Melhorando A Qualidade de Documentos Coloridos com Interferência Frente-Verso

J. M. M. da Silva, Rafael Dueire Lins e G. F. P. e Silva

***Resumo*— A interferência frente-verso ocorre em documentos escritos (ou impressos) em ambos os lados de papel translúcido. Tal interferência, dificulta sua transcrição automática e binarização. Este artigo apresenta uma nova técnica de filtragem de documentos coloridos com interferência frente-verso que objetiva a melhoraria da legibilidade do mesmo.**

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

***Abstract*— The back-to-front interference occurs on documents written (or printed) on both sides of a translucent paper. Such interference makes more difficult their transcription and binarization. This paper presents a new technique to filter out such interference in color documents, enhancing readability.**

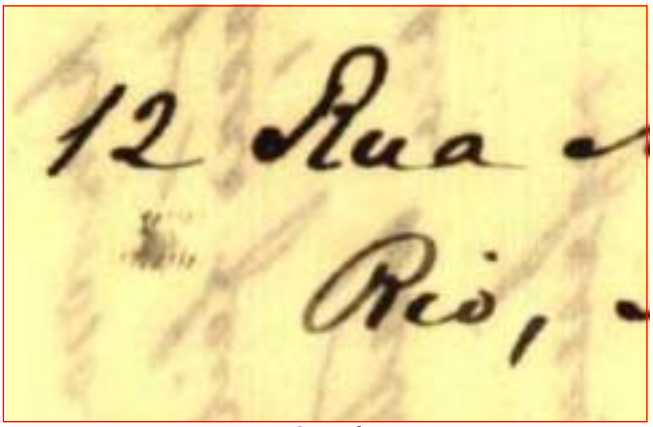
***Keywords*— Back-to-front interference, bleeding, show-through, historical documents, segmentation, interpolation.**

I. INTRODUÇÃO

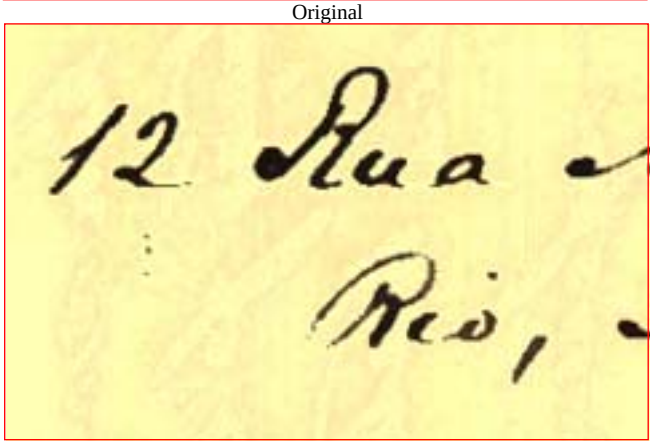
No início dos anos 90, processou-se a digitalização do acervo de correspondências de Joaquim Nabuco, através de trabalho conjunto realizado entre a Fundação Joaquim Nabuco e a Universidade Federal de Pernambuco [1]. Do rico acervo de aproximadamente 6.500 cartas, cerca de 10% das imagens dos documentos digitalizados apresentavam uma característica não anteriormente descrita na literatura e que passou a ser conhecida como interferência frente-verso (do inglês *back-to-front interference*) [1]. Posteriormente, outros autores utilizaram os nomes de *bleeding* [2] e *show-through* [3] para este mesmo efeito.

A interferência frente-verso ocorre quando o conteúdo da face do verso de um documento se faz presente na face frontal. Para que tal interferência apareça em um documento é necessário que este seja escrito (ou impresso) em ambos os lados de papel translúcido (vide primeira imagem da Figura 1). Essa interferência em documentos degrada seus processos de transcrição automática e binarização. Em documentos históricos, o envelhecimento do papel é mais um fator de dificuldade, pois o seu escurecimento diminui o "grau de separação" entre a tinta de cada um dos lados e o papel.

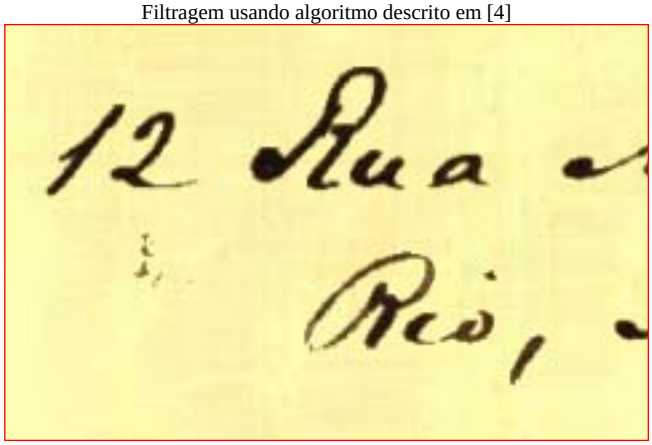
Este artigo apresenta uma nova estratégia de filtragem da interferência frente-verso em imagens de documentos coloridos, tendo melhores resultados do que os apresentados em [4]. Nas duas estratégias, a idéia é discriminar a área interferente e realizar um preenchimento sobre a mesma.



