

Um Método Quantitativo de Avaliação para Filtragem da Interferência Frente-Verso em Documentos Coloridos

João Marcelo Monte da Silva, Rafael Dueire Lins e Ismael Gomes Netto

Resumo— Sempre que se tem um documento escrito em ambos os lados de papel translúcido diz-se que há uma interferência frente-verso. A literatura técnica descreve vários algoritmos para filtragem da interferência frente-verso em documentos. Este artigo apresenta um novo método quantitativo para avaliar tais algoritmos. Este método é baseado na síntese de imagens com tal interferência e em uma posterior comparação entre as imagens filtradas e uma imagem de referência fazendo uso do PSNR.

Palavras-Chave— Análise de Algoritmos, Interferência frente-verso, Documentos Históricos.

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Keywords— Algorithm Assessment, Back-to-front interference, Bleeding, Show-through, Historical Documents.

I. INTRODUÇÃO

Quando um documento é escrito ou impresso em ambos os lados de um papel cuja opacidade permite que a impressão do verso seja vista na frente há uma degradação da legibilidade desse documento, dificultando sua transcrição e binarização. Esse fenômeno, primeiramente citado na literatura técnica por Lins em 1994 [1], recebeu o nome de “interferência frente-verso” (do inglês *back-to-front interference*), mais adiante, outros autores também o chamaram de *bleeding* e *show-through*. Em documentos históricos há vários casos em que a tinta, devido a sua alta fluidez, permeia o papel, intensificando ainda mais a visibilidade do conteúdo impresso do verso na face frontal do documento. A carta de Joaquim Nabuco mostrada na Figura 1, pertencente ao acervo da Fundação Joaquim Nabuco – FUNDAJ [2], exemplifica tal interferência.

Na literatura há vários algoritmos que foram projetados para filtrar a interferência frente-verso [3], [4], [5], [6], [7], [8], [9]. Este artigo descreve um método quantitativo para a avaliação de tais algoritmos quando aplicados em imagens de documentos coloridos.

João Marcelo Monte da Silva, 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: joaommsilva@gmail.com, rdl@ufpe.br, ismaelgomesn@gmail.com.

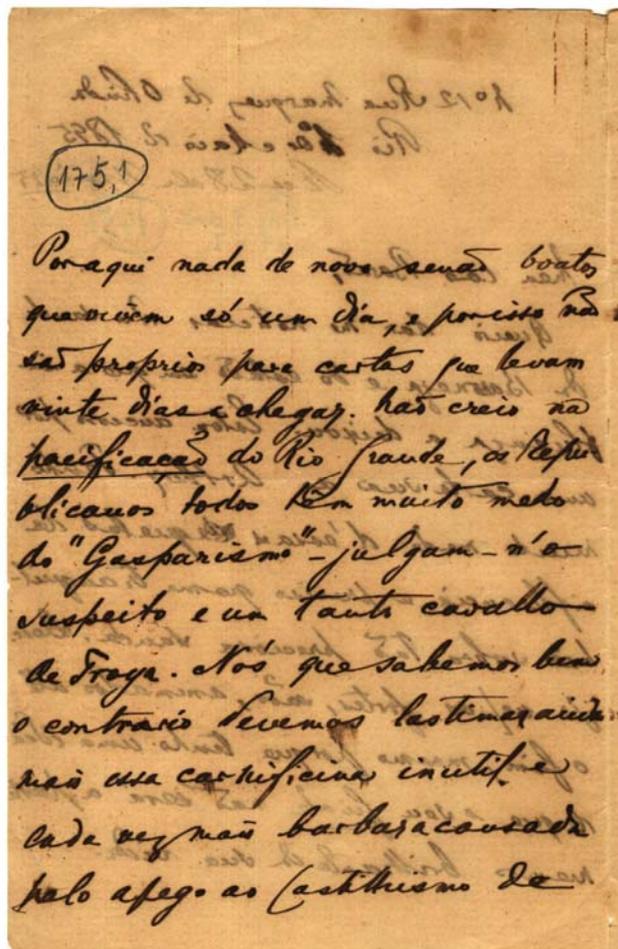


Fig. 1. Imagem de um documento histórico com interferência frente-verso.

O processo aqui desenvolvido é uma “generalização” do apresentado em [10] que lida com imagens em níveis de cinza. A partir do método aqui descrito, é possível comparar os algoritmos de filtragem de uma forma justa.

