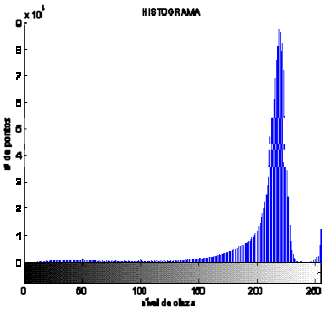
$$|e(t)| = \left| \frac{H'(t)}{H/\log(256)} - \alpha(H/\log(256)) \right|, \qquad (5)$$

onde α é um fator de perda determinado experimentalmente e dado por:

$$\alpha(H/\log(256)) = \begin{cases} -\frac{3}{7}H/\log(256) + 0{,}8 & \text{if } H/\log(256) < 0{,}7 \\ H/\log(256) - 0{,}2 & \text{if } H/\log(256) \geq 0{,}7 \end{cases} \qquad (6)$$



(a)



(b)

Fig. 2. (a) versão em níveis de cinza da imagem da Figura 1 e (b) seu histograma.

2) Otimização no algoritmo Silva-Lins-Rocha

Duas otimizações no algoritmo Silva-Lins-Rocha mostraram-se efetivas na melhoria da qualidade da segmentação da imagem. Uma diz respeito ao melhoramento do desempenho do algoritmo em imagens com um número de níveis de cinza (luminâncias) inferior a 256. A outra foi feita para que a separação entre o papel e a tinta interferente, segunda aplicação do algoritmo, seja feita de forma mais eficaz, visto que o algoritmo foi projetado para separar uma distribuição de tinta da frente do resto do documento.

Primeira Otimização

Observando-se os histogramas das imagens em níveis de cinza dos documentos do acervo de Joaquim Nabuco [5] foi constatado que há várias imagens que não apresentam os 256 níveis de cinza, algumas, por exemplo, com 232, 188, até 167 níveis. Dessa forma, a primeira alteração proposta é no cálculo da *entropia normalizada* que era dada por H/log(256) e agora passa a ser:

$$H_{N_{\text{níveis-de-cinza}}} = \frac{H}{\log(N_{\text{níveis-de-cinza}})}, \qquad (7)$$

onde $N_{\text{níveis-de-cinza}}$ é o número de níveis de cinza existentes na imagem. Assim a Equação 5 é agora substituída por:

$$|e(t)| = \left| \frac{H'(t)}{H_{N_{\text{níveis-de-cinza}}}} - \alpha\left(H_{N_{\text{níveis-de-cinza}}}, s, N_{\text{níveis-de-cinza}}\right) \right|. \qquad (8)$$

Isso faz com que no valor calculado da entropia normalizada se considere apenas a quantidade de níveis de cinza existentes na imagem.

Outra modificação que já aparece na Equação 8 é a mudança no fator de perda α, tal fator leva em consideração, além da entropia normalizada, o número de níveis de cinza presentes na imagem ($N_{\text{níveis-de-cinza}}$) e a medida s que é baseada no desvio padrão da distribuição a priori. Dessa forma, no lugar da Equação 6, passamos a utilizar:

$$\alpha\left(H_{N_{\text{níveis-de-cinza}}}, s, N_{\text{níveis-de-cinza}}\right) = \begin{cases} -\frac{3}{7} H_{N_{\text{níveis-de-cinza}}} + 0{,}8 & \text{if } H_{N_{\text{níveis-de-cinza}}} < 0{,}7 \\ H_{N_{\text{níveis-de-cinza}}} - 1.7s / N_{\text{níveis-de-cinza}} & \text{if } H_{N_{\text{níveis-de-cinza}}} \geq 0{,}7 \end{cases} \qquad (9)$$

onde s é dado por:

$$s = \sqrt{\sum_{j=0}^{N_{\text{níveis-de-cinza}}-1} (j-m)^2 q_j} \qquad (10)$$

sendo q_0, q_1, ..., $q_{N\text{níveis-de-cinza}-1}$ as probabilidades da distribuição a priori (Equação 3) que são diferentes de zero, preservando-se a ordem, e m um valor de média dado por:

$$m = \sum_{j=0}^{N_{\text{níveis-de-cinza}}-1} j \cdot q_j \,. \qquad (11)$$

A medida s é utilizada ao invés do desvio padrão para que o algoritmo seja imune a escalonamentos feitos no histograma.

