Um Ambiente para Filtrar a Interferência Frente-Verso em Documentos Históricos

Rafael Dueire Lins e Ismael Gomes Netto

***Resumo*— Freqüentemente documentos são estritos em ambos os lados de um papel translúcido fazendo a tinta de um lado ser visível do outro. Este fenômeno foi chamado de "interferência frente-verso". Este artigo apresenta uma nova forma de remover tal interferência deixando o usuário comandar o processo de filtragem através da seleção de amostras do documento e da interferência. Esta nova técnica de filtragem trás bons resultados.**

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

***Abstract*— Very often documents are written on both sides on translucent paper making visible the ink from one side on the other. This artifact was called "back-to-front interference". This paper presents a new way to remove such interference by letting the user drive the filtering process by selecting samples of the document and the interference. This new filtering technique yields good results.**

***Keywords*— Back-to-front interference, Historical Documents.**

I. INTRODUÇÃO

O algoritmo apresentado aqui é parte de um grande projeto para filtrar documentos históricos de origem do século XIX pertencentes ao acervo de Joaquim Nabuco, escritor, diplomata e estadista, uma das figuras mais importantes na campanha de libertação dos escravos no Brasil (1861-1910). Este rico arquivo é mantido pela FUNDAJ – Fundação Joaquim Nabuco [1] (um instituto de pesquisa em ciências sociais em Recife-Brasil) – e contém mais de 6.000 cartas. Estas cartas são de suma importância para o entendimento da formação da estrutura política e social dos países das Américas e suas relações com outros países.

Embora a humanidade use papéis de mais de 2.000 anos para guardar informações, a fabricação do papel na época de Nabuco acrescentou muito mais branqueador e estes documentos correm o risco de uma decomposição mais rápida. Assim, o objetivo do Projeto Nabuco é preservar tal informação para futuras gerações e torná-la completamente acessível disponibilizando-as on-line. Para permitir uma boa compactação de tais arquivos e uma eficiente transmissão via rede, os documentos são disponibilizados para consulta em suas versões monocromáticas. A binarização deste tipo de documento é mais difícil do que os mais recentes porque enquanto o papel escurece com o tempo, a parte impressa, tanto datilografada como escrita, tende a esmaecer. Uma dificuldade especial aparece em documentos escritos ou datilografados em ambos os lados e a opacidade do papel é tal que permite a impressão do verso ser visualizada na frente. Um novo conjunto de cores de papel e impressão aparece complicando o processo de binarização de forma que a aplicação direta de algoritmos gerais é completamente inapropriada e traz documentos ilegíveis. Este fenômeno primeiramente referenciado em [2] foi chamado de interferência frente-verso, posteriormente chamado de *bleeding* [3] ou *show-through* [4]. A Figura 1 apresenta uma carta com tal interferência e sua binarização direta aparece na Figura 2.

A literatura apresenta diversos esquemas automáticos para remover a interferência frente-verso. A de melhores resultados parece ser a técnica de limiarização baseada em entropia [5], embora outros pesquisadores tenham sugerido modelos de filtragem baseados em *waterflow* [6] e *wavelet* [7]. Outra abordagem encontrada é baseada na técnica de filtragem em espelho originalmente sugerida em [2] e adotada com sucesso em [4]. Uma dificuldade aparece nesta última técnica para se fazer o alinhamento das imagens dos dois lados. Todos os algoritmos reportam limitações em diferentes tipos de imagens (papel de fundo muito escuro, impressão muito esmaecida, interferência restrita a parte do documento, etc.). Um esquema de filtragem mais complexo é proposto por Nishida e Suzuki [8] onde primeiro os componentes do objeto são separados do fundo e interferência através de uma binarização local adaptativa para cada componente de cor e limiarização de contorno [9]. Cores do fundo são estimadas localmente através limiares de cor para gerar uma imagem restaurada, e então corrigir adaptativamente usando análise *multi-scale* através da comparação das distribuições de contorno entre as imagens original e restaurada. Devido a natureza dos documentos do acervo de Nabuco detecção de contorno parece ser de pouca ajuda na eliminação da interferência frente-verso, assim o método de Nishida-Suzuki parece não ser apropriado para este tipo de documento, embora isto ainda não tenha sido provado por experimentos.

Este artigo propõe uma nova forma de filtrar a interferência frente-verso em documentos coloridos onde o usuário comanda o processo de filtragem através da seleção de amostras das três componentes distintas do documento: papel, escrita e interferência.