Original



Filtragem usando algoritmo descrito em [4]



Filtragem usando estratégia proposta

Fig. 1. Zoom em parte de documento do acervo de Nabuco com interferência frente-verso.

A principal diferença entre a nova estratégia e a descrita em [4] está no processo de preenchimento da área interferente. A anterior preenche a área destacada aleatoriamente, utilizando *pixels* previamente classificados

João Marcelo Monte da Silva, Rafael Dueire Lins e Gabriel de França Pereira e Silva¸ Departamento de Eletrônica e Sistemas, Centro de Tecnologia e Geociências, Universidade Federal de Pernambuco, Recife, Brasil, E-mails: joao.mmsilva@ufpe.br, rdl@ufpe.br, gfps@cin.ufpe.br.

como papel, enquanto a nova realiza uma interpolação "linear". Para que se possa ter uma idéia prévia da melhoria implantada, é mostrado na Figura 1 um mesmo trecho ampliado das imagens de um documento. Percebe-se que o preenchimento promovido pela nova estratégia (terceira imagem) possui um aspecto mais natural. Além disso, o "contorno da interferência" remanescente no resultado da filtragem usando o algoritmo anterior (vide segunda imagem), não mais aparece no da nova.

Na seção II, é feita a descrição desse novo sistema de filtragem. Os resultados e as análises estão apresentados na seção III. Finalmente, na seção IV, são apresentadas as conclusões e linhas para trabalhos futuros.

II. Sistema de Filtragem

Esta seção apresenta a nova estratégia para remover a interferência frente-verso de imagens de documentos coloridos, melhorando os resultados apresentados em [4].

A idéia básica é discriminar a área correspondente à interferência frente-verso (primeira etapa) e preenchê-la com *pixels* cujas cores mais se assemelham às cores do papel (segunda etapa). A melhoria proposta aqui, se dá nas duas etapas mencionadas, as quais serão tratadas separadamente.

A. *Discriminação dos Pixels da Interferência*

Para se encontrar a área interferente, o algoritmo de segmentação Silva-Lins-Rocha [5] é utilizado duas vezes: a primeira, para separar o texto do resto do documento; e a segunda, para destacar a interferência do papel. Em linhas gerais, as características das distribuições do texto e da interferência são distintas, sendo a da segunda mais dispersa.

O *fator de perda* (α) é um parâmetro do algoritmo de segmentação utilizado que deve garantir um melhor ajuste estatístico entre as distribuições das imagens original e binarizada, baseado na entropia de Shannon [6]. Para a segunda aplicação, propõe-se uma pequena alteração nesse fator, que passa a ser constante (α=1), garantindo uma melhor separação entre a interferência e o papel.

Em suma, para detectar a área interferente através desta estratégia:

1. aplica-se o algoritmo de segmentação [5] para separar a tinta da frente do resto do documento (vide Figuras 2a e 2b); e
2. aplica-se novamente o algoritmo, agora com a alteração do *fator de perda* acoplada, para separar a tinta interferente do papel (vide Figuras 2c e 2d).

Agora, têm-se identificados os *pixels* interferentes.

Para uma melhor visualização de como é feita a segmentação, a Figura 3 apresenta o histograma da versão em níveis de cinza da imagem de um documento com interferência frente-verso. 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 nível de cinza é inferior a T_L são classificados como tinta da face frontal, 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.