Para demonstrar o método quantitativo que será descrito na seção II, far-se-á uso de três algoritmos. O primeiro é um algoritmo que será introduzido neste trabalho, este faz uso do conceito de filtragem em espelho [1] e será descrito na seção III. Os outros dois são algoritmos de filtragem por limiar, sendo um com seleção manual de pontos [3] e o outro automático [4]. Os resultados e as análises serão discutidos na seção IV. Finalmente, na seção V, conclusões e algumas linhas para trabalhos futuros.

II. DESCRIÇÃO DO MÉTODO DE AVALIAÇÃO QUANTITATIVO

O método aqui descrito é dividido em dois passos:

1. síntese de imagens com interferência frente-verso – baseado em duas imagens sem interferência; e
2. comparação das imagens filtradas através dos algoritmos com uma imagem de referência – fazendo uso do PSNR (*Peak Signal to Noise Ratio*) para cada componente de cor no espaço RGB.

O PSNR foi escolhido por ter uma baixa complexidade computacional frente a parâmetros que definem a qualidade da imagem no sentido perceptual [11].

A. Síntese de Imagens com Interferência Frente-Verso

Esta etapa tem início tomando-se duas imagens de documentos coloridos sem interferência frente-verso, como mostra a Figura 2: F – representa a frente do documento; e V – o verso do mesmo.

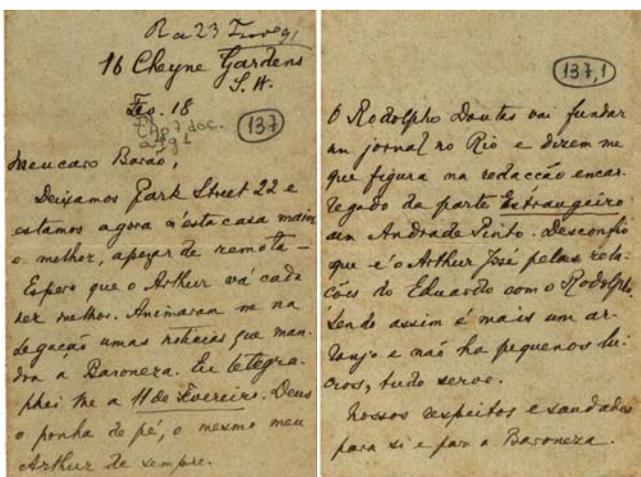


Imagem Frente – F Imagem Verso – V

Fig. 2. Imagens originais de dois documentos sem interferência frente-verso.

Através da composição das imagens F e V, fazendo-se uso do canal alfa, uma terceira imagem FV_α é gerada (vide Figura 3).

As cores dos *pixels* dessa nova imagem são determinadas por:

$$\begin{cases} FV_\alpha^{(R)} = \alpha F^{(R)} + (1-\alpha) V^{(R)} \\ FV_\alpha^{(G)} = \alpha F^{(G)} + (1-\alpha) V^{(G)} \\ FV_\alpha^{(B)} = \alpha F^{(B)} + (1-\alpha) V^{(B)} \end{cases} \quad (1)$$

onde os sobrescritos (R), (G) e (B) denotam as componentes de vermelho, verde e azul das imagens FV_α , F e V, e α é o coeficiente de opacidade da imagem F (frente) que varia de 0 a 100%. Ainda, o fator $(1-\alpha)$ indica o coeficiente de transparência da imagem F.

Para se ter uma melhor visualização do que está ocorrendo, imagine a imagem F como um vidro colorido. A cor de cada ponto é a cor do vidro que tem opacidade α mais a cor que está atrás reduzida pelo fator $(1-\alpha)$ que representa a transparência do vidro.

Para finalizar, faz-se uma operação de “mais escuro” entre as imagens F e FV_α , gerando-se a imagem final S_α (vide Figura 4). Essa operação é feita *pixel a pixel*, comparando-se as luminâncias dos *pixels* correspondentes das duas imagens. A seguir, temos um detalhamento dessa operação.

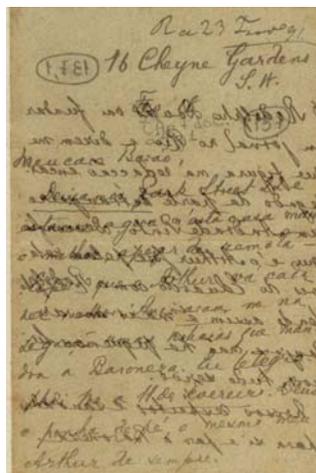


Imagem FV_α

Fig. 3. Composição, usando canal alfa das imagens F e V.

