Sobre Adequação da Similaridade Estrutural Média em Codificadores Baseados em DCT

Ronaldo de Freitas Zampolo e Rui Seara

***Resumo*— Este trabalho considera o uso da SSIM em codificadores de imagem baseados na DCT. Mais especificamente, os autores propõem uma abordagem para adaptar a SSIM média (*mean SSIM* – MSSIM) à estrutura pré-existente do codificador. Tal abordagem também resulta na redução da complexidade computacional associada à avaliação da métrica em questão. O desenvolvimento matemático da proposta bem como os resultados evidenciando sua aplicabilidade são apresentados e discutidos.**

***Palavras-Chave*— Avaliação de qualidade visual, codificação perceptual, métricas perceptuais.**

***Abstract*— This paper addresses the use of SSIM metric in DCT-based image coding. Specifically, the authors propose a strategy for adapting the mean SSIM (MSSIM) to the previous coder structure. Such a strategy also reduces computational complexity related to the assessment of the referred metric. Mathematical development of the proposed approach are presented and discussed. Experimental results verify its feasibility.**

***Keywords*— Visual quality assessement, perceptual coding, perceptual metrics.**

I. Introdução

Em função do crescente número de aplicações em vídeo e imagem destinadas ao consumo humano, sistemas com orientação perceptual são atualmente de grande interesse. Justificada pela sua ampla utilização em sistemas de transmissão e armazenamento, a codificação de imagem/vídeo vem estimulando um número considerável de pesquisas em codificadores perceptuais, ou seja, codificadores cujos parâmetros sejam otimizados de acordo com alguma métrica que modele, em determinado sentido, a percepção humana em termos de qualidade visual.

A avaliação de qualidade visual aplicada a sistemas de codificação se beneficia do fato de a imagem original encontrar-se disponível. Nesse caso, métricas de referência completa podem ser adaptadas à estrutura do codificador, o que não ocorre em outras aplicações como, por exemplo, em restauração e realce de imagens, nas quais a avaliação de qualidade em situações práticas deve ser feita através de estratégias de referência parcial ou mesmo sem referência.

Dentre os trabalhos de maior relevância nesta área de pesquisa, as estratégias que abordam sistemas baseados na transformada do cosseno discreta (*discrete cosine transform* – DCT) ocupam lugar de destaque em virtude da ampla utilização dos codificadores JPEG [1]. Em geral, as técnicas propostas buscam uma representação no domínio transformado de alguma métrica perceptual. O mapeamento de limiares de percepção em diferentes freqüências angulares, expressos comumente por uma versão da função de limiar de contraste (*contrast threshold function* – CTF), para a DCT surge naturalmente, pois ambos estão fortemente relacionados. Assim, uma análise combinada entre as amplitudes para cada tom DCT e os correspondentes valores de limiar de percepção é usada na definição da matriz de quantização dos coeficientes DCT para uma dada imagem [2], [3], [4], [5].

Recentemente, foi apresentado um estudo em que a métrica de similaridade estrutural (*structural similarity index*, SSIM) [6], [7] é mapeada para o domínio da DCT e usada para estimar faixas-limite para taxas em codificadores [8].

Este trabalho discute o uso da SSIM em codificadores de imagem baseados na DCT. Mais especificamente, é proposta uma abordagem para adaptar a SSIM média (*mean SSIM* – MSSIM) à estrutura pré-existente do codificador e que também resulta na redução da complexidade computacional associada à avaliação da métrica em questão. O desenvolvimento matemático da proposta bem como os resultados evidenciando sua aplicabilidade são apresentados e discutidos.

Este artigo é organizado como segue. A Seção II discute brevemente codificadores perceptuais baseados em DCT. Na Seção III, é apresentada uma revisão sobre a SSIM e sua versão no domínio da DCT. Na Seção IV, a amostragem do mapa de similaridade é considerada como uma maneira de reduzir a complexidade computacional associada ao cálculo convencional da MSSIM. Resultados que evidenciam a aplicabilidade da técnica proposta são apresentados na Seção V. Finalmente, as conclusões do trabalho são apresentadas na Seção VI.