Rafael Dueire Lins e Ismael Gomes Netto¸ Departamento de Eletrônica e Sistemas, Centro de Tecnologia e Geociências, Universidade Federal de Pernambuco, Recife, Brasil, E-mails: rdl@ufpe.br, ismaelgomesn@gmail.com.

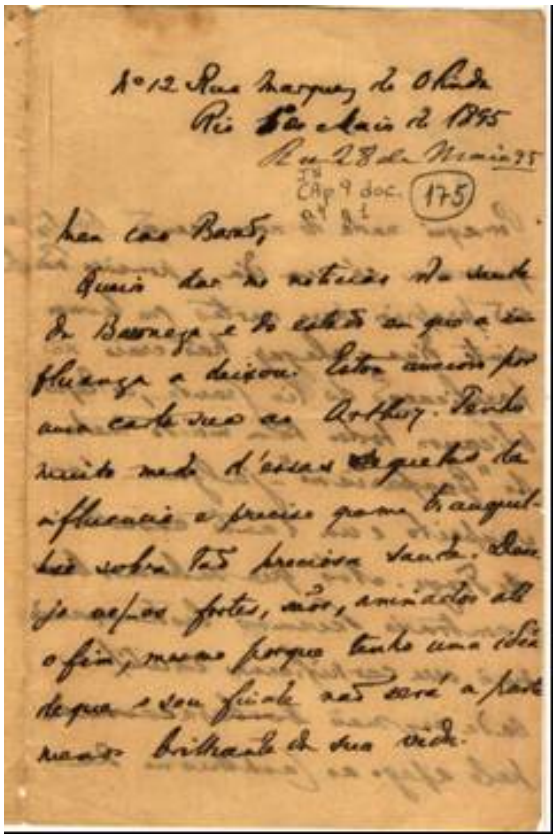


Fig. 1. Documento do acervo de Joaquim Nabuco.

Fig. 2. Documento da Figura 1 binarizado.

II. DESCRIÇÃO DO MÉTODO

A estratégia de filtragem proposta aqui é dividida em dois passos:

1. *seleção de pixels*: o usuário seleciona alguns *pixels* da imagem original;
2. *algoritmo de filtragem*: este passo usa os pontos selecionados para determinar a área da interferência frente-verso e a preenche com pixels do papel.

Os dois passos do método proposto são detalhados a seguir.

A. Seleção de Pixels

A ferramenta de filtragem abre o documento a ser filtrado e pede ao usuário que clique em áreas representativas do documento:

- 1 amostra da "Interferência Escura", a área da tinta proveniente do verso do documento que pareça mais escura para o usuário;
- 1 amostra da "Interferência Clara", a área onde a tinta proveniente do verso do documento pareça mais clara para o usuário;
- 2 amostras da "Informação Clara", as áreas da tinta provenientes da frente do documento que pareçam mais claras para o usuário;
- 4 amostras do "Papel", áreas não escritas do papel do documento.

A Figura 3 mostra um exemplo do processo de seleção.

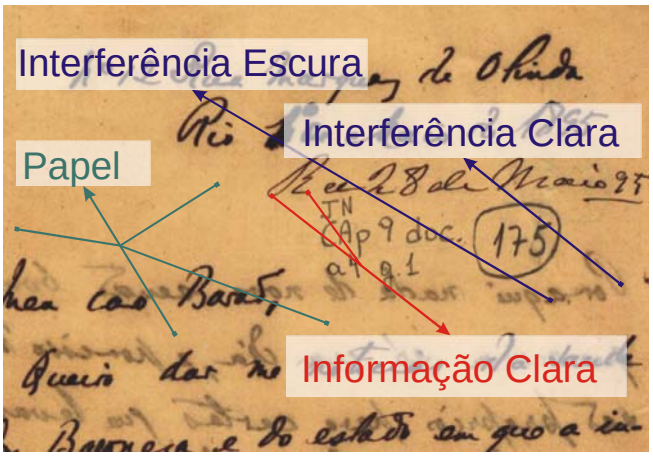


Fig. 2. Exemplo da seleção das amostras na imagem da Figura 1.

Das amostras selecionadas são armazenadas suas posições.

B. Algoritmo de Filtragem

O algoritmo de filtragem começa tomando a versão em níveis de cinza da imagem original. A transformação de *true-color* para níveis de cinza é obtida por

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

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 em *true-color*; e $gray(i,j)$ é o nível de cinza correspondente na imagem em níveis de cinza. A Figura 4

apresenta a versão em níveis de cinza da imagem da Figura 1 e seu histograma é apresentado na Figura 5. Pode-se observar que a imagem em níveis de cinza permanece legível.

Na imagem em níveis de cinza localizam-se as amostras selecionadas e para cada uma delas armazena-se uma janela 3x3, centrada na posição selecionada.