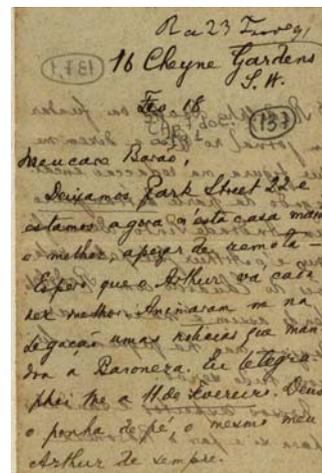


Imagem S_α

Fig. 4. Resultado final da síntese.

- Sejam $f^{(Y)}(i,j)$ e $fV_\alpha^{(Y)}(i,j)$ os valores de luminância dos *pixels* das imagens F e FV_α , respectivamente, na posição (i,j) determinados por:

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

onde $f^{(R)}(i,j)$, $f^{(G)}(i,j)$, $f^{(B)}(i,j)$ e $fV_\alpha^{(R)}(i,j)$, $fV_\alpha^{(G)}(i,j)$, $fV_\alpha^{(B)}(i,j)$ são as componentes (vermelho, verde, azul) dos *pixels* das imagens F e FV_α , respectivamente, na posição (i,j) .

- Se $f^{(Y)}(i,j) \leq fV_\alpha^{(Y)}(i,j)$, ou seja, se o *pixel* proveniente de F tiver uma luminância mais escura ou igual a do *pixel* vindo de FV_α , então o *pixel* da imagem final S_α terá a cor do *pixel* proveniente de FV_α . Caso contrário, ele terá a cor do *pixel* proveniente de F. Isto é mostrado na Equação 3.

$$\begin{cases} \text{se } f^{(Y)}(i,j) \leq fV_\alpha^{(Y)}(i,j) \text{ então } \rightarrow \begin{cases} s_\alpha^{(R)}(i,j) = f^{(R)}(i,j) \\ s_\alpha^{(G)}(i,j) = f^{(G)}(i,j) \\ s_\alpha^{(B)}(i,j) = f^{(B)}(i,j) \end{cases} \\ \text{se } f^{(Y)}(i,j) > fV_\alpha^{(Y)}(i,j) \text{ então } \rightarrow \begin{cases} s_\alpha^{(R)}(i,j) = fV_\alpha^{(R)}(i,j) \\ s_\alpha^{(G)}(i,j) = fV_\alpha^{(G)}(i,j) \\ s_\alpha^{(B)}(i,j) = fV_\alpha^{(B)}(i,j) \end{cases} \end{cases} \quad (3)$$

onde $s_\alpha^{(R)}(i,j)$, $s_\alpha^{(G)}(i,j)$ e $s_\alpha^{(B)}(i,j)$ são as componentes vermelho, verde e azul, respectivamente, do *pixel* da imagem S_α na posição (i,j) .

Na Figura 5 pode ser visto um diagrama de bloco do processo de síntese. A imagem indicada por V é a imagem V da Figura 2 espelhada.

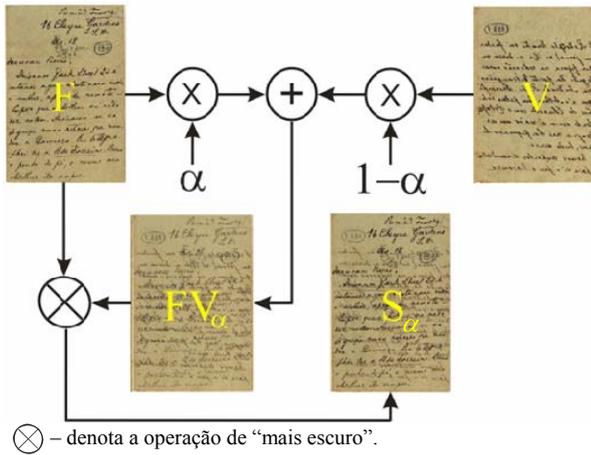


Fig. 5. Diagrama de bloco da etapa de síntese.

Para avaliar a capacidade de filtragem dos algoritmos variamos o fator de opacidade α de 0 a 100% ($\alpha = 0, 1, \dots, 100\%$). O efeito da variação de α na imagem final S_α gerada pelos documentos mostrados na Figura 2 com V espelhada é apresentado na Figura 6.


 Fig. 6. Partes de imagens sintetizadas S_α com diferentes valores de α .