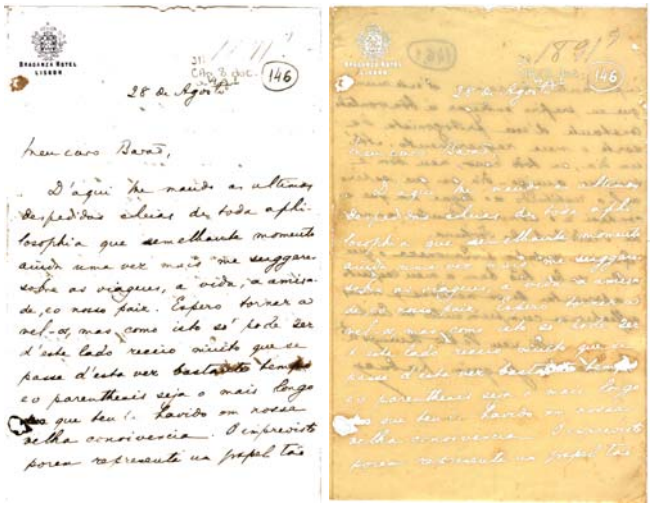
A aplicação do algoritmo com esta primeira otimização é utilizada para separar a tinta da frente do resto do documento.

Segunda Otimização

A segunda otimização é proposta para que tenhamos uma separação mais eficiente entre tinta interferente e papel. Aqui também é incorporada a mudança da entropia normalizada da mesma forma que foi feita para a primeira otimização. A diferença entre as duas está no valor do fator α. Para a primeira aplicação do algoritmo temos um fator de perda que assegura estarmos destacando apenas a tinta da frente do resto do documento. Para a segunda, é a interferência que deve ser destacada. Em linhas gerais, as características das distribuições do texto e da interferência são distintas, sendo a da segunda mais dispersa devido a sua natureza. Dessa forma, o valor de α na segunda aplicação do algoritmo é α=1.

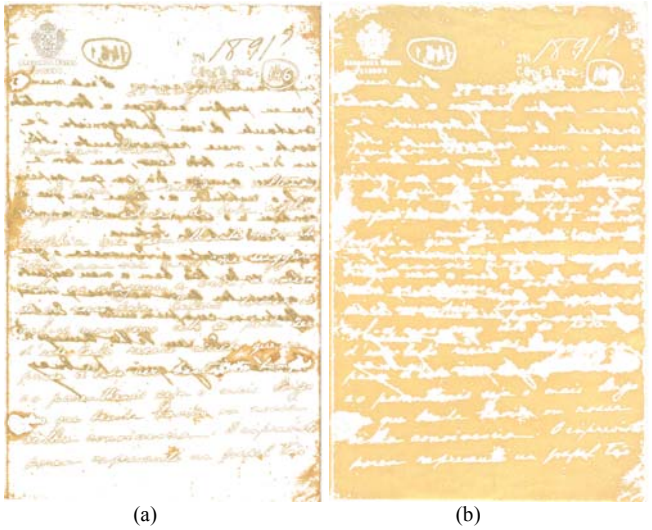
A partir das otimizações explicitadas acima, deve-se proceder da seguinte forma:

1. aplica-se o algoritmo de segmentação, fazendo uso da primeira otimização, para separar a tinta da frente do resto do documento (vide Figura 3); e
2. aplica-se novamente o algoritmo, agora com a segunda otimização, para separar a tinta interferente do papel (vide Figura 4).



(a) (b)

Fig. 3. Segmentos da imagem da Figura 1: (a) tinta da frente e (b) papel com interferência.



(a) (b)

Fig. 4. Segmentos da imagem Figura 3b: (a) interferência e (b) papel.

Dessa forma, temos identificados os *pixels* interferentes.

Para visualizarmos melhor como é feita a segmentação, observe a Figura 5. Essa figura apresenta o mesmo histograma da Figura 2b, pois o algoritmo trabalha apenas com a imagem na versão em níveis de cinza. O primeiro limiar T_L é obtido na primeira aplicação do algoritmo e o

segundo T_H a partir da segunda. Os *pixels* cujo valor de luminância é inferior a T_L são classificados como tinta da frente, os superiores a T_H são ditos pertencer ao papel e os maiores que T_L e menores que T_H são discriminados como interferência.