II. Codificadores Perceptuais Baseados em DCT

A estrutura básica de um codificador de imagem baseado em DCT é mostrada na Fig. 1. Essa estrutura é utilizada em codificadores JPEG, nos quais o sinal de entrada consiste em blocos de dimensão 8×8 não superpostos retirados da imagem a ser codifcada [1].

Apesar da quantização JPEG ter sido concebida a fim de tirar vantagem das características do sistema visual humano, em especial daqueles aspectos relacionados à dependência em freqüência da sensibilidade ao contraste, a adição de métricas de qualidade visual com inspiração perceptual pode melhorar o desempenho desses sistemas. Preferencialmente, tais métricas devem ser integradas aos codificadores sem, ou com um mínimo de, alteração/modificação nas suas estruturas

Ronaldo de Freitas Zampolo, Laboratório de Processamento de Sinais - LaPS, Faculdade de Engenharia da Computação, Universidade Federal do Pará, Belém - PA, Brasil, E-mail: zampolo@ufpa.br. Rui Seara, Laboratório de Circuitos e Processamento de Sinais - LINSE, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis - SC, Brasil, E-mail: seara@linse.ufsc.br.

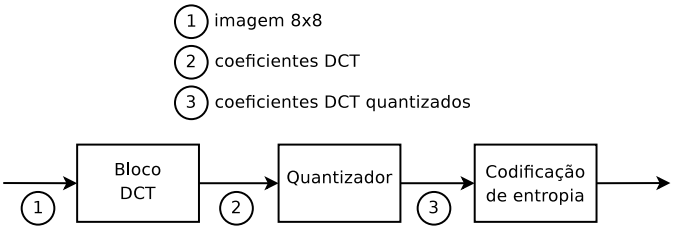


Fig. 1. Codificador baseado em DCT.

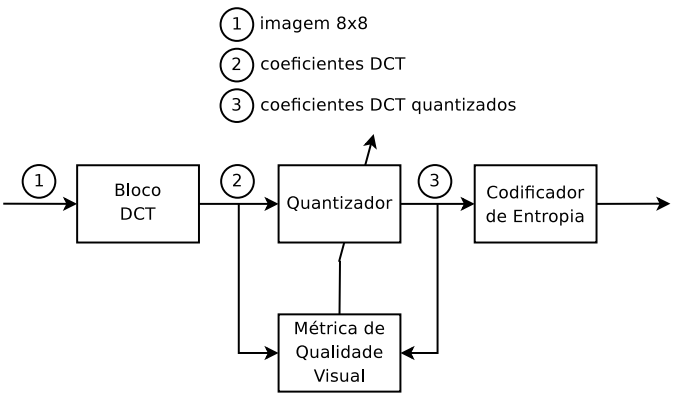


Fig. 2. Codificador perceptual baseado em DCT.

previamente definidas. Uma representação simplificada de um codificador perceptual baseado em DCT é ilustrada na Fig. 2. Nesse codificador, os coeficientes DCT da imagem a ser codificada originais e seus correspondentes quantizados formam as entradas de um procedimento de avaliação de qualidade, cujo resultado, em conjunto com restrições de taxa e/ou níveis de qualidade desejados, orienta o cálculo dos elementos da tabela de quantização.

O codificador perceptual ilustrado na Fig. 2 pressupõe que a métrica de qualidade visual possa ser calculada no domínio da DCT, permitindo assim que a implementação do codificador perceptual não requeira muitas alterações em relação à estrutura convencional. Assim, ao se pensar em utilizar uma das diversas métricas disponíveis para avaliação de qualidade visual, é pertinente verificar se tal métrica possui ou é passível de ter uma versão no domínio transformado.

III. Similaridade Estrutural Média

A métrica adotada neste trabalho, devido à sua simplicidade matemática, baixa complexidade computacional e bons resultados na caracterização da qualidade percebida, é a similaridade estrutural média (*mean structural similarity*, MSSIM) [6], [9].