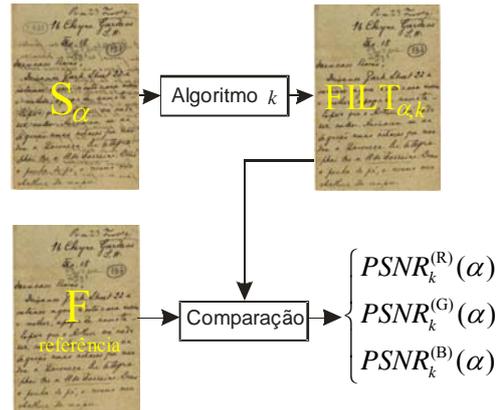
B. Cálculo dos PSNR's

Após o processo de síntese, as imagens sintetizadas são filtradas pelos algoritmos a serem avaliados. Os resultados das filtragens são comparados com a imagem F (vide Figura 2). O PSNR (*Peak Signal-to-Noise Ratio*) será usado nesta etapa para medir a qualidade das imagens resultantes do processo de filtragem. Como se está trabalhando com imagens em *true-color* o PSNR será computado para cada componente de cor RGB como segue

$$PSNR_k^{(cor)}(\alpha) = 20 \log_{10} \frac{255}{\sqrt{\frac{\sum_{i=1}^A \sum_{j=1}^L [f^{(cor)}(i,j) - filt_{\alpha,k}^{(cor)}(i,j)]^2}{A \times L}}}, \quad (4)$$

onde A é a altura da imagem, L é a largura da imagem, $f^{(cor)}(i,j)$ é a intensidade da componente cor na posição (i,j) da imagem F (referência) e $filt_{\alpha,k}^{(cor)}(i,j)$ é a intensidade da componente cor na posição (i,j) da imagem $FILT_{\alpha,k}$ que é o resultado da filtragem da imagem S_α pelo algoritmo k .

A seguir, na Figura 7, é apresentado um diagrama de bloco que ilustra o processo de comparação entre a imagem de referência F e a filtrada $FILT_{\alpha,k}$.


 Fig. 7. Diagrama de bloco que ilustra o processo de comparação entre a imagem de referência F e a filtrada $FILT_{\alpha,k}$.

III. ALGORITMO BASEADO EM FILTRAGEM EM ESPELHO

Salienta-se que o algoritmo aqui descrito foi concebido apenas para a demonstração do método.

Este algoritmo além de usar a imagem a ser filtrada, ele necessita da imagem do verso do documento para inferir quais *pixels* pertencem à interferência. Dessa forma, só é indicada sua utilização quando se têm as imagens da frente e do verso do documento a ser filtrado.

O algoritmo funciona da seguinte forma:

1. tomam-se as duas faces de um documento, a frente e o verso refletido (vide Figuras 8a e 8b);
2. alinham-se, o máximo possível, as duas imagens de forma que a interferência constante na imagem da frente fique na mesma posição que o texto da imagem do verso refletido;
3. comparam-se, *pixel a pixel*, as imagens em níveis de cinza da frente e do verso refletido, destacando-se como interferência os *pixels* cujos níveis de cinza são mais baixos na segunda do que na primeira, ou seja, se um dado *pixel* da imagem do verso refletido é mais escuro do que seu correspondente na imagem da frente ele é classificado como interferência (Vide Figura 8c);
4. preenche-se a área destacada como interferência com *pixels* cujas cores são próximas da moda do documento (vide Figura 8d).

Deve-se salientar que o segundo passo acima descrito não é uma tarefa trivial, pois um detalhe pode fazer com que as imagens nunca tenham um bom alinhamento, por exemplo, dobraduras e amassados no documento original quando digitalizadas impossibilitam o perfeito alinhamento das duas imagens.

O preenchimento da área destacada (quarto passo do algoritmo) é feito da seguinte forma:

- i. toma-se uma grande área do documento, tendo o cuidado de não incluir sua borda;
- ii. observa-se, na área tomada, a luminância mais frequente (y_{moda}), a luminância de maior valor (y_{max}) e a diferença entre elas $d = y_{max} - y_{moda}$ (vide Figura 9);
- iii. calcula-se o valor do limiar $T = y_{moda} - d$;
- iv. 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 ;

- v. preenche-se, finalmente, a área interferente escolhendo-se aleatoriamente *pixels* do vetor *amostra_do_papel*.