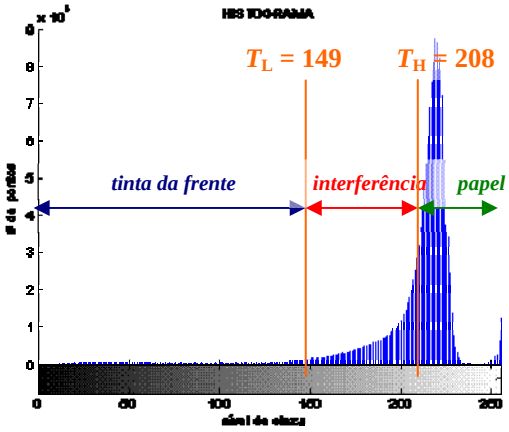


Fig. 5. Histograma da imagem da Figura 2a – detalhes da segmentação.

B. *Preenchimento da Área Interferente*

Na seção II.A foi apresentada uma estratégia para identificar os *pixels* pertencentes à interferência frente-verso. Agora vamos fazer o preenchimento desta área com *pixels* existentes no papel. A idéia é bem simples:

1. toma-se uma grande área do documento, tendo o cuidado de não incluir sua borda (foi escolhida neste trabalho uma janela de largura 60% da largura da imagem, altura 60% da altura da imagem e centralizada no centro do documento);
2. varre-se a área definida armazenando-se em um vetor – *amostra_do_papel* – os *pixels* (RGB) que tiverem um valor de luminância – nível de cinza – maior que T_H (vide Figura 5), essa condição garante que os *pixels* selecionados são provenientes do papel;
3. preenche-se, finalmente, a área interferente escolhendo-se aleatoriamente *pixels* do vetor *amostra_do_papel*.

Deve-se salientar que a área citada no ítem 1, logo acima, não deve conter a borda do documento pelo fato da mesma, geralmente, ser muito clara, ou seja, ter altos valores de luminância (nível de cinza), o que a classifica como papel (vide Figura 5). Sendo assim, seus *pixels* fariam parte do vetor *amostra_do_papel* e, consequentemente, eles estariam presentes no preenchimento da área interferente.

É importante destacar que devido à interferência ter valores de luminância dispersos, seus *pixels* mais escuros têm intensidades luminosas (níveis de cinza) semelhantes às dos *pixels* mais claros da tinta da frente, por outro lado, os mais claros se confundem com os *pixels* mais escuros do papel. Dessa forma, a imagem da Figura 4a traz, além da interferência, o contorno mais claro da tinta da frente e as partes mais escuras do papel. Esse fato não degrada a qualidade visual do resultado final da aplicação do algoritmo de filtragem aqui proposto, como pode ser visto na Figura 6, pois os *pixels* pertencentes ao papel, classificados como interferência, são preenchidos por *pixels* do próprio papel, já o preenchimento dos contornos da tinta da frente tornam as letras do documento ligeiramente mais estreitas.

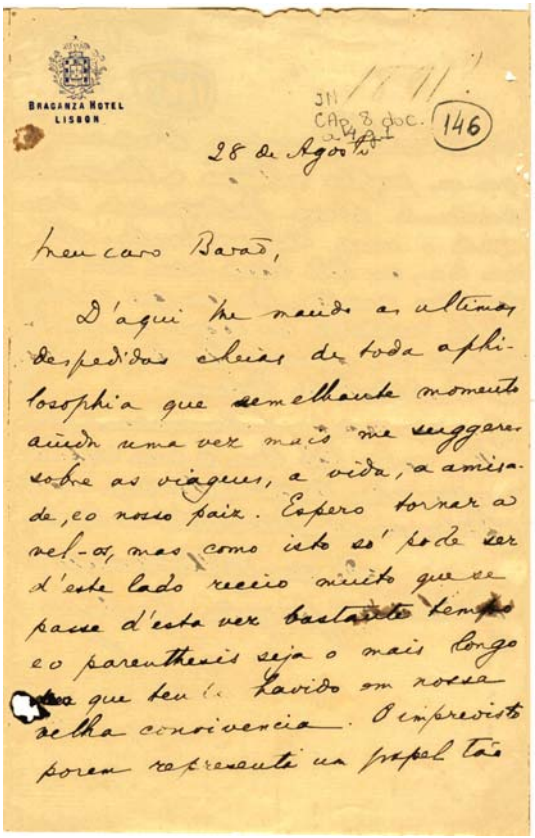
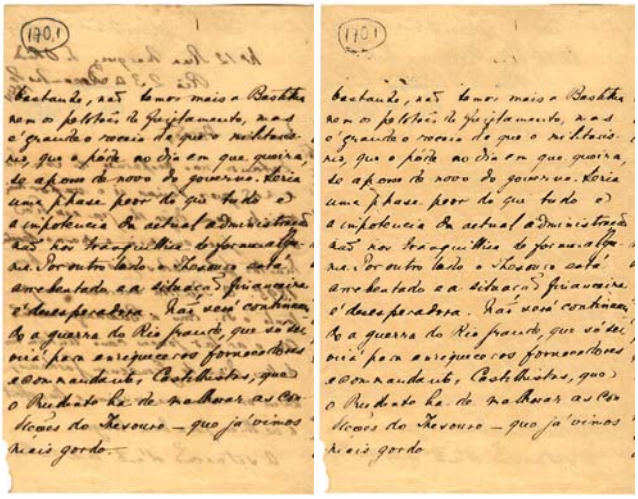


