Estudo Comparativo entre Métricas para Avaliação da Qualidade de Imagens

Ronaldo de Freitas Zampolo e Rui Seara

***Resumo*— Neste artigo, são comparadas tres medidas usadas para avaliar a qualidade de imagens: erro quadratico medio (mean-square error MSE), medida de qualidade de ruıdo (noise quality measure NQM), e medida de informação estrutural (structural information metric SIM). Tal comparação e realizada para avaliar o desempenho das referidas metricas em situações praticas. Resultados experimentais e ındices estatısticos de desempenho são apresentados. Esses resultados sugerem a existencia de situações nas quais o MSE supera em desempenho as duas outras medidas.**

***Palavras-Chave*— Banco de imagens LIVE, metricas para qualidade de imagem, NQM, qualidade perceptual de imagem, SIM.**

***Abstract*— In this paper three image quality metrics are compared: mean-square error (MSE), noise quality measure (NQM), and structural information metric (SIM). Such a comparison is made in order to evaluate the performance of those different metrics under practical conditions. Experimental results along with statistical indices of performance are provided. The referred results suggest that there are situations in which the MSE-based metric outperforms the two other considered metrics.**

***Keywords*— LIVE image database, image quality metrics, NQM, perceptual image quality, SIM.**

I. INTRODUÇÃO

Neste artigo, três métricas para avaliação da qualidade de imagens são comparadas usando os dados subjetivos experimentais referentes às imagens JPEG2000 da base LIVE [1], a fim de verificar o desempenho das referidas métricas em situações práticas.

Medidas para qualidade de imagens são importantes no processamento de imagens, uma vez que o desempenho de uma dada aplicação é fortemente influenciado pelo seu associado, e comumente subjacente, modelo de imagem. Considerando sistemas projetados para o consumo humano, medidas que sejam capazes de representar a percepção humana da qualidade de imagens são desejáveis. Tais medidas deveriam ser capazes de considerar as condições de visualização (por exemplo, iluminação do ambiente, distância de visualização, etc.) bem como as características visuais daquele que observa a imagem.

Assim, desde os trabalhos pioneiros em medidas de contraste, devidos a Michelson [2], Weber e Fechner [3], aos dias atuais, muita pesquisa vem sendo realizada no desenvolvimento de métricas para qualidade de imagem no sentido perceptual. Atualmente, a pesquisa nessa área vem sendo direcionada para a predição da qualidade de cenas coloridas e naturais através de basicamente três maneiras: quando a cena isenta de degradação encontra-se disponível (referência plena ou total); quando somente informação parcial sobre a cena original está disponível (referência parcial); e quando a cena original é indisponível (sem referência) [4]. Em qualquer uma das estratégias mencionadas, o procedimento para avaliação da qualidade de imagens pode ser encarado como sendo composto por um mesmo conjunto de etapas: pré-processamento, filtragem usando a função de sensibilidade ao contraste (contrast sensitivity function – CSF), decomposição em canais perceptuais, normalização do erro, e *error pooling* [5], [6]. Esse conjunto de etapas, necessário para conferir às métricas características que as façam semelhantes ao sistema visual humano (human visual system – HVS), resulta na maior complexidade computacional apresentada pelas métricas psicovisuais em relação àquelas baseadas no erro quadrático médio (mean-square error – MSE) como, por exemplo, a razão entre sinal e ruído de pico (peak signal-to-noise ratio – PSNR) e a razão entre sinal e ruído (signal-to-noise ratio – SNR), citando as mais comuns. Contudo, tal desvantagem (complexidade mais alta) espera-se ser compensada pela maior correlação entre qualidade subjetiva e predita.

Devido às evidências estatísticas favoráveis em diversos trabalhos na área [4]–[8], há poucas dúvidas da boa correlação da qualidade predita pelas métricas psicovisuais propostas até então e a qualidade subjetiva obtida experimentalmente. Por outro lado, as medidas baseadas no MSE são consideradas como de pobre representação da qualidade percebida de imagens [4], [9], [10].

Este fato seria suficiente para justificar todos os esforços computacionais e de pesquisa associados às medidas perceptuais, se o número de estudos comparativos entre medidas perceptuais e aquelas baseadas no MSE não fosse tão reduzido. Então, surge a questão: quão melhor uma medida baseada no HVS é em relação a qualquer outra medida baseada no MSE?