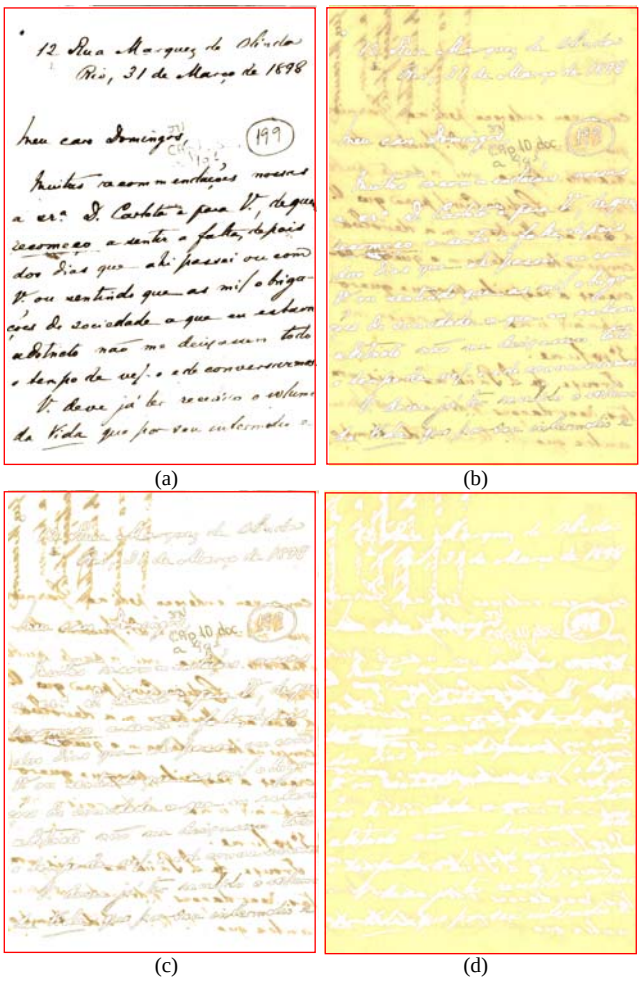


Fig. 2. Segmentos da imagem de um documento com interferência frente-verso: (a) tinta da frente e (b) papel com interferência. Segmentos da imagem da Figura 2b: (c) interferência e (d) papel.

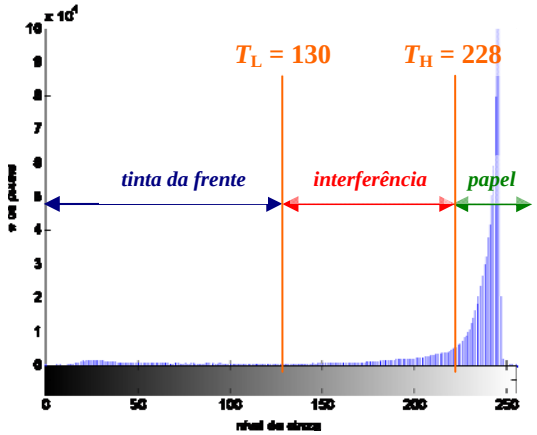
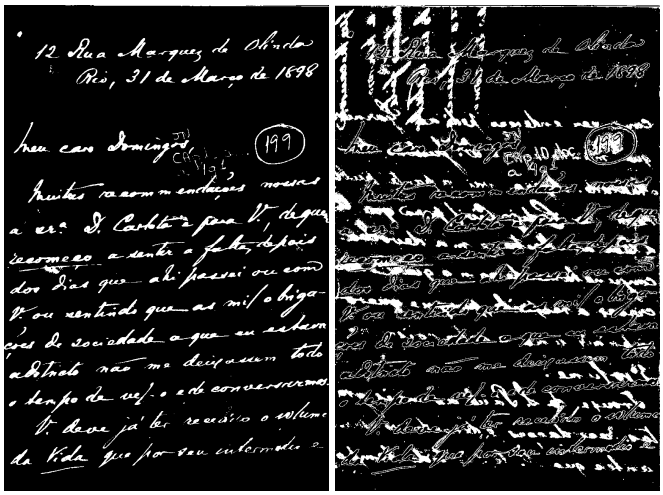


Fig. 3. Histograma da imagem de um documento com interferência frente-verso – detalhes da segmentação.

B. *Preenchimento da Área Interferente*

O preenchimento da área interferente aqui proposto utiliza uma interpolação "linear" para preencher a área interferente, diferentemente do algoritmo apresentado em [4] que preenche tal área com *pixels* pertencentes ao papel escolhidos aleatoriamente.



(a) (b)

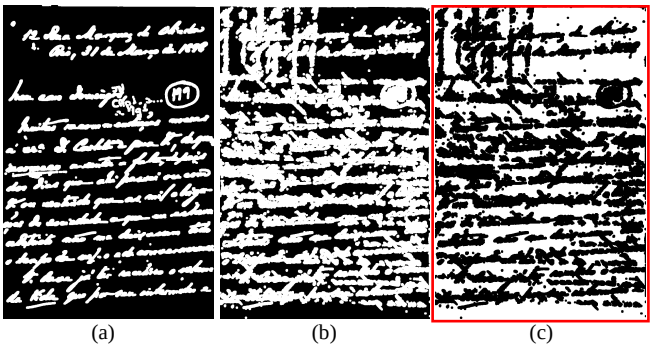
Fig. 4. Máscaras que identificam (a) o texto e (b) a interferência.