Fig. 6. Resultado da aplicação do algoritmo de filtragem proposto na imagem da Figura 1.

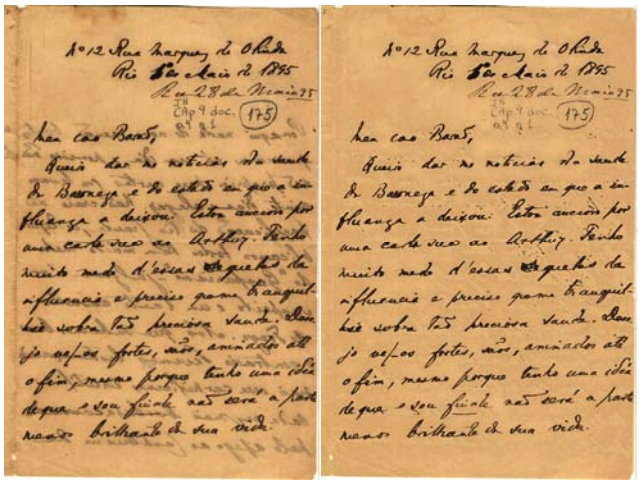
III. Análises e Resultados

O algoritmo apresentado na última seção foi testado em 260 imagens do acervo de documentos digitalizados de Joaquim Nabuco [5] trazendo resultados satisfatórios. Três amostras das 260 imagens são mostradas nas Figuras 7a, 8a e 9a e os resultados da aplicação do algoritmo proposto sobre tais imagens são mostrados nas Figuras 7b, 8b e 9b respectivamente.



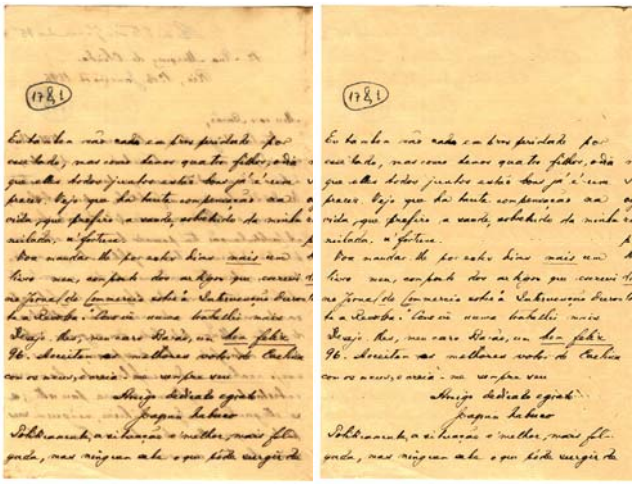
(a) (b)

Fig. 7. Imagens dos documentos
(a) original e (b) filtrado com o novo algoritmo.



(a) (b)

Fig. 8. Imagens dos documentos
(a) original e (b) filtrado com o novo algoritmo.