Como uma primeira abordagem a este assunto, o presente trabalho compara três medidas da qualidade de imagens: o MSE, a medida de qualidade de ruído (noise quality measure – NQM) [11], e a medida de informação estrutural (structural information metric – SIM) [4], sendo as duas últimas medidas perceptuais. As imagens JPEG2000 e suas respectivas avaliações usadas vêm da base LIVE [1]. Para cada medida sob avaliação, são consideradas modelagens dos tipos linear e não-linear. Para todos os casos, são calculados índices estatísticos

Ronaldo de Freitas Zampolo, LaPS – Laboratório de Processamento de Sinais, Departamento de Engenharia Elétrica e de Computação, Universidade Federal do Pará, Belém, PA, E-mail: zampolo@ieee.org.

Rui Seara, LINSE – Laboratório de Circuitos e Processamento de Sinais, Departamento de Engenharia Elétrica, Universidade Federal de Santa Catarina, Florianópolis, SC, E-mail: seara@linse.ufsc.br.

Este trabalho foi parcialmente financiado pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq).

de desempenho, seguindo as orientaçãoes estabelecidas pelo *video quality experts group* (VQEG) [12].

O restante desse texto está estruturado como segue. A Seção II trata sucintamente das métricas de qualidade consideradas. A Seção III define os testes e índices estatísticos de desempenho usados na modelagem das avaliações de qualidade. A Seção IV apresenta os resultados obtidos para as três métricas testadas. E na Seção V, as principais conclusões do artigo são relacionadas.

II. Medidas para Avaliação da Qualidade de Imagens

Nesta seção, as medidas de qualidade usadas são comentadas brevemente.

A. Erro Quadrático Médio (Mean-square error – MSE)

O MSE, assim como as métricas dele derivadas, está entre as métricas mais usadas, devido em grande parte à simplicidade computacional das expressões resultantes, adequadas a sistemas com grande volume de dados a serem processados ou quando o tempo de processamento é um elemento crítico. O MSE é definido por

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N}[x_1(i) - x_2(i)]^2, \quad (1)$$

onde $x_1(i)$ e $x_2(i)$ denotam a imagem de referência (original) e a imagem de teste (degradada), respectivamente; e N é o número de amostras da imagem.

A principal objeção ao MSE (bem como às métricas nele baseadas) está na impossibilidade de representar as condições de visualização bem como as características de percepção particulares de um dado observador.

B. Medida de Qualidade de Ruído (Noise quality measure – NQM)

A NQM foi originalmente proposta para avaliar a qualidade de imagens degradadas somente por injeção de ruído [11]. Todavia, o procedimento da NQM também apresenta resultados aceitáveis na presença de outros tipos de degradação [7], [8]. Essa métrica é baseada no sistema visual humano, permitindo seu ajuste à distância de visualização, dimensões e resolução do monitor, etc. A NQM é também sensível aos chamados efeitos de mascaramento da percepção de contraste, sendo dada por

$$NQM = 10\log_{10}\left\{\frac{\sum\limits_{(m,n)} O_s(m,n)^2}{\sum\limits_{(m,n)}[O_s(m,n) - I_s(m,n)]^2}\right\}, \quad (2)$$

onde $O_s(m,n)$ e $I_s(m,n)$, no contexto deste trabalho, são denominadas versões sintetizadas da imagem original e imagem sob avaliação, respectivamente.

Tanto $O_s(m,n)$ quanto $I_s(m,n)$ são calculadas usando a função de limiar de contraste (contrast threshold function – CTF) (3) e a medida de contraste de Peli [3], [13].

A CTF é definida pela expressão

$$CTF(f) = \frac{k}{CSF(f)}, \quad (3)$$

onde f denota a freqüência espacial radial (ciclos/grau), k é uma constante de escalamento e a CSF é a função de sensibilidade ao contraste (contrast sensitivity function - CSF).

A CSF adotada nas simulações a serem apresentadas é dada por

$$CSF(f) = 2,6\,(0,0192 + 0,114f)\exp\left[-(0,114f)^{1,1}\right], \quad (4)$$

onde f representa a freqüência espacial radial (ciclos/grau).