O novo processo define duas máscaras binárias: TEXTO e INTERF. A primeira identifica os *pixels* provenientes do texto da frente (vide Figura 4a), a segunda destaca a área interferente (vide Figura 4b). Pode-se supor que apenas a máscara INTERF seja suficiente para o processo de preenchimento, pois os *pixels* que se deseja substituir "já estão discriminados". Contudo, há o aparecimento de algumas dificuldades.

A idéia é substituir as cores dos *pixels* da interferência por cores o mais próximo possível do papel naquela região. Isto é conseguido através de uma interpolação, a qual faz uso das cores dos *pixels* que estão "no contorno" da área a ser interpolada. Se o texto estiver muito próximo (ou seja, na periferia) da área interferente, seus *pixels* participarão do processo de interpolação, o que trará, novamente, cores relativamente escuras para a área interferente, isso vai contra o objetivo de tornar tal área "o mais próximo possível do papel". Dessa forma, antes de interpolar deve-se também retirar o texto. Isso justifica a utilização da máscara TEXTO.

Outro problema que surge após a segmentação é a permanência (no "papel") do contorno do texto e da interferência (vide Figuras 2d e 4b). Este fato certamente torna o processo de interpolação ineficiente, pois as cores dos pixels interferentes seriam substituídas pelas presentes no contorno do texto e da interferência. Para solucionar esse problema, deve-se aplicar uma operação morfológica de dilatação nas máscaras, isso faz com que os contornos do texto e da interferência sejam corretamente classificados como "texto" e "interferência", respectivamente (vide Figuras 5a e 5b).

Como mencionado anteriormente os *pixels* que farão parte do processo de interpolação são os que contornam a área interferente e pertencem ao papel. Dessa forma, cria-se uma nova máscara que destaca os *pixels* pertencentes apenas ao papel. Tal máscara PAPEL pode ser adquirida pelo complemento da resultante da operação lógica OU entre as máscaras TEXTO e INTERF já dilatadas (vide Figura 5c).



(a) (b) (c)

Fig. 5. Máscaras dilatadas: (a) texto (T) e (b) interferência (I). (c) $\overline{\mathrm{T\,OR\,I}}$.

Os *pixels* que serão utilizados no processo de interpolação estão destacados na máscara PAPEL (Figura 5c); e os que serão interpolados são os que aparecem na máscara INTERF dilatada (Figura 5b), mas não estão presentes na máscara TEXTO (Figura 5a). Essa condição do último uso da máscara de TEXTO é imposta para que o processo de interpolação não altere os pixels classificados como texto. Se tal condição não fosse imposta, parte do texto seria "apagada".

Agora, será apresentado o processo de interpolação. Para um melhor entendimento, deve-se observar a imagem da Figura 6.

Sejam as coordenadas:

- (x_0,y_0) de um ponto P do intervalo a ser interpolado;
- (x_0,y_1) do ponto P_N – primeiro ponto ao norte de P;
- (x_0,y_2) do ponto P_S – primeiro ponto ao sul de P;
- (x_1,y_0) do ponto P_O – primeiro ponto à oeste de P;
- (x_2,y_0) do ponto P_L – primeiro ponto à leste de P.

Seja $i_C(x,y)$ a intensidade da componente C (R, G ou B) do *pixel* nas coordenadas (x,y). A intensidade do *pixel* (P) interpolado é dada por

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

onde os valores d_k's e i_k's (k=1,...,4) representam as intensidades e as distâncias dos pontos definidos – P_N, P_S, P_O e P_L – ao ponto P, seguindo uma ordem crescente em relação às distâncias. Por exemplo, o ponto mais próximo de P tem distância d_1 e intensidade i_1, o segundo mais próximo tem distância d_2 e intensidade i_2, e assim por diante.

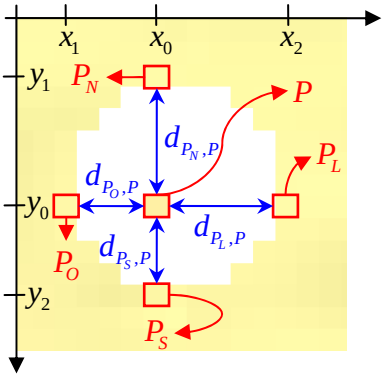
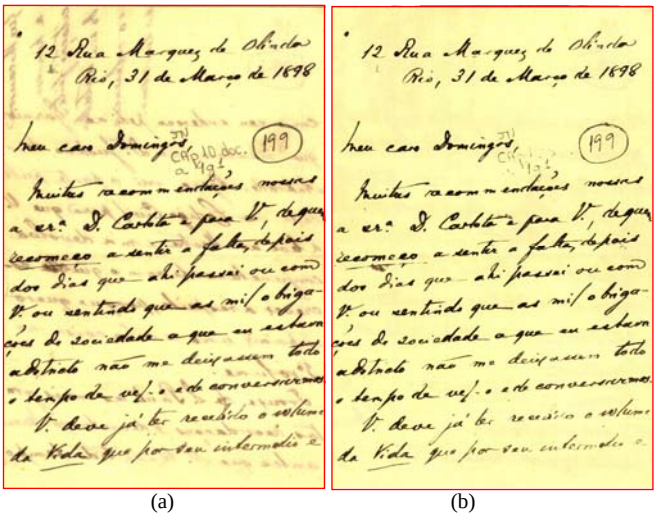