Na imagem em níveis de cinza, localizam-se as posições selecionadas e armazenam-se os níveis de cinza da janela 3x3, centrada na posição amostrada. No caso das posições referentes às amostras do papel devem ser armazenados os níveis de cinza da janela 7x7, centrada na posição amostrada.

Os níveis de cinza armazenados são utilizados para determinar dois limiares: o baixo T_L e o alto T_H. O limiar baixo separa a informação da interferência frente-verso e o alto separa a última do papel, em outras palavras, assume-se que todo *pixel* da interferência tem seu nível de cinza maior que T_L e menor que T_H como ilustrado na Figura 5.

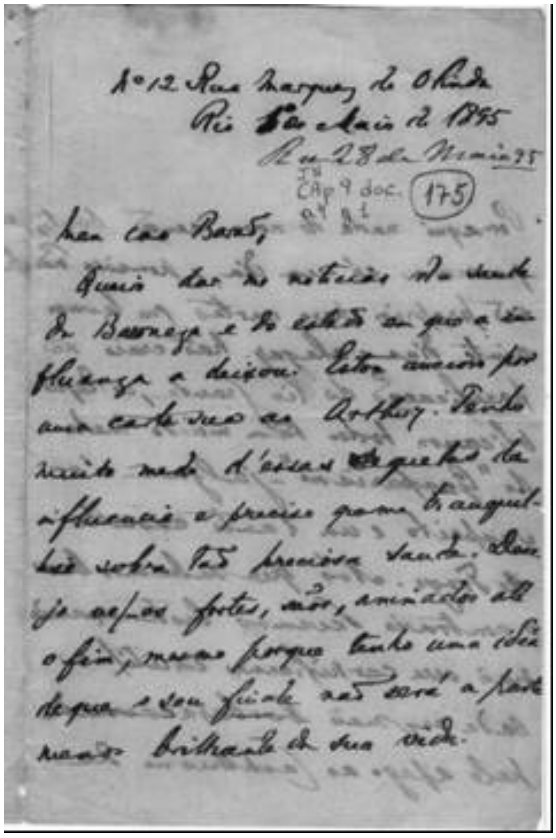


Fig. 4. Versão em níveis de cinza da imagem da Figura 1.

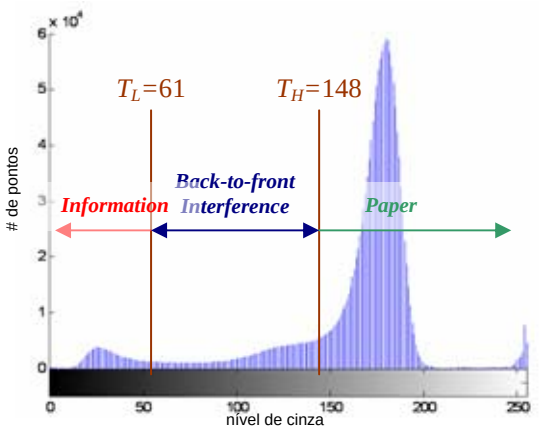


Fig. 5. Histograma da imagem da Figura 4.

Os parâmetros para determinar o limiar baixo são as janelas da "informação clara" e da "interferência escura". De cada janela seleciona-se o *pixel* mais escuro definido por:

- $P_1^{(\text{informação clara})}$ – o *pixel* mais escuro presente na primeira janela da "Informação Clara";
- $P_2^{(\text{informação clara})}$ – o *pixel* mais escuro encontrado na segunda janela da "Informação Clara";
- $P^{(\text{interferência escura})}$ – o *pixel* mais escuro na janela da "Interferência Escura";

Diretamente, o limiar baixo (T_L) é o nível de cinza de maior valor dentre os três definidos acima.

O limiar alto é obtido usando as janelas de "Papel" e "Interferência Clara". Similarmente, definem-se os seguintes *pixels*:

- $P^{(\text{interferência clara})}$ – o *pixel* mais claro presente na janela da "Interferência Clara";
- $P_n^{(\text{papel})}$ – o *pixel* mais escuro na janela *n* do "Papel", onde n=1,2,3,4.

Se $P^{(\text{interferência clara})}$ é o *pixel* mais escuro dos cinco definidos acima, o limiar (T_H) será o mais escuro dentre os $P_n^{(\text{papel})}$, n=1,2,3,4. Caso contrário, ele será o limiar.

Depois da identificação dos *pixels* da interferência frente-verso, eles são substituídos por *pixels* das quatro janelas de "Papel" escolhidas. Esta etapa é realizada na versão em *true-color* da imagem original.