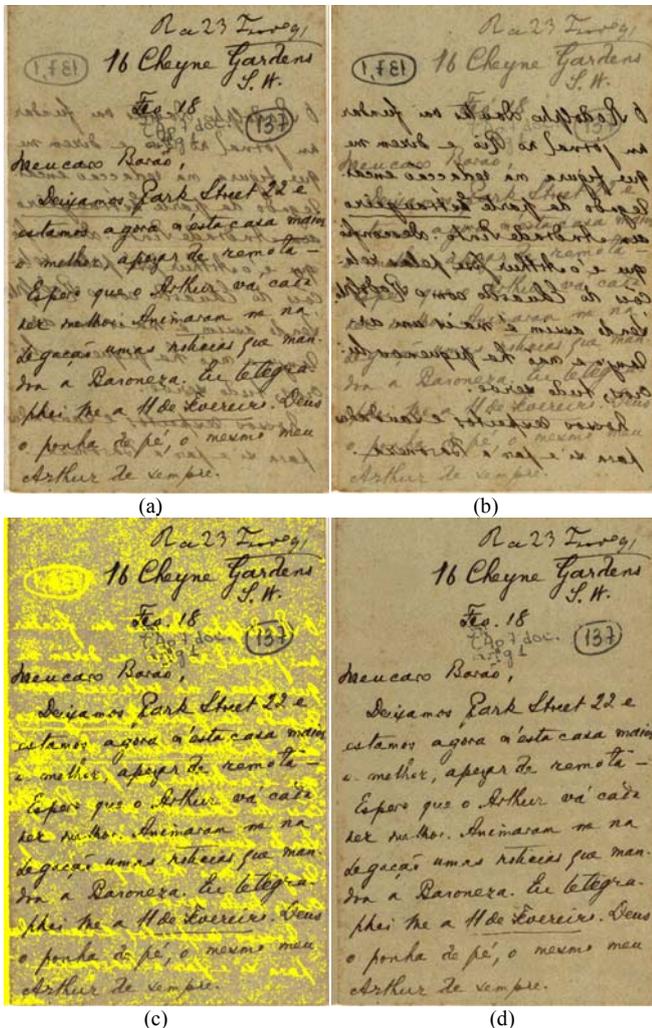


Fig. 8. (a) Face da frente e (b) do verso refletida de um documento sintetizado com interferência frente-verso. (c) Imagem da Figura 8a com os pixels da interferência destacados de amarelo e (d) resultado final da filtragem.

O passo iii. acima descrito é justificado pelo fato da distribuição dos níveis de cinza do papel terem uma certa simetria em relação a sua moda (vide Figura 9).

A forma de filtragem em espelho foi primeiramente sugerida por Lins [1]. Este conceito foi utilizado nos algoritmos propostos por Sharma [5] e por Cao, Tan e Shen [6].

As imagens mostradas na Figura 8 são imagens sintetizadas pelo processo descrito na seção anterior. Tais imagens foram escolhidas para a demonstração deste algoritmo por estarem “naturalmente” alinhadas, devido ao processo de síntese utilizado, pois como já mencionado o alinhamento não é uma tarefa trivial.

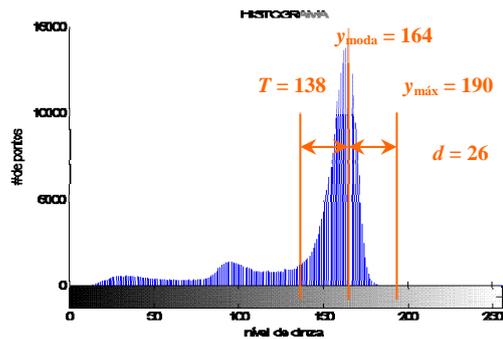


Fig. 9. Histograma da janela referente à versão em níveis de cinza da imagem da Figura 8a.

IV. ANÁLISES E RESULTADOS

Foram escolhidos três algoritmos para demonstração do método de avaliação quantitativo descrito na seção II.

- Algoritmo 1: filtragem em espelho – faz uso da imagem do verso do documento para classificar os *pixels* da interferência (este algoritmo está descrito na seção III);
- Algoritmo 2: filtragem por limiar manual [3] – baseado na seleção manual de alguns *pixels* da imagem original para encontrar a interferência;
- Algoritmo 3: filtragem por limiar automático [4] – utiliza um algoritmo de limiarização para demarcar a área interferente.