Fig. 6. Processo de interpolação.



(a) (b)

Fig. 7. Imagens (a) original e (b) resultado da aplicação da nova estratégia de filtragem proposta.

A distância entre dois pontos quaisquer P_a e P_b com coordenadas (x_a,y_a) e (x_b,y_b), respectivamente, é definida por

$$d_{Pa,Pb} = \sqrt{(x_a - x_b)^2 + (y_a - b_b)^2}. \quad (2)$$

A Equação 1, efetua uma média ponderada, onde a intensidade do *pixel* mais próximo do ponto *P* tem um peso maior. Isto é perfeitamente razoável visto que em uma "pequena" vizinhança, geralmente, quanto mais próximo um ponto está de outro, mais próximo são seus valores de intensidade. O resultado da aplicação desta estratégia de filtragem para a imagem da Figura 7a é apresentado na Figura 7b.

III. Análises e Resultados

O algoritmo proposto foi testado em 260 imagens do acervo de documentos digitalizados de Joaquim Nabuco [7] trazendo resultados de melhor qualidade que o algoritmo em [4].

Do total testado, três resultados são mostrados nas Figuras 8, 9 e 10. As imagens apresentadas representam um mesmo trecho ampliado das imagens originais, e resultantes da filtragem com os algoritmos: anterior e proposto. Pode-se constatar a superioridade da nova técnica de filtragem, pois a detecção da interferência e a qualidade do preenchimento foram melhoradas, isso implica dizer que as duas propostas de aperfeiçoamento atingiram seus objetivos, tornando o aspecto do documento mais natural.

Deve-se relatar que, semelhante ao que ocorreu com o algoritmo anterior, o aqui proposto não teve um desempenho tão bom nas imagens cuja interferência era muito dispersa, ou seja, muito "borrada" (vide Figura 10). Também cabe ressaltar que no acervo observado há poucos casos de imagens com essa característica.

A detecção efetiva de toda interferência se torna uma tarefa complexa, além disso, mesmo se detectando "quase toda interferência" (o que foi conseguido efetuando-se uma dilatação maior na máscara INTERF) a área para preenchimento é grande (pois a interferência é bastante dispersa). Com uma grande área a ser preenchida, a interpolação utilizada não traz um aspecto tão natural para a imagem final.

A Figura 11 ilustra o problema exposto no parágrafo anterior. A primeira imagem contém o mesmo trecho observado na Figura 10, entretanto ele corresponde à imagem filtrada através da aplicação da nova estratégia, fazendo uso da máscara INTERF com uma dilatação maior. Observando-se as Figuras 10 e 11a percebe-se que houve uma melhor filtragem no trecho em questão. No entanto, avaliando-se outro trecho da imagem (Figura 11b), observa-se que este não parece tão natural, dessa forma, constata-se o problema que surge quando se tenta interpolar uma área "relativamente grande".

Para tentar amenizar tal problema, pretende-se utilizar uma interpolação que leve em consideração não apenas os "quatro pontos vizinhos", mas sim todo um "intervalo vizinho", além disso, a consideração de estatísticas da distribuição dos pixels do papel também pode contribuir para um melhor preenchimento. É esperado que o uso dessa nova interpolação faça com que o documento tenha um aspecto mais natural após a filtragem.

IV. Conclusões e Trabalhos Futuros

Neste artigo é proposta uma nova estratégia de filtragem da interferência frente-verso em imagens de documentos coloridos, que apresenta melhores resultados que os algoritmos anteriormente descritos na literatura. Esse sistema utiliza o algoritmo de segmentação proposto em [5] para discriminar os *pixels* provenientes da interferência. Após a discriminação, a área interferente é interpolada de forma a tornar as cores dos seus *pixels* mais próximas das do papel.