A MSSIM é calculada a partir do mapa de similaridade estrutural (*similarity map*), obtido ao se determinar o índice de similaridade (*structural similarity index*, SSIM) para cada ponto da imagem-teste em relação a uma imagem de referência. O SSIM entre duas imagens $\mathbf{x}$ e $\mathbf{y}$ é dado pela expressão

$$SSIM(\mathbf{x},\mathbf{y}) = [l(\mathbf{x},\mathbf{y})]^{\alpha}[c(\mathbf{x},\mathbf{y})]^{\beta}[s(\mathbf{x},\mathbf{y})]^{\gamma} \quad (1)$$

onde α, β e γ são índices de ponderação; e $l(\mathbf{x},\mathbf{y})$, $c(\mathbf{x},\mathbf{y})$ e $s(\mathbf{x},\mathbf{y})$ são termos que avaliam, respectivamente, luminância, constraste e a correlação estrutural das imagens consideradas. Estes três últimos termos são definidos por

$$l(\mathbf{x},\mathbf{y}) = \frac{2\mu_x\mu_y + C_1}{\mu_x^2 + \mu_y^2 + C_1} \quad (2)$$

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sigma_x\sigma_y + C_2}{\sigma_x^2 + \sigma_y^2 + C_2} \quad (3)$$

$$s(\mathbf{x},\mathbf{y}) = \frac{\sigma_{xy} + C_3}{\sigma_x\sigma_y + C_3} \quad (4)$$

onde C_1, C_2 e C_3 são constantes usadas para melhorar o condicionamento numérico; μ_x e μ_y representam as médias de $\mathbf{x}$ e $\mathbf{y}$, respectivamente; σ_x^2 e σ_y^2 denotam as variâncias de $\mathbf{x}$ e $\mathbf{y}$, respectivamente; e σ_{xy} é a covariância de $\mathbf{x}$ e $\mathbf{y}$.

A fim de melhor caracterizar o comportamento local das imagens avaliadas, $\mathbf{x}$ e $\mathbf{y}$ representam vizinhanças em torno de um determinado *pixel* da imagem de referência e de seu correspondente na imagem de teste, respectivamente. O procedimento descrito em [6], [7] adota uma janela deslizante para determinação das vizinhanças $\mathbf{x}$ e $\mathbf{y}$. Dessa forma, um valor de SSIM é associado ao *pixel* central da janela que, após percorrer toda a imagem, resulta na atribuição de um valor de SSIM para cada *pixel*. Tal conjunto de SSIMs é chamado de mapa de similaridade.

Por sua vez, a MSSIM é obtida a partir do mapa de similaridade como segue:

$$MSSIM = \frac{1}{MN}\sum_{m,n} SSIM[m,n] \quad (5)$$

onde $SSIM[m,n]$ representa o valor de SSIM associado ao elemento $[m,n]$ do mapa de similaridade; M e N são as dimensões da imagem de referência.

A Fig. 3 ilustra o procedimento de obtenção da MSSIM para uma imagem de teste. Note que a MMSIM é um único número passível de ser representativo da qualidade visual percebida da imagem de teste, considerando uma dada referência.

A. Cálculo da SSIM no Domínio da DCT

Em [8], é desenvolvida uma representação do SSIM no domínio da DCT como parte de uma proposta para estimar faixas de valores para a taxa de bits em codificadores baseados na referida transformada.

A seguir, são apresentadas as versões DCT para os termos $l(\mathbf{x},\mathbf{y})$, $c(\mathbf{x},\mathbf{y})$ e $s(\mathbf{x},\mathbf{y})$ que compõem a expressão (1) para o cálculo do SSIM.

$$l(\mathbf{x},\mathbf{y}) = \frac{2X_{00}Y_{00} + 64C_1}{X_{00}^2 + Y_{00}^2 + 64C_1} \quad (6)$$

$$c(\mathbf{x},\mathbf{y}) = \frac{2\sqrt{\left[\sum_{(u,v)\neq(0,0)} X_{uv}^2\right]\left[\sum_{(u,v)\neq(0,0)} Y_{uv}^2\right]} + 64C_2}{\sum_{(u,v)\neq(0,0)} [X_{uv}^2 + Y_{uv}^2] + 64C_2} \quad (7)$$

$$s(\mathbf{x},\mathbf{y}) = \frac{\sum_{(u,v)\neq(0,0)} X_{uv}Y_{uv} + 64C_3}{\sqrt{\left[\sum_{(u,v)\neq(0,0)} X_{uv}^2\right]\left[\sum_{(u,v)\neq(0,0)} Y_{uv}^2\right]} + 64C_3} \quad (8)$$