C. Medida de Informação Estrutural (Structural information metric – SIM)

Esta métrica é proposta em [4], onde é validada experimentalmente. Consiste em comparar a informação estrutural da imagem com distorções com a da imagem original. Para tanto, cada uma das referidas imagens é processada pelos seguintes blocos que representam o modelo do HVS adotado:

- normalização dos componentes cromáticos, para simular a adaptação à intensidade luminosa;
- transformação dos valores RGB em luminâncias;
- projeção em mapa de cores perceptual (espaço de cores de Krauskopf ACr1Cr2);
- ponderação perceptual do espectro via filtragem usando CSF;
- decomposição dos sinais em canais perceptuais;
- extração de segmentos orientados e medida de características;
- transformação de parâmetros para gerar representação estrutural invariante à translação, rotação e zoom.

Em seguida, é calculada a medida de similaridade estrutural entre as imagens degradada e original. Esse método pode ser usado com referências reduzidas, mas aqui é aplicado em situações de referência plena. A implementação SIM usada neste artigo, denominada Quality Assesser for Windows Application, pode ser obtida em http://www.dcapplications.t2u.com.

III. Modelagem da Qualidade de Imagem

Esta seção aborda alguns conceitos relacionados à modelagem de dados referentes à avaliação de qualidade, considerando as orientações sugeridas pelo VQEG em [12].

A. Modelagem através da função qui-quadrado

O critério adotado nesse trabalho é a minimização do qui-quadrado (5). Tal critério pode ser visto como o método dos mínimos quadrados ponderado, no qual os coeficientes de ponderação são os inversos dos desvios padrões dos dados coletados y_i. A vantagem da abordagem qui-quadrado em relação à de mínimos quadrados é que amostras de maior variância, logo menos confiáveis, são conseqüentemente de menor importância no cômputo de χ^2. Os parâmetros assim obtidos correspondem à estimativa de máxima verossimilhança do modelo considerado. A função qui-quadrado é dada por

$$\chi^2 = \sum_{i=1}^{N}\left[\frac{y_i - g(x_i)}{\sigma_i}\right]^2, \quad (5)$$

onde N é o número de pontos avaliados, (y_i, x_i) referem-se aos pares de dados coletados experimentalmente, σ_i é o desvio padrão de um ponto específico i, e $g(x_i)$ indica o modelo escolhido que relaciona y_i e x_i. Nesse trabalho, i representa uma imagem JPEG2000 da base LIVE; y_i denota a média das avaliações (mean opinion score – MOS) da imagem associada; x_i é o valor de uma das três métricas (MSE, NQM ou SIM) para uma dada imagem i; e σ_i é o desvio padrão das avaliações subjetivas registradas na base LIVE. A função $g(x_i)$ é discutida nas próximas seções.

B. Regressão linear

Quando y_i é aproximado por uma linha reta dependente de x_i, a função $g(x_i)$ é chamada regressão linear e é dada por

$$g(x_i) = ax_i + b, \tag{6}$$

onde a e b são os coeficientes angular e linear de $g(x_i)$, respectivamente.

C. Modelagem não-linear

Neste caso, $g(x_i)$ não é restrito a ser uma linha reta. A partir daqui, a função não-linear $g(x_i)$ será baseada na função proposta em [12], definida como

$$g(x_i) = \frac{\beta_1 - \beta_2}{1 + \exp\left\{-\frac{x_i - \beta_3}{|\beta_4|}\right\}} + \beta_2, \tag{7}$$

onde β_j são coeficientes a serem estimados de tal forma que $g(x_i)$ melhor modele as avaliações experimentais, com condições iniciais: $\beta_1 = \max(y_i)$; $\beta_2 = \min(y_i)$; $\beta_3 = \bar{x}$, este último representando o valor médio do vetor $[x_1, \ldots, x_N]$; e $\beta_4 = 1$.

As seções restantes discutem alguns índices estatísticos usados para qualificar os parâmetros estimados.

D. Medida de qualidade de modelagem

Assumindo que os dados obtidos experimentalmente possam ser considerados amostras de um processo de distribuição Gaussiana, a probabilidade q de que um qui-quadrado exceda um valor particular χ^2 aleatoriamente é dada por [14]

$$q = Q[(N - M)/2, \chi^2/2], \tag{8}$$

onde N representa o número de pontos avaliados, M é o número de parâmetros do modelo, e Q denota a função gama incompleta, definida por

$$Q(a, b) = 1 - \frac{1}{\Gamma(a)} \int_0^b e^{-t} t^{a-1} dt, \qquad a > 0, \tag{9}$$

onde $\Gamma(a)$ é a função gama.