Um dos passos que precede o processo de interpolação é a dilatação das máscaras TEXTO e INTERF. Para as 260 filtragens realizadas, foi utilizada uma mesma dilatação. Visto que há imagens com interferência mais (ou menos) borrada, seria mais eficiente o uso de uma dilatação específica para cada documento. Esta possível melhoria está sendo estudada, e se baseia na tentativa do dimensionamento do grau de dispersão da interferência. Tal grau pode ser obtido através da observação do gradiente entre a interferência e o papel.

Outro aspecto ainda não mencionado é o fato do aparecimento de componentes de altas freqüências na imagem resultante da filtragem. Isso ocorre devido à "quebra inercial" da variação das intensidades que havia na imagem original. Esse fato também contribui para um aspecto menos natural da imagem final. Para se ter uma melhoria neste sentido, pode-se verificar a freqüência máxima que aparece no documento original, e com esta, fazer uso de um filtro passa-baixas para filtrar a imagem final. Isso suavizará, por exemplo, transições entre áreas interpoladas e áreas de texto (não alteradas), trazendo um aspecto mais "natural" para o documento final.

AGRADECIMENTOS

Ao CNPq (Conselho Nacional de Pesquisas e Desenvolvimento Tecnológico) e à CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) pelo suporte financeiro. 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. Kasturi, L. O´Gorman and V. Govindaraju, "Document image analysis: A primer", *Sadhana*, (27):3-22, 2002.

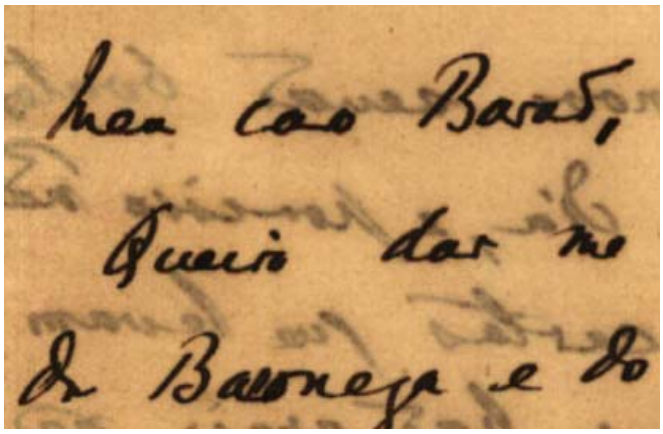
[3] G.Sharma, "Show-through cancellation in scans of duplex printed documents", *IEEE Trans. Image Processing*, v10(5):736-754, 2001.

[4] J. M. M. da Silva e R. D. Lins. "Um Novo Método de Filtragem de Interferência Frente-Verso em Documentos Coloridos", *XXV SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES - SBrT 2007*, Recife, Brasil, 2007.

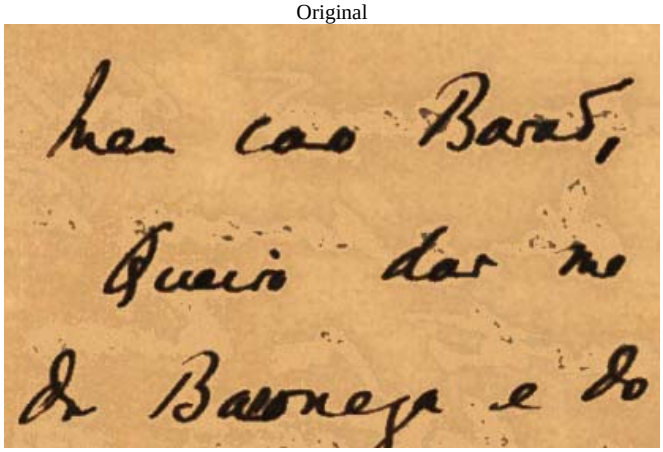
[5] J. M. M. da Silva; R. D. Lins; F. M. J. Martins; R. Wachenchauzer. "A New and Efficient Algorithm to Binarize Document Images Removing Back-to-Front Interference". *Journal of Universal Computer Science*, v. 14, p. 299-313, 2008.

[6] N. Abramson, "Information Theory and Coding", McGraw-Hill Book Co, 1963.