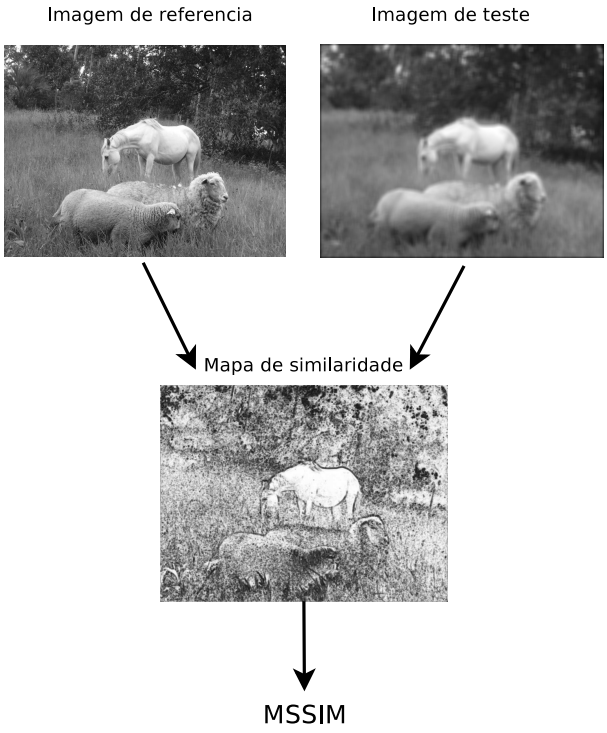


Fig. 3. Obtenção da MSSIM para uma imagem de teste.

onde u e v denotam as coodenadas no domínio transformado; e X_{uv} e Y_{uv} representam as transformadas DCT de $\mathbf{x}$ e $\mathbf{y}$, respectivamente.

Até aqui nada há de diferente do que é apresentado em [8], exceto que as expressões (6)–(8) estão em um formato mais geral, permitindo valores de α, β e γ diferentes do *default* ($\alpha = \beta = \gamma = 1$).

IV. Amostragem do Mapa de Similaridade

As expressões (6)–(8) mostram que é possível calcular o SSIM e, conseqüentemente, a MSSIM tanto no domínio da seqüência quanto no da DCT. Contudo, pensando na utilização da MSSIM para a implementação de codificadores perceptuais baseados em DCT, deve-de notar que a MSSIM considera uma janela deslizante, enquanto os blocos 8×8 da entrada do codificador são obtidos sem superposição da imagem a ser codificada, isto é, usando uma janela saltitante.

Dentre as alternativas possíveis na tentativa de adaptar o cálculo da MSSIM ao codificador DCT, foi escolhido o caminho que causa menor impacto na estrutura do codificador: abandonar a janela deslizante da abordagem original do MSSIM e adotar a janela saltitante que divide a imagem em blocos 8×8 sem superposição. As conseqüências de tal escolha são discutidas a seguir.

Primeiramente, usar a janela 8×8 sem superposição resulta em um conjunto de SSIMs que é uma versão dizimada do mapa de similaridade obtido com a janela deslizante. Mais especificamente, de cada 64 elementos, correspondentes a uma região de dimensão 8×8 do mapa de similaridade, apenas um é preservado (Fig. 4).