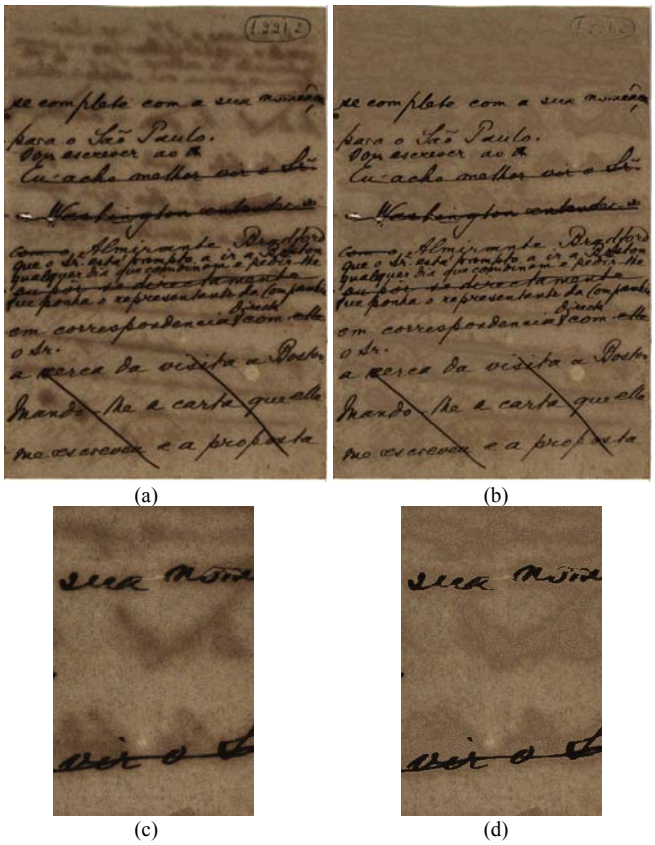
(a) (b)

Fig. 9. Imagens dos documentos
(a) original e (b) filtrado com o novo algoritmo.

É importante relatar que o algoritmo aqui proposto não teve um desempenho tão bom nas imagens cuja interferência era muito dispersa (muito "borrada") no que diz respeito à separação da interferência do papel, como pode ser visto nas imagens da Figura 10. As ampliações de um mesmo trecho das imagens das Figuras 10a (original) e 10b (filtrada) são apresentadas nas Figuras 10c e 10d, respectivamente. Através das últimas é possível observar que boa parte do contorno da interferência permaneceu na imagem filtrada.

Para amenizar tal problema, na segunda aplicação do algoritmo pode-se utilizar um fator de perda α que leve em conta algumas informações sobre a distribuição a ser particionada, semelhante ao feito na primeira aplicação.

Foi observado que há poucos casos em que uma grande porção da distribuição da interferência está imersa na distribuição do papel, em termos do histograma em níveis de cinza. Em tais casos, a interferência é muito fraca e não é indicada a utilização de algoritmos de limiarização globais. A utilização de tais algoritmos resultará em imagens com preservação do papel, mantendo boa parte da interferência, ou terá a interferência retirada, pagando o preço de não preservar detalhes do papel.



(a) (b) (c) (d)

Fig. 10. Imagens dos documentos (a) original e (b) filtrado com o novo algoritmo. Ampliação de um mesmo trecho das imagens (c) original e (d) filtrada pelo novo algoritmo.

IV. CONCLUSÕES E TRABALHOS FUTUROS

Foi introduzido neste artigo um sistema de remoção da interferência frente-verso em imagens de documentos coloridos. A aplicação desse sistema em tais documentos melhora suas legibilidades. Esse sistema utiliza o algoritmo Silva-Lins-Roch, com duas otimizações, para discriminar os *pixels* provenientes da interferência. Após a discriminação é feito um preenchimento da área interferente com cores existentes no papel do documento em questão. O sistema aqui introduzido é uma automatização do apresentado em [2].

O sistema de filtragem aqui proposto pode ter seu processo de preenchimento da área interferente melhorado, substituindo os *pixels* provenientes da interferência pelos seus vizinhos pertencentes ao papel.

Outra melhoria, mencionada na seção anterior, é a utilização de um fator de perda α, na segunda aplicação do algoritmo, que leve em conta algumas informações sobre a distribuição a ser particionada, semelhante ao feito na primeira aplicação.

Além da apresentação do sistema de filtragem, este trabalho trouxe uma melhoria para o algoritmo de segmentação Silva-Lins-Rocha [3].

AGRADECIMENTOS

Ao CNPq – Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico, Governo do Brasil – pelo suporte financeiro e à FUNDAJ – Fundação Joaquim Nabuco – pela permissão de utilização das imagens.

REFERÊNCIAS

[1] R. D. Lins, et al. “An Environment for Processing Images of Historical Documents. Microprocessing & Microprogramming”, pp. 111-121, North-Holland, 1994.

[2] R. D. Lins and I. G. Netto. “Um Ambiente para Filtrar a Interferência Frente-Verso em Documentos Históricos”. *SBrT-2007*, Recife, Brasil, 2007.

[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-Document Engineering-2006*, Dijon, France, 2006.

[4] N. Abramson, “Information Theory and Coding”, McGraw-Hill Book Co, 1963.

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