Como mencionado na descrição do método foram sintetizadas 101 imagens com intensidades diversas de interferência frente-verso ($\alpha = 0, 1, \dots, 100\%$). Essas imagens foram filtradas pelos algoritmos aqui mencionados. Em seguida, foram calculados os PSNR's para cada componente de cor de cada imagem, como indicado na seção II, e seus resultados podem ser vistos nos gráficos da Figura 10. Tais gráficos trazem no eixo das abscissas o valor da opacidade α utilizado para sintetizar a imagem S_α correspondente, e no eixo das ordenadas os valores dos PSNR's das componentes de cor RGB.

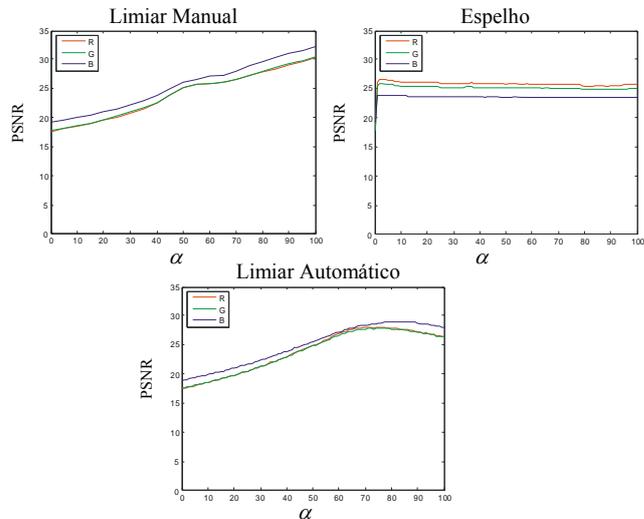


Fig. 10. Gráficos dos PSNR's das imagens filtradas $FILT_{\alpha k}$ em relação à imagem F.

É possível observar através dos gráficos apresentados na Figura 10 que o algoritmo de filtragem em espelho é o mais estável, e, em linhas gerais, tem o melhor desempenho. Este método é bastante eficiente na separação da interferência do resto do documento. Contudo, além do algoritmo necessitar da imagem do verso do documento para fazer a discriminação dos *pixels* interferentes, há a dificuldade no alinhamento das imagens, uma tarefa não trivial. Seu alto desempenho é justificado pelo fato das imagens já estarem alinhadas devido à forma com que foi realizado o processo de síntese.

O segundo colocado foi o algoritmo de filtragem por limiar manual. Este por selecionar manualmente alguns *pixels* (tinta da frente, interferência e papel) do documento faz uma melhor discriminação dos *pixels* interferentes do que o método de filtragem por limiar automático. Contudo, o último tem a vantagem de não fazer uso da imagem do verso do documento nem de precisar da intervenção do usuário na escolha de *pixels* do mesmo.

Observa-se, ainda na Figura 10, que o algoritmo de filtragem por limiar manual aumenta seu desempenho quando a interferência vai se tornando mais fraca, ou seja, quando o valor de α cresce. O mesmo não ocorre com o algoritmo de filtragem em espelho, pois mesmo em imagens com baixa interferência este algoritmo classifica como interferência os *pixels* que são mais escuros na imagem do verso. Dessa forma, sendo a interferência quase inexistente, tal algoritmo classifica *pixels* que pertencem ao papel, como interferência. Para melhorar seu desempenho quanto a este fato, pode-se utilizar a diferença δ entre os *pixels* da frente e do verso, e dispor de um limiar T_δ para determinar se o *pixel* pertence ou não à interferência, em outras palavras, se a diferença δ for positiva e menor do que T_δ ($0 < \delta < T_\delta$) o *pixel* em questão será classificado como interferência, caso contrário não será classificado como tal. Assim, em imagens com interferência muito fraca, o algoritmo não classificará como interferência *pixels* que pertencem ao papel, pois a diferença calculada provavelmente será maior que o limiar estabelecido.

Deve-se destacar que o algoritmo de filtragem por limiar automático tem um comportamento semelhante ao algoritmo de filtragem em espelho em relação à classificação dos *pixels* interferentes para interferência fraca ($\alpha > 80$). Sendo a interferência quase inexistente o algoritmo classifica como interferência *pixels* do papel. Entretanto ele tem uma tendência semelhante a do manual para valores de $\alpha < 80$.

Do conjunto de imagens sintetizadas foram selecionadas três amostras. As amostras escolhidas foram tais que representassem interferências existentes em documentos reais do acervo de documentos digitalizados de Joaquim Nabuco [2]. As amostras com $\alpha = 40\%$, $\alpha = 65\%$ e $\alpha = 90\%$ (vide Figura 11) representam imagens com interferências forte, média e fraca, respectivamente.