A MSSIM (5) pode ser rescrita como

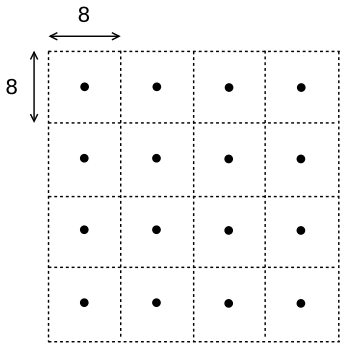


Fig. 4. Resultado devido ao uso de janela 8×8 sem superposição. De um conjunto de 64 elementos do mapa de similaridade, apenas um elemento (círculo preto) é preservado.

$$MSSIM = \frac{1}{W}\sum_{w}\frac{1}{64}\sum_{\substack{0\leq m\leq 7\\ 0\leq n\leq 7}} SSIM_w[m,n] \qquad (9)$$

onde w faz referência a uma dada região 8×8 do mapa de similaridade e W corresponde ao número total dessas regiões. A variável $SSIM_w[m,n]$ caracteriza os valores de similaridade de uma região w específica.

Com o uso da referida janela 8×8 sem superposição e a conseqüente dizimação do mapa de similaridade, pode-se definir um outro MSSIM dado por

$$MSSIM' = \frac{1}{W}\sum_{w} SSIM_w \qquad (10)$$

onde $SSIM_w$ é o único valor de $SSIM$ que restou após o processo de dizimação mencionado.

A diferença entre (9) e (10) avalia o impacto do abandono da janela deslizante. Tal diferença é dada por

$$MSSIM - MSSIM' = \frac{1}{W}\sum_{w} e[w] \qquad (11)$$

onde $e[w]$ é o erro entre a similaridade média da vizinhança w e o $SSIM_w$, expresso como segue:

$$e[w] = \frac{1}{64}\sum_{\substack{0\leq m\leq 7\\ 0\leq n\leq 7}} SSIM_w[m,n] - SSIM_w. \qquad (12)$$

V. Resultados

Os resultados apresentados nesta seção são obtidos a partir de um conjunto de 10 imagens de referência consideradas clássicas em processamento de sinais (Fig. 5). Tais imagens, neste trabalho, foram obtidas da base de dados IVC disponível para *download* em [10]. São produzidas 11 versões das imagens de referência, usando quatro tipos de degradação (má focalização, ruído impulsivo, *sal e pimenta*, e ruído gaussiano branco aditivo) em diferentes níveis, totalizando 395 imagens no conjunto de teste[1] (incluindo as imagens de referência). A Fig. 6 ilustra os tipos de degradação utilizados para a imagem de referência *Clown*.

[1]Não foram produzidas versões degradadas por adição de ruído gaussiano aditivo para as 10 imagens de referência, mas apenas para 5.

(a) Plane

(b) Barbara

(c) Boats

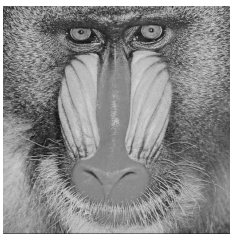
(d) Clown

(e) Fruits

(f) House

(g) Isabel

(h) Lena



(i) Mandril

(j) Peppers

Fig. 5. Imagens de referência a partir das quais o conjunto de teste foi gerado.

Para cada uma das imagens degradadas, é determinado o erro $e[w]$ (12).

A Tabela I mostra o comportamento médio do erro $e[w]$ para os diferentes tipos de degradação e imagens de referência. Os valores da Tabela I são obtidos pela expressão

$$\mu_e = \frac{1}{11} \sum_i \frac{1}{W} \sum_w e_i[w] \tag{13}$$

onde $e_i[w]$ denota o erro $e[w]$ para uma dada imagem de teste i, considerando um tipo de degradação e uma imagem de referência específicos.

(a) Má focalização

(b) Ruído impulsivo

(c) Ruído *sal e pimenta*

(d) Ruído gaussiano branco aditivo

Fig. 6. Exemplos dos tipos de degradação utilizados.