De acordo com o apresentado anteriormente, quanto maior o valor de q, maior o grau de validade do modelo estimado. Da mesma forma, valores de q baixos indicam modelos estimados de validade questionável.

Em situações práticas, $q \geq 10^{-3}$ indica estimações válidas, enquanto que se $q < 10^{-3}$, outro modelo deve ser escolhido para representar os dados obtidos.

E. Coeficiente de correlação linear (Pearson's r)

O coeficiente de correlação linear é amplamente usado para quantificar a correlação entre duas variáveis. No contexto da avaliação da qualidade de imagens, o coeficiente de correlação linear é definido como

$$r = \frac{\sum_{i=1}^{N} [g(x_i) - \bar{g}][y_i - \bar{y}]}{\sqrt{\sum_{i=1}^{N} [g(x_i) - \bar{g}]^2} \sqrt{\sum_{i=1}^{N} [y_i - \bar{y}]^2}}, \tag{10}$$

onde $\bar{g}$ e $\bar{y}$ denotam o valor médio dos vetores $[g(x_1), \ldots, g(x_N)]$ e $[y_1, \ldots, y_N]$, respectivamente.

O coeficiente de Pearson r varia de -1 a 1. Valores absolutos próximos a zero de r indicam que as duas variáveis apresentam fraca associação (pouco correlacionadas). Se $r = 1$ ($r = -1$), correlação positiva (negativa) completa, então $g(x_i) = ay_i + b$ com $a > 0$ ($a < 0$).

F. Coeficiente de correlação ponderada (r_w)

Basicamente, r_w é o coeficiente de correlação linear ponderado pelo inverso da variância das medidas. O objetivo dessa abordagem é o mesmo do caso do critério de mínimos quadrados ponderados. O coeficiente r_w é dado por [12]

$$r_w = \frac{\sum_{i=1}^{N} w_i [g(x_i) - \bar{g}_w][y_i - \bar{y}_w]}{\sqrt{\sum_{i=1}^{N} w_i [g(x_i) - \bar{g}_w]^2} \sqrt{\sum_{i=1}^{N} w_i [y_i - \bar{y}_w]^2}}, \tag{11}$$

onde $w_i = \frac{1}{\sigma_i^2}$.

G. Correlação de Spearman (r_s)

O coeficiente de Spearman r_s é uma medida de correlação não-paramétrica, definida em [14] como

$$r_s = \frac{\sum_i (G_i - \bar{G})(Y_i - \bar{Y})}{\sqrt{\sum_i (Gi - \bar{G})^2} \sqrt{\sum_i (Y_i - \bar{Y})^2}}, \tag{12}$$

onde G_i e Y_i denotam as posições ordenadas de $g(x_i)$ e y_i, respectivamente.

A correlação de Spearman é usada como medida de monotonicidade [12] e, tal qual outras medidas de correlação não-paramétricas, é mais robusta do que a correlação linear [14].

H. Razão de Outliers

A razão de outliers (outlier ratio – OR) é uma medida de consitência [12], definida como

$$\text{OR} = \frac{n\_outliers}{N}, \tag{13}$$

onde $n\_outliers$ denota o número de outliers, os quais podem ser obtidos como segue:

$n\_outliers = 0$;
para $i = 1$ até N,
início
    $e_i = y_i - g(x_i)$;
    se abs$(e_i) > (2S_i)$ então $n\_outliers = n\_outliers + 1$;
fim

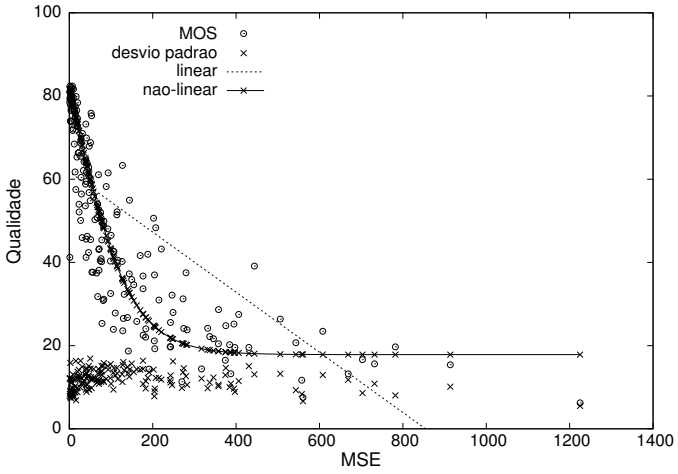

Fig. 1. Qualidade subjetiva *versus* MSE: MOS e desvio padrão para cada imagem avaliada, bem como as curvas de regressão linear e não-linear.