Na Figura 12 são apresentadas as imagens da Figura 11 filtradas pelos três algoritmos aqui comparados.

Todo este estudo tem a finalidade de indicar qual o algoritmo mais apropriado para atuar em uma determinada faixa de valores de α . Dessa forma, é interessante saber determinar o valor de α em um dado documento.

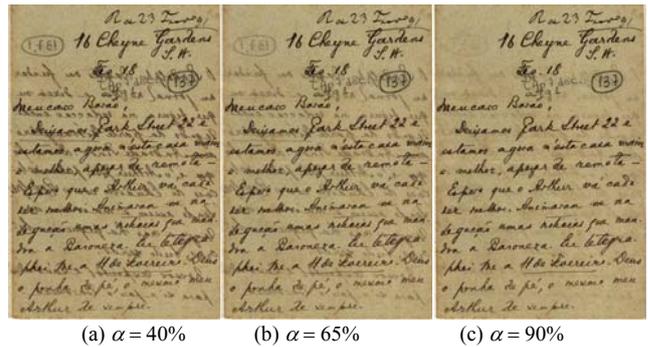


Fig. 11. Imagens sintetizadas (a) $\alpha = 40\%$, (b) $\alpha = 65\%$ e (c) $\alpha = 90\%$.

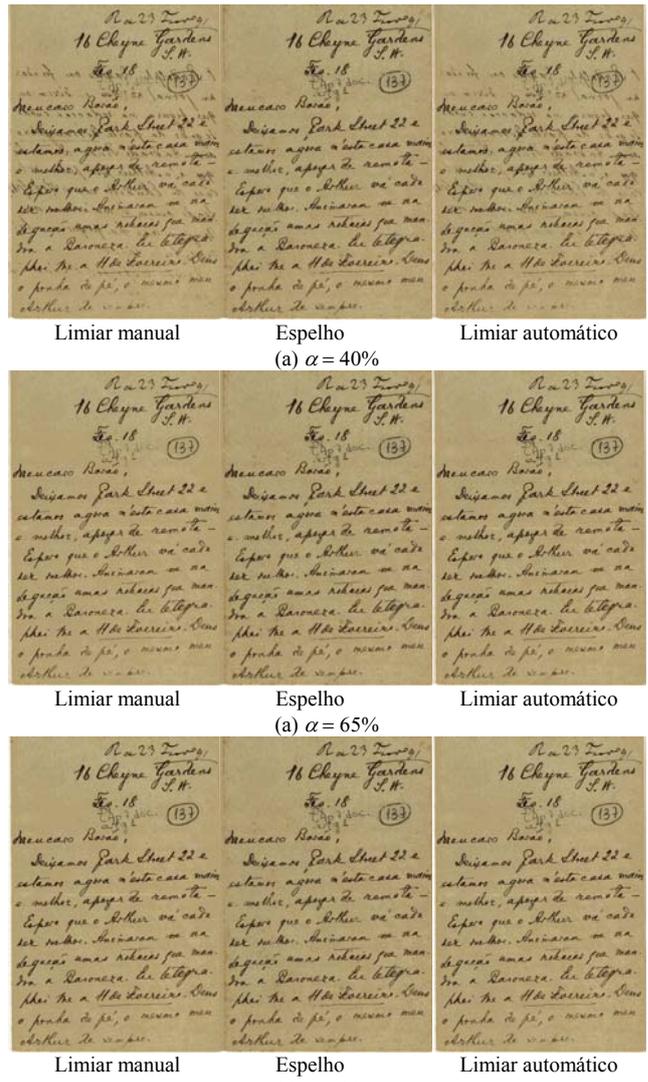


Fig. 12. Imagens das Figuras (a) 11a, (b) 11b e (c) 11c filtradas pelos algoritmos comparados.