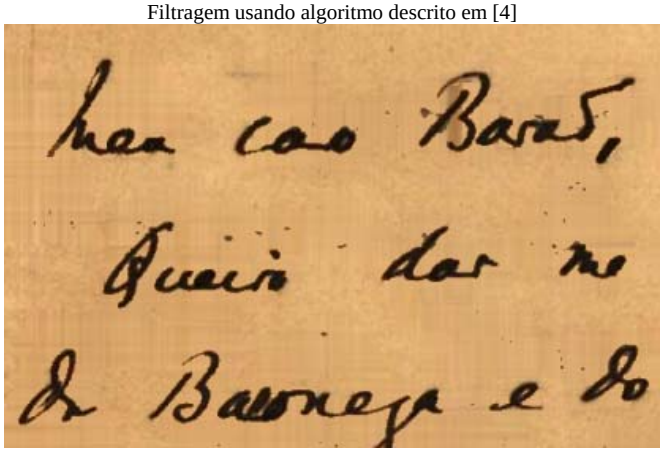
[7] FUNDAJ: www.fundaj.gov.br



Original

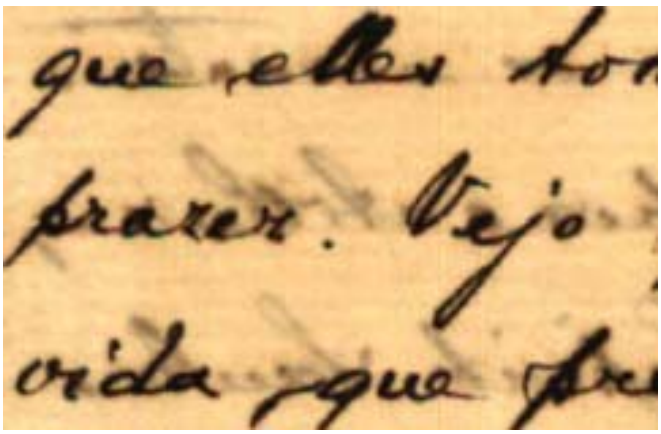


Filtragem usando algoritmo descrito em [4]

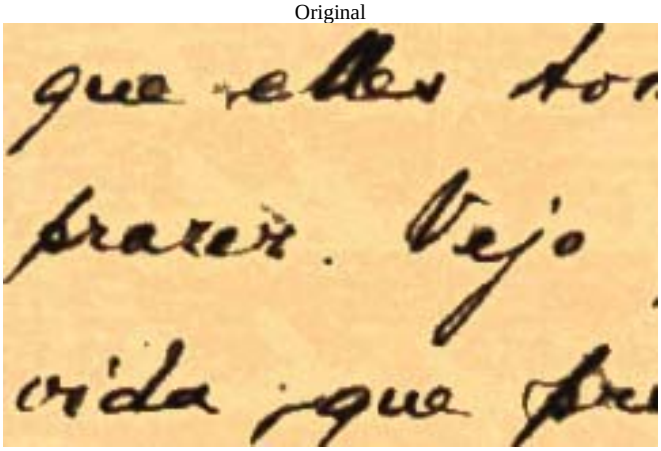


Filtragem usando estratégia proposta

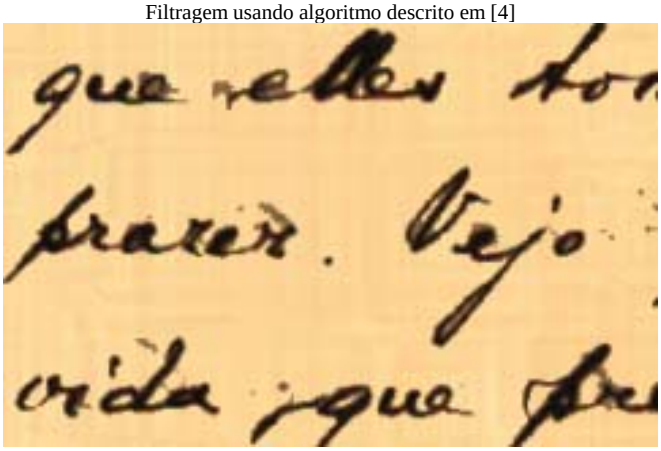
Fig. 8. Zoom em parte de documento do acervo de Nabuco com interferência frente-verso.



Original

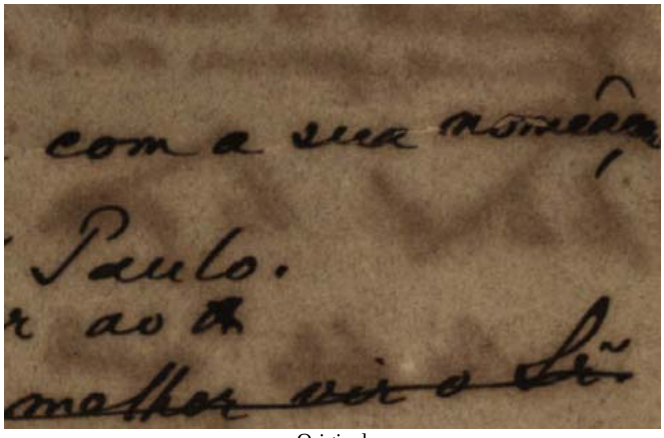


Filtragem usando algoritmo descrito em [4]