O termo S_i refere-se ao desvio padrão do erro dado por

$$S_i = \sqrt{\frac{1}{J_i - 1} \sum_{j=1}^{J_i} (e_{ij} - \mu_i)^2}, \qquad (14)$$

onde μ_i denota a média de e_{ij}, J_i é o número de vezes que a qualidade da imagem é avaliada pelos sujeitos do experimento, e e_{ij} é um erro definido por

$$e_{ij} = score_{ij} - g(x_i), \qquad (15)$$

com o índice j referindo-se a uma determinada avaliação da imagem i.

IV. Resultados

As curvas e as estatísticas apresentadas resumem os resultados obtidos para as imagens JPEG2000 da base LIVE e avaliações de qualidade correspondentes [1]. Para cada imagem, os valores de MSE, NQM e SIM foram determinados. Após tal etapa, foi realizada a modelagem do MOS em função de cada uma das métricas de qualidade usando (6) e (7). Figs. 1, 3 e 5 apresentam o MOS e desvio padrão de cada imagem e os modelos estimados, por regressão linear e não-linear, em função das métricas MSE, NQM e SIM, respectivamente. Nas Figs. 2, 4 e 6, ao invés do MOS, são mostradas as avaliações subjetivas propriamente ditas e modelos estimados em função das métricas usadas. Nas Tabelas I e II, são apresentados os índices estatísticos para $g(x_i)$ obtidos de (6) e (7), respectivamente.

TABELA I
MODELAGEM POR REGRESSÃO LINEAR

	MSE	NQM	SIM
q	$1,0576 \times 10^{-18}$	$0,059838$	$0,97208$
r	$0,71736$	$0,84797$	$0,89570$
r_w	$0,74261$	$0,87501$	$0,91664$
r_s	$0,89526$	$0,86887$	$0,89791$
OR	$0,17160$	$0,071006$	$0,029586$
χ^2	$375,79$	$196,34$	$133,86$

TABELA II
MODELAGEM NÃO-LINEAR

	MSE	NQM	SIM
q	$0,98792$	$0,82587$	$0,98300$
r	$0,89086$	$0,87188$	$0,90062$
r_w	$0,92552$	$0,90744$	$0,91995$
r_s	$0,89526$	$0,86887$	$0,89791$
OR	$0,023669$	$0,059172$	$0,029586$
χ^2	$126,79$	$147,92$	$128,85$

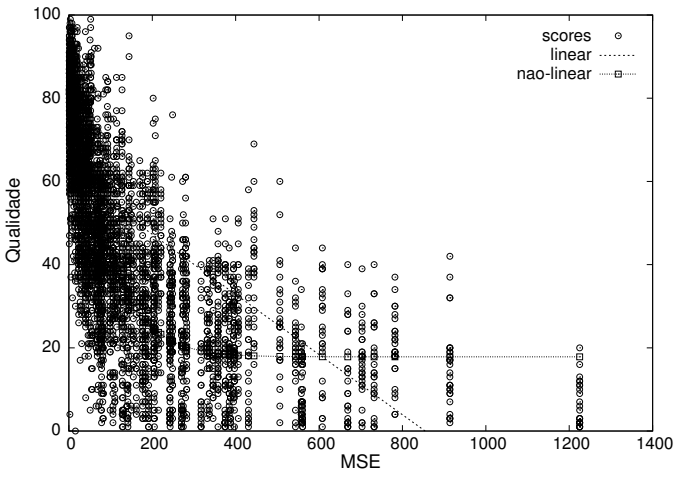

Fig. 2. Qualidade subjetiva *versus* MSE: *scores* para cada imagem avaliada, bem como as curvas de regressão linear e não-linear.

A. Comentários

Observando a Tabela I, a modelagem por regressão linear para o caso do MSE deve ser rejeitada, uma vez que $q < 10^{-3}$, enquanto que para as outras métricas os modelos estimados podem ser considerados válidos. Os melhores índices são obtidos com a métrica SIM.