III. RESULTADOS

Nesta seção serão apresentados três dos cinqüenta experimentos feitos usando o método proposto aqui. É importante dizer que os outros quarenta e sete resultados tiveram qualidades semelhantes.

As Figuras 6 e 7 mostram a imagem da Figura 1 filtrada pelo algoritmo proposto e seu histograma, respectivamente.

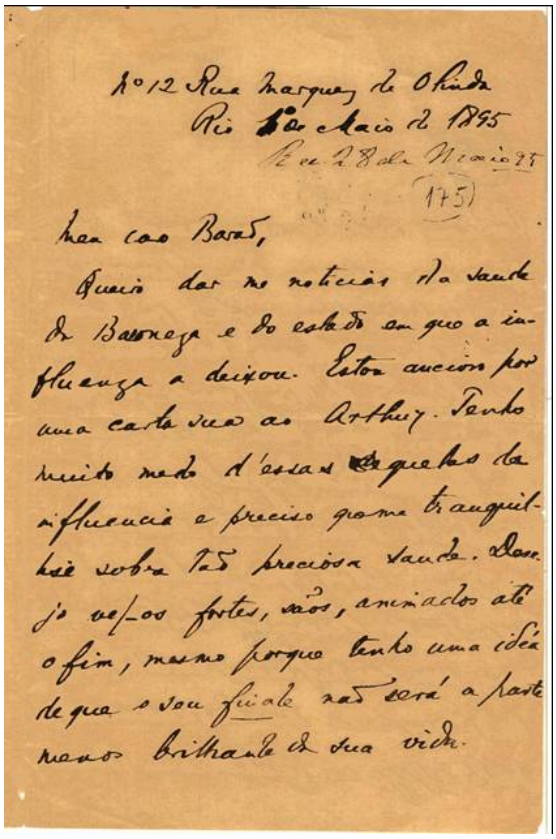


Fig. 6. Imagem da Figura 1 filtrada

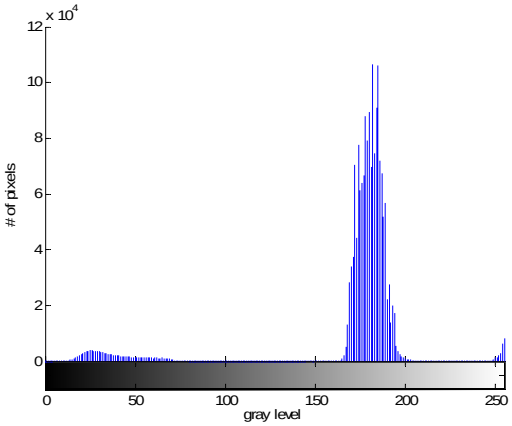
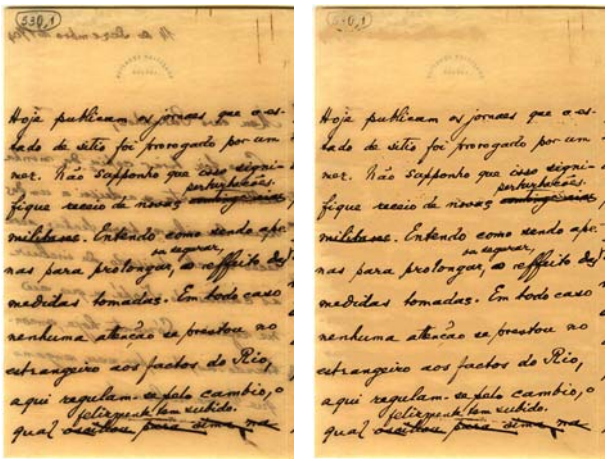


Fig. 7. Histograma da versão em níveis de cinza da imagem da Figura 6.



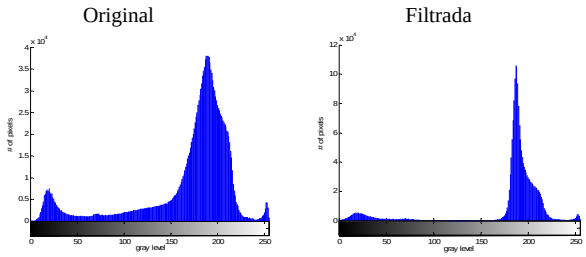


Fig. 8. Imagens original e filtrada, e histogramas em níveis de cinza.