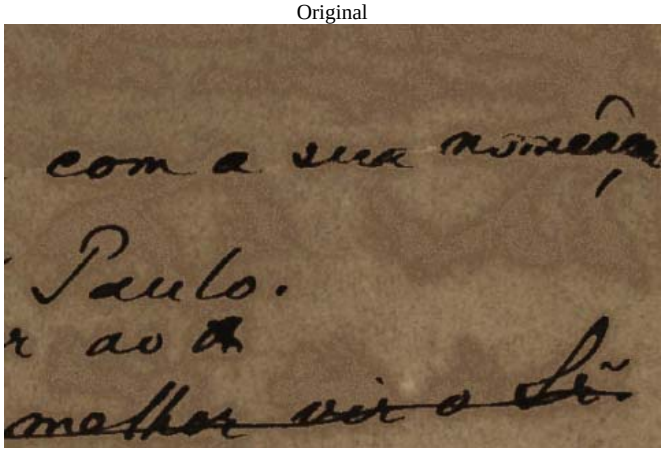


Filtragem usando estratégia proposta

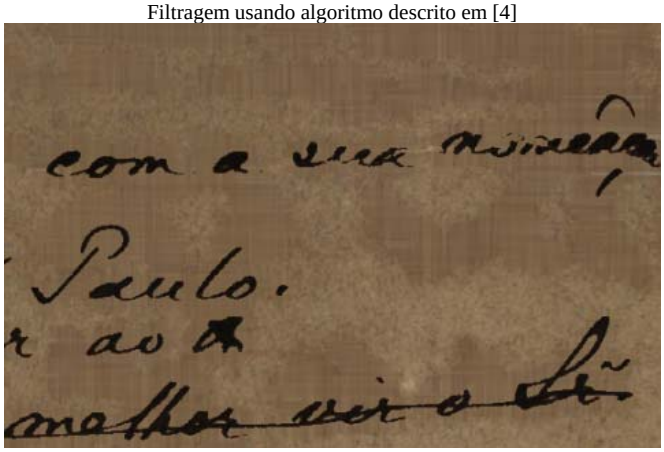
Fig. 9. Zoom em parte de documento do acervo de Nabuco com interferência frente-verso.



Original

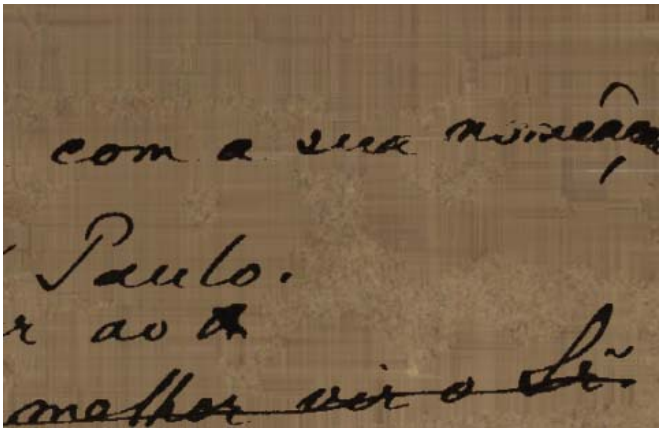


Filtragem usando algoritmo descrito em [4]

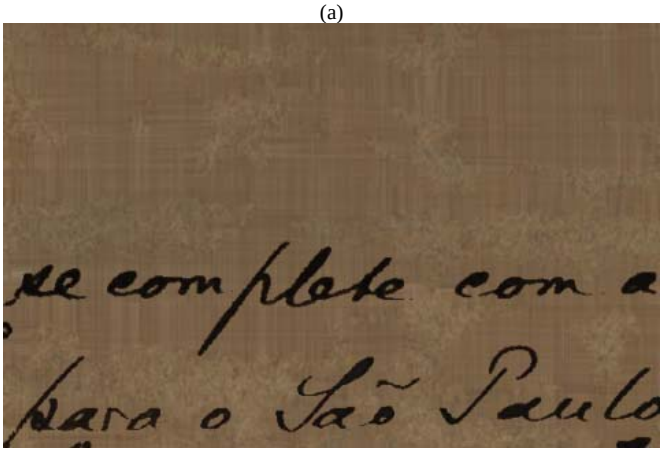


Filtragem usando estratégia proposta

Fig. 10. Zoom em parte de documento do acervo de Nabuco com interferência frente-verso.



(a)



(b)

Fig. 11. (a) Mesmo trecho da Figura 11, correspondente à imagem filtrada com o novo algoritmo, fazendo uso da máscara INTERF com uma dilatação maior. (b) Outro trecho ampliado da mesma imagem.