Ao analisar a Tabela II, modelos válidos são estimados para todas as três métricas; contudo, o modelo referente ao MSE possui um desempenho ligeiramente melhor como mostram os índices estatísticos. Tal resultado, a princípio inesperado, revela que há situações nas quais o MSE supera as métricas psicovisuais testadas. Deve-se notar que este resultado depende da base de dados usada (LIVE), a qual por sua vez depende do respectivo procedimento experimental adotado.

Na opinião dos autores, testes adicionais devem ser realizados com diferentes bancos de imagem, outras medidas de qualidade baseadas no sistema visual humano e procedimentos experimentais, no sentido de estabeler qualquer generalização de conclusões.

As rotinas usadas estão disponíveis para *download* em www.laps.ufpa.br/zampolo/engl/dwnl.html.

V. Conclusões

Neste trabalho, três métricas para qualidade de imagens (MSE, NQM e SIM) foram comparadas, usando imagens JPEG2000 e suas respectivas avaliações do banco de imagens LIVE. Em geral, o desempenho da modelagem não-linear mostrou-se melhor que o da modelagem linear para cada uma das métricas consideradas. Dentre as métricas comparadas,

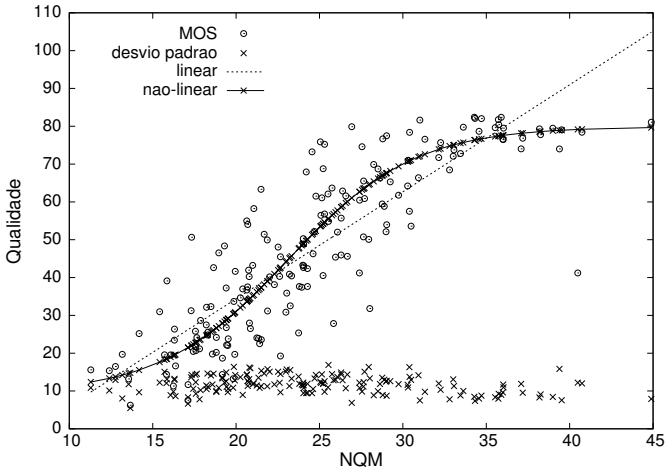

Fig. 3. Qualidade subjetiva *versus* NQM: MOS e desvio padrão para cada imagem avaliada, bem como as curvas de regressão linear e não-linear.

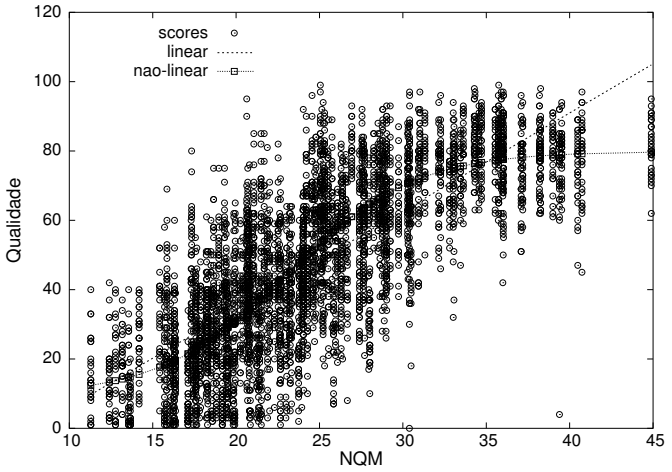

Fig. 4. Qualidade subjetiva *versus* NQM: *scores* para cada imagem avaliada, bem como as curvas de regressão linear e não-linear.

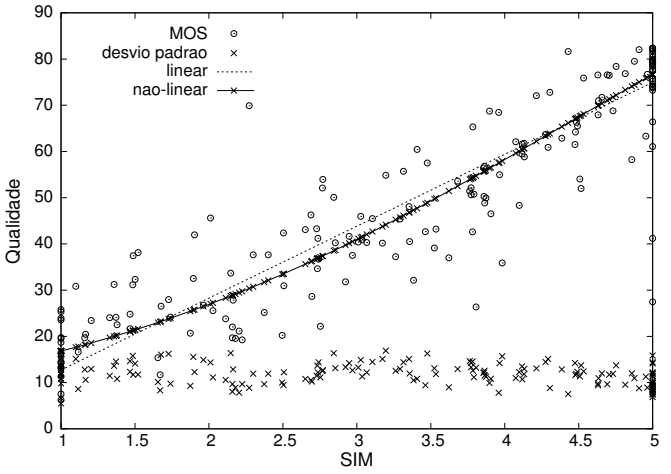

Fig. 5. Qualidade subjetiva *versus* SIM: MOS e desvio padrão para cada imagem avaliada, bem como as curvas de regressão linear e não-linear.