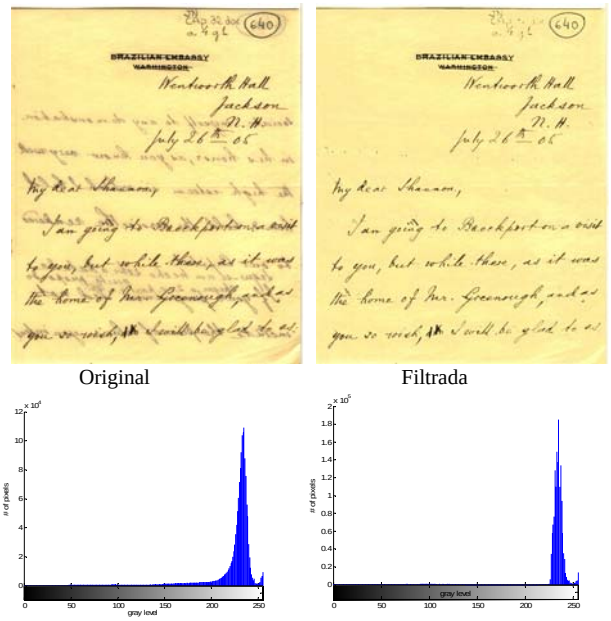


Fig. 9. Imagens original e filtrada, e histogramas em níveis de cinza.

As Figuras 8 e 9 apresentam dois experimentos, cada um deles traz a imagem original, a imagem filtrada e seus respectivos histogramas.

Nos histogramas das imagens filtradas pode-se observar que aparece um salto entre T_L e T_H. Isto ocorre pelo fato de todos os *pixels* da interferência serem substituídos por *pixels* do papel que são, em sua maioria, maiores que T_H. Assim, a distribuição referente à informação permanece inalterada e a referente ao papel sofre uma ligeira mudança.

IV. Conclusões e Trabalhos Futuros

Este artigo apresenta uma nova estratégia para filtrar a interferência frente-verso em imagens de documentos. A idéia é achar dois limiares no espaço de luminância para identificar a interferência e substituir sues *pixels* por *pixels* do papel.

O primeiro passo do processo é selecionar as diferentes áreas do documento. Este passo é feito manualmente por um operador, mas trabalhos em andamento tentam torna-lo automático. Vários algoritmos na literatura ([5], [10], [11], [12]) tentam separar o texto do papel em um dado documento. Este tipo de algoritmo pode ter sucesso quando usado para calcular automaticamente o limiar baixo (T_L). Contudo, encontrar o segundo limiar (T_H) é mais difícil. Este é um dos pontos deixados para trabalhos futuros.

Se o primeiro passo é feito automaticamente, o processo

como um todo pode ser feito também. Se isso for conseguido, pode-se aplicar o método proposto em uma imagem dividida, o que pode melhorar o resultado como um todo.

Uma forma de se calcular os limiares T_L e T_H sem a necessidade da coleta de amostras da imagem é proposta em [ref], contudo o método de avaliação quantitativo introduzido em mostra que as imagens filtradas pelo método aqui proposto apresentam maior qualidade.

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] FUNDAJ: www.fundaj.gov.br

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

[3] R.Kasturi, L. O´Gorman and V.Govindaraju, “Document image analysis”: A primer, Sadhana, (27):3-22, 2002.

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

[5] 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.

[6] O.Hyun-Hwa, et al. “An improved binarization algorithm based on a waterflow model for document image with inhomogeneous backgrounds”. Pattern Recognition 38 (2005) 2612 – 2625, 2005.

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

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

[9] A. Cumani, “Edge detection in multispectral images”, G. Models and Image Processing, 53(1):40-51, 1991.

[10] E. Kavallieratou and H. Antonopoulou, “Cleaning and Enhancing Historical Document Images”, Int. Vision Systems, LNCS3708: 681-688, Springer-Verlag 2005.

[11] G. Leedham, et. al. “Separating text and background in degraded document images—a comparison of global thresholding techniques for multi-stage thresholding”, 8th WS Front. in Handwritten Recog., 244–249, 2002.

[12] B.Sankur and M.Sezgin, “A survey over image thresholding techniques and quantitative performance evaluation”. J. Elec. Imaging, 13(1), 146-165 (2004).

[13] J. M. M. da Silva e R. D. Lins. “Um Sistema de Filtragem de Interferência Frente-Verso em Documentos Coloridos”. *SUBMETIDO AO SBrT-2007*, Recife, Brasil, 2007.

[14] R. D. Lins, J. M. M. da Silva e I. G. Netto, “Método de Avaliação Quantitativo para Algoritmos que Filtram Interferência Frente-Verso em Documentos Coloridos” ”. *SUBMETIDO AO SBrT-2007*, Recife, Brasil, 2007.