É possível estimar a opacidade α de um documento real da seguinte forma:

- pelo modelo apresentado, um *pixel* de interferência tem intensidade luminosa $interf(i, j)$ dado por:

$$interf(i, j) = \alpha \cdot papel(i, j) + (1 - \alpha) \cdot tinta_{verso}(i, j) \quad (5)$$

onde $papel(i, j)$ e $tinta_{verso}(i, j)$ são as intensidades dos *pixels* do papel e da tinta do verso, respectivamente;

- na grande maioria dos casos pode-se assumir que a intensidade da tinta do verso $tinta_{verso}$ é aproximadamente a mesma da tinta da frente $tinta_{frente}$, dessa forma, pode-se coletar amostras da interferência, do papel e da tinta da frente para representarem as intensidades $interf$, $papel$ e $tinta_{frente}$, respectivamente;
- por fim, usando a Equação 5 e assumindo que $tinta_{verso} \approx tinta_{frente}$, podemos estimar o valor de α por:

$$\alpha = \frac{interf - tinta_{frente}}{papel - tinta_{frente}} \quad (6)$$

Para exemplificar, toma-se a imagem da Figura 1. Os valores coletados foram:

$$\begin{cases} tinta_{frente} = 23 \\ interf = 106 \\ papel = 201 \end{cases} \Rightarrow \alpha = \frac{106 - 23}{201 - 23} \approx 0,47 \quad (7)$$

assim, pode-se dizer que tal imagem tem um coeficiente de opacidade $\alpha \approx 47\%$.

Uma maneira automática de se saber qual o valor da interferência pode ser obtida através da binarização direta da imagem colorida. A experiência mostra que cerca de no máximo 8% dos pixels de um documento são traduzidos em pixels pretos. A binarização direta implicará em uma quantidade maior de pixels, proporcional ao coeficiente de opacidade. O relacionamento das faixas com a quantidade de pixels ainda não está estabelecida, ficando para trabalhos futuros.

V. CONCLUSÕES E TRABALHOS FUTUROS

Foi introduzido neste trabalho um método quantitativo para a avaliação de algoritmos que filtram a interferência frente-verso em imagens de documentos coloridos. Esse método permite a realização de um estudo sobre tais algoritmos com a finalidade de saber qual o algoritmo mais indicado para filtrar um determinado documento, tomando como base a opacidade α . Para se automatizar essa escolha é necessário se estimar, aproximadamente, valores das amostras da tinta da frente, do papel, e da tinta interferente para então se calcular a opacidade α . Outra forma é binarizar diretamente a imagem colorida, e através de uma relação pré-estabelecida determinar a opacidade α através da quantidade de pixels pretos da imagem binarizada. Com o valor da opacidade determinado, escolhe-se o melhor algoritmo para a filtragem, tomando como base um estudo

previamente feito.

O PSNR foi utilizado para medir a qualidade da imagem final. Em trabalhos futuros, pretende-se definir medidas que informe a legibilidade da informação útil, a preservação do papel e a qualidade do preenchimento da área interferente de forma independente.

Em muitas imagens a interferência aparece “borrada”. Para se levar em conta esse efeito em um estudo, deve-se passar a imagem que irá representar o verso por um filtro passa baixa antes do processo de síntese. O modelo pode ser ainda mais sofisticado, fazendo uso de um filtro adaptativo que leva em conta dados da imagem da frente.

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] FUNDAJ: www.fundaj.gov.br
- [3] R. D. Lins e I. G. Netto. “Um Ambiente para Filtrar a Interferência Frente-Verso em Documentos Históricos”. *SBrT-2007*, Recife, Brasil, 2007.
- [4] J. M. M. da Silva e R. D. Lins. “Um Novo Método de Filtragem de Interferência Frente-Verso em Documentos Coloridos”. *SBrT-2007*, Recife, Brasil, 2007.
- [5] G. Sharma, “Show-through cancellation in scans of duplex printed documents”, *IEEE Trans. Image Processing*, v10(5):736-754, 2001.
- [6] 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.
- [7] Feng Su and Ali Mohammad-Djafari, “Bayesian Separation of Document Images with Hidden Markov Model”, *2nd International Conference on Computer Vision Theory and Applications*, Barcelona, Spain, 2007.
- [8] B. Ophir and D. Malah, “Improved cross-talk cancellation in scanned images by adaptive decorrelation”, *23rd IEEE Convention of Electrical and Electronics Engineers in Israel*, 2004, pp. 388-391.
- [9] H.Nishida and T.Suzuki, “A Multiscale Approach to Restoring Scanned Color Document Images with Show-through Effects”, Proc. of. ICDAR 2003, 2003.
- [10] R. D. Lins and J. M. M. da Silva. “Assessing Algorithms to Remove Back-to-Front Interference in Documents”, *ITS-2006*, Fortaleza, Brazil, 2006.
- [11] R. F. Zampolo e R. Seara. “Estudo Comparativo entre Métricas para Avaliação da Qualidade de Imagens”. *XXII SIMPÓSIO BRASILEIRO DE TELECOMUNICAÇÕES*, pp. 237-241, Campinas-SP, 2005.