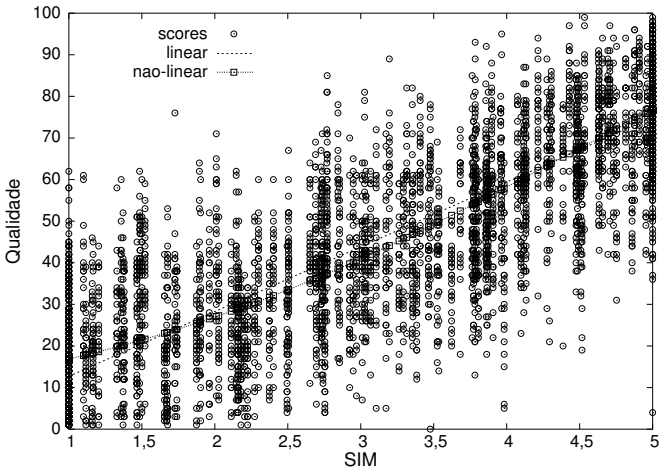

Fig. 6. Qualidade subjetiva *versus* SIM: *scores* para cada imagem avaliada, bem como as curvas de regressão linear e não-linear.

a baseada no MSE obteve melhor desempenho para o caso não-linear. Esse resultado inesperado mostra que há situações nas quais o MSE supera métricas baseadas no sistema visual humano (ou é, no mínimo, tão bom quanto). Para futuros trabalhos, sugere-se a realização de comparações usando outras imagens e métricas.

REFERÊNCIAS

[1] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik. Live image quality assessment database. Last access in Dec. 2004. [Online]. Available: http://live.ece.utexas.edu/research/quality

[2] A. A. Michelson, *Studies in Optics*, 3rd ed. Dover Publications, 1995.

[3] E. Peli, "Contrast in complex images," *J. Opt. Soc. Am. A*, vol. 7, no. 10, pp. 2032–2040, Oct. 1990.

[4] M. Carnec, P. L. Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *Proc. IEEE Int. Conf. Image Process.*, Barcelona, Spain, Sept. 2003.

[5] B. Fontaine, A. Saadane, and A. Thomas, "Perceptual quality metrics: Evaluation of individual components," in *Proc. IEEE Int. Conf. Image Process.*, Singapore, Oct. 2004, pp. 3507–3510.

[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 Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[7] R. F. Zampolo and R. Seara, "A measure for perceptual image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Barcelona, Spain, Sept. 2003.

[8] ——, "Perceptual image quality assessment based on Bayesian networks," in *Proc. IEEE Int. Conf. Image Process.*, Singapore, Oct. 2004, pp. 753–756.

[9] B. Girod, "What's wrong with mean squared error?" in *Digital Images and Human Vision*, A. B. Watson, Ed. Cambridge, MA: MIT Press, 1993, pp. 207–220.

[10] P. C. Teo and D. J. Heeger, "Perceptual image distortion," in *Proc. IEEE Int. Conf. Image Process.*, vol. 2, Austin, USA, Nov. 1994, pp. 982–986.

[11] N. Damera-Venkata, T. D. Kite, W. S. Geisler, *et al.*, "Image quality assessment based on a degradation model," *IEEE Trans. Image Processing*, vol. 9, no. 4, pp. 636–650, Apr. 2000.

[12] A. M. Rohaly, P. Corriveau, J. Libert, A. Webster, V. Baroncini, J. Beerends, J. L. Blin, L. Contin, T. Hamada, D. Harrison, A. Hekstra, J. Lubin, Y. Nishida, R. Nishihara, J. Pearson, A. F. Pessoa, N. Pickford, A. Schertz, M. Visca, A. Watson, and S. Winkler, "Video quality experts group: Current results and future directions," in *Proc. SPIE Visual Communications and Image Process.*, vol. 3, Perth, Australia, June 2000, pp. 742–753.

[13] E. Peli, "In search of a contrast metric: Matching the perceived contrast of gabor patches at different phases and bandwidths," *Vision Res.*, vol. 37, no. 23, pp. 3217–3224, Oct. 1990.

[14] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling, *Numerical Recipes in Pascal*. Cambridge University Press, 1996.