TABELA I

VALORES DE μ_e PARA DIFERENTES IMAGENS DE REFERÊNCIA E TIPOS DE DEGRADAÇÃO (D1: MÁ FOCALIZAÇÃO; D2: RUÍDO IMPULSIVO; D3: "SAL E PIMENTA"; D4: RUÍDO GAUSSIANO BRANCO ADITIVO)

	D1	**D2**	**D3**	**D4**
Plane	$-1{,}03$ E-3	$-4,34$ E-3	$-4,71$ E-3	–
Barbara	$-2{,}16$ E-4	$-2,71$ E-3	$-2,20$ E-3	$-9,19$ E-4
Boats	$6{,}78$ E-6	$-6,70$ E-3	$-5,53$ E-3	–
Clown	$8{,}30$ E-4	$-1,50$ E-3	$-1,77$ E-3	$-5,36$ E-4
Fruits	$1{,}03$ E-3	$-3,44$ E-4	$-5,60$ E-4	$-3,72$ E-4
House	$-4{,}58$ E-4	$-5,97$ E-4	$-2,70$ E-3	–
Isabel	$3{,}85$ E-4	$-4,26$ E-3	$-5,03$ E-3	$-1,86$ E-3
Lena	$8{,}85$ E-4	$-4,61$ E-3	$-4,94$ E-3	–
Mandril	$7{,}66$ E-5	$-1,96$ E-3	$-2,34$ E-3	$-7,14$ E-4
Peppers	$-1{,}63$ E-3	$-2,36$ E-3	$-4,27$ E-3	–

Por sua vez, a Tabela II mostra o comportamento médio da variância do erro $e[w]$ para os diferentes tipos de degradação e imagens de referência. Os valores da Tabela II são obtidos pela expressão

$$\bar{\sigma}_e^2 = \frac{1}{11} \sum_i \frac{1}{W-1} \sum_w \{e_i[w] - \mu_e\}^2 . \tag{14}$$

A Fig. 7 mostra uma região 8×8 de um mapa de similaridade, cujo comportamento pode ser considerado típico dos resultados obtidos. Devido à dizimação ocasionada pela não adoção da janela deslizante, apenas o elemento em negrito

TABELA II

VALORES DE $\bar{\sigma}_e^2$ PARA DIFERENTES IMAGENS DE REFERÊNCIA E TIPOS DE DEGRADAÇÃO (D1: MÁ FOCALIZAÇÃO; D2: RUÍDO IMPULSIVO; D3: "SAL E PIMENTA"; D4: RUÍDO GAUSSIANO BRANCO ADITIVO)

	D1	**D2**	**D3**	**D4**
Plane	$2,83$ E-3	$1,03$ E-2	$1,91$ E-2	–
Barbara	$2,34$ E-3	$1,31$ E-2	$1,56$ E-2	$7,12$ E-4
Boats	$3,68$ E-3	$1,25$ E-2	$1,78$ E-2	–
Clown	$2,88$ E-3	$1,45$ E-2	$1,80$ E-2	$8,16$ E-4
Fruits	$2,56$ E-3	$1,39$ E-2	$1,96$ E-2	$9,22$ E-4
House	$2,87$ E-3	$1,08$ E-2	$1,71$ E-2	–
Isabel	$2,51$ E-3	$1,33$ E-2	$1,87$ E-2	$9,70$ E-4
Lena	$1,93$ E-3	$1,48$ E-2	$1,85$ E-2	–
Mandril	$2,02$ E-3	$9,51$ E-3	$1,20$ E-2	$5,99$ E-4
Peppers	$2,67$ E-3	$1,47$ E-2	$1,95$ E-2	–

$$\begin{bmatrix} 0,99 & 0,99 & 0,98 & 0,98 & 0,98 & 0,98 & 0,98 & 0,97 \\ 0,99 & 0,99 & 0,98 & 0,98 & 0,98 & 0,98 & 0,98 & 0,97 \\ 0,99 & 0,99 & 0,98 & 0,98 & 0,98 & 0,98 & 0,98 & 0,97 \\ 0,99 & 0,99 & 0,98 & \mathbf{0,98} & 0,98 & 0,98 & 0,98 & 0,97 \\ 0,99 & 0,99 & 0,99 & 0,99 & 0,99 & 0,99 & 0,99 & 0,99 \\ 0,97 & 0,98 & 0,98 & 0,98 & 0,97 & 0,98 & 0,98 & 0,98 \\ 0,94 & 0,94 & 0,94 & 0,94 & 0,94 & 0,94 & 0,95 & 0,95 \\ 0,92 & 0,93 & 0,93 & 0,93 & 0,93 & 0,93 & 0,94 & 0,95 \end{bmatrix}$$

Fig. 7. Exemplo de região 8×8 em um mapa de similaridade (os valores estão truncados para fins de exibição). Apenas o elemento em negrito ($SSIM_w$) é preservado devido à dizimação no mapa de similaridade.

($SSIM_w$) é preservado. Neste exemplo, o erro $e[w]$ (12) associado é igual a $-0,010803$. Pode-se notar que os valores de SSIM da região são suficientemente próximos para que um único elemento seja um bom representante de toda a vizinhança. Apesar de não se poder generalizar, tal comportamento mostra-se razoável para imagens naturais em que os *pixels* de uma vizinhança 8×8 são fortemente correlacionados.

A Tabela I mostra que o erro médio μ_e possui valores pequenos (da ordem de 10^{-3} em sua maioria) quando comparados com a faixa de valores que a MSSIM pode assumir (de 0 a 1). A Tabela II, por sua vez, exibe variâncias também baixas. Tais resultados revelam que não há comprometimento da qualidade medida em relação à MSSIM quando a janela deslizante original é trocada pelas vizinhanças 8×8 não superpostas.

Os dados apresentados indicam que não só a MSSIM pode ser usada em codificadores baseados em DCT sem muito impacto na estrutura original do codificador, evitando a necessidade de calcular DCTs adicionais para a avaliação de qualidade e, mais ainda, sugerem a possibilidade de reduzir a complexidade da abordagem original da MSSIM.

VI. CONCLUSÕES

Neste trabalho, foi apresentado um estudo sobre o uso da MSSIM na implementação de codificadores perceptuais baseados em DCT. Primeiramente, as expressões que definem a MSSIM foram reescritas no domínio da DCT. Em seguida, discutiu-se a adaptação da MSSIM à estrutura convencional de codificadores DCT. Foi mostrado que o uso de regiões 8×8 sem superposição, típicas das estratégias de codificação, ao invés de janelas deslizantes, utilizadas na MSSIM original, resulta em dizimação do mapa de similaridade. Os resultados apresentados, contudo, indicam não haver perdas significativas na caracterização da qualidade visual em função da referida dizimação para o conjunto de teste avaliado. Conseqüentemente, além de ter mostrado que a MSSIM é bastante adequada à estrutura de codificadores baseados em DCT, os resultados deste trabalho sugerem uma maneira de reduzir a complexidade computacional associada ao cálculo da referida métrica. Como proposta de continuidade, pretende-se investigar possibilidades para o uso da MSSIM em codificação JPEG.

REFERÊNCIAS

[1] G. K. Wallace, "The jpeg still picture campression standard," *Communications of the ACM*, vol. 34, no. 4, pp. 30–44, April 1991.

[2] J. Solomon, A. Watson, and A. Ahumada, "Visibility of dct basis functions: effects of contrast masking," in *Proc. Data Compression Conference DCC '94*, Snowbird, Utah, USA, Mar. 1994, pp. 361–370.

[3] A. Watson, "Perceptual optimization of dct color quantization matrices," in *Proc. ICIP-94. IEEE International Conference Image Processing*, vol. 1, Austin, Texas, USA, Nov. 1994, pp. 100–104 vol.1.

[4] H. R. Wu and K. R. Rao, Eds., *Digital Video Image Quality and Perceptual Coding*. Taylor and Francis, 2006.

[5] A. B. Watson, Ed., *Digial Images and Human Vision*. Bradford Books, 1993.

[6] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[7] Z. Wang and A. C. Bovik, *Modern Image Quality Assessment*. Morgan & Claypool, 2006.

[8] S. Channappayya, A. Bovik, and R. Heath, "Rate bounds on ssim index of quantized images," *IEEE Trans. Image Process.*, vol. 17, no. 9, pp. 1624–1639, 2008.

[9] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[10] P. Le Callet and F. Autrusseau, "Subjective quality assessment irccyn/ivc database," 2005, http://www.irccyn.ec-nantes.fr/ivcdb/.