Avaliação e Comparação de Métricas de Referência Completa na Caracterização de Limiares de Detecção em Imagens

Ronaldo de Freitas Zampolo, Diego de Azevedo Gomes e Rui Seara

***Resumo*— Este artigo propõe uma metodologia de avaliação de desempenho como também compara algumas métricas de qualidade visual para caracterização de limiar de percepção em imagens degradadas por ruído aditivo Gaussiano. São avaliadas tanto métricas não-perceptuais (SNR e PSNR) quanto perceptuais (MSSIM, IFC, VIF e C4). O experimento de avaliação subjetiva que serve de base para os resultados apresentados é realizado com aproximadamente 120 observadores, considerando um conjunto de teste com 60 imagens, repartidas em cinco grupos. A análise dos dados busca estimar um valor de limiar de detecção para cada uma das métricas avaliadas bem como a qualidade obtida da correspondente estimativa.**

***Palavras-Chave*— Avaliação de qualidade de imagem, comparação entre diferentes métricas, detecção de diferenças entre imagens, limiares de percepção, métricas perceptuais.**

***Abstract*— This paper proposes a performance assessment procedure as well as compares some visual quality metrics for characterizing the perception threshold of images degraded by additive Gaussian noise. Both conventional (SNR and PSNR) and perceptual (MSSIM, IFC, VIF, and C4) metrics are evaluated. The subjective assessment experiment used for upholding the presented results is carried out with approximately 120 subjects, considering a test set with 60 images divided into 5 groups. The data analysis provides a perception threshold for each tested metric as well as assesses the quality obtained from the corresponding estimates.**

***Keywords*— Image quality assessment, comparison between different metrics, image difference detection, perception thresholds, perceptual metrics.**

I. INTRODUÇÃO

Muito interesse tem havido recententemente nas pesquisas sobre avaliação de qualidade visual. Em sistemas de processamento de imagem ou vídeo destinados ao consumo humano, é desejável que tais sistemas tenham seus parâmetros otimizados visando maximizar a qualidade da informação visual percebida pelos usuários finais. Nesse sentido, não há melhor forma de avaliação do que a subjetiva, na qual são exibidas imagens ou vídeos para um grupo de avaliadores que emitem julgamentos sobre a qualidade do que está sendo exibido. Contudo, tal estratégia, se por um lado é plenamente representativa da qualidade percebida, por outro requer um tempo considerável, demandando geralmente um elevado custo, portanto, não se prestando, na prática, à avaliação de novos algoritmos ou melhoramentos em sistemas de mídia visual. Uma alternativa à avaliação subjetiva é a bem conhecida avaliação objetiva, em que a inspeção humana é substituída por uma função matemática que realiza, então, uma avaliação automática da qualidade visual, reduzindo o tempo e o custo em comparação à avaliação subjetiva. Esse modelo matemático (ou métrica) pode ser baseado ou não em experimentos de avaliação subjetiva (métricas perceptuais e não-perceptuais, respectivamente). Como exemplos de métricas não-perceptuais mais utilizadas, podem ser citados: o erro quadrático médio (*mean-square error* – MSE), a razão sinal-ruído (*signal-to-noise ratio* – SNR) e a razão sinal-ruído de pico (*peak signal-to-noise ratio* – PSNR). As métricas não-perceptuais possuem como virtude a baixa complexidade computacional, mas apresentam baixa correlação com a qualidade visual percebida quando comparadas com as métricas perceptuais. Estas últimas representam melhor a percepção visual humana ao custo de uma maior complexidade computacional.

As métricas, tanto perceptuais quanto não-perceptuais, podem ser classificadas em: (i) de referência completa, quando uma imagem de referência está disponível e é utilizada para fins de comparação com uma imagem-teste (imagem cuja qualidade se quer avaliar); (ii) de referência parcial ou reduzida, quando se dispõe de algumas informações da imagem de referência, as quais são comparadas com o mesmo tipo de informação obtido da imagem-teste; e (iii) sem referência, quando a avaliação de qualidade é realizada apenas usando a imagem-teste.

Este trabalho propõe uma metodologia de avaliação de desempenho e compara métricas perceptuais e não-perceptuais de referência completa na caracterização de limiares de percepção de imagens degradadas por adição de ruído branco Gaussiano. Um conjunto-teste composto de 60 imagens, repartido em cinco grupos, é avaliado por aproximadamente 120 indivíduos através de um experimento subjetivo que fornece os dados para as análises apresentadas neste artigo.

As demais seções são organizadas como segue. Na Seção II, são apresentados os principais aspectos associados ao desenvolvimento e avaliação de métricas para qualidade visual, tal como sugerido pelo ITU e VQEG (*Video Quality Experts Group*). A Seção III discute de forma sucinta as métricas comparadas neste trabalho. O procedimento experimental é descrito na Seção IV. A Seção V apresenta a metodologia de avaliação proposta, bem como os resultados obtidos pelas

Ronaldo de Freitas Zampolo e Diego de Azevedo Gomes, Laboratório de Processamento de Sinais - LaPS, Faculdade de Engenharia da Computação, Universidade Federal do Pará, Belém - PA, Brasil, E-mails: zampolo@ufpa.br, diagomes@yahoo.com.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. O presente trabalho foi realizado com apoio do programa de bolsas PIBIC/UFPA.

análises do experimento subjetivo. Finalmente, as conclusões e propostas para continuidade do trabalho são consideradas na Seção VI.

II. Desenvolvimento e avaliação de métricas de qualidade visual

Esta seção descreve o processo atualmente utilizado no desenvolvimento e avaliação de métricas de qualidade visual em conformidade com as recomendações do ITU [1] e VQEG [2]. Tal processo desenvolve-se em três fases: (a) formação de uma base de dados obtida experimentalmente em que usuários são solicitados a emitir opiniões pessoais sobre a qualidade das imagens ou vídeos exibidos; (b) desenvolvimento de funções matemáticas que possuam expressiva correlação com os dados experimentais; e (c) avaliação de desempenho da métrica desenvolvida e eventual comparação com outras métricas propostas na literatura.

A primeira fase é de grande relevância para as restantes, uma vez que tem influência nos aspectos e/ou tipos de artefato que serão evidenciados pelo conjunto de sinais de teste e caracterizados pelas opiniões dos avaliadores humanos. Por exemplo, as imagens-teste podem retratar cenas *in-door* ou *outdoor*, conter imagens de pessoas ou paisagens naturais, objetos manufaturados, dentre outros. Com respeito aos tipos de degradação, o conjunto de sinais de teste pode apresentar adição de ruído Gaussiano ou ruído impulsivo, ou ser obtido via codificação JPEG2000 ou JPEG. Durante a obtenção das avaliações subjetivas, as imagens-teste podem ser exibidas com ou sem uma imagem de referência e as condições de visualização, tais como dimensão e resolução de monitores de vídeo, distância média entre monitores e usuários, iluminação do ambiente, dentre outras, devem ser, pelo menos, caracterizadas a fim de se definir o contexto em que os dados foram obtidos. Compreende-se que as análises realizadas após formada a base de dados aplicam-se, em princípio, somente ao referido contexto. Outro fator preponderante é o tipo de instrução dada aos participantes do experimento. Por exemplo, pode ser solicitado ao usuário que emita opinião sobre a qualidade de uma imagem apresentada, associando seu conceito de avaliação a um dado número em uma escala de 0 a 100. Em outro experimento, pode-se perguntar ao usuário se ele percebe alguma diferença entre duas imagens exibidas simultaneamente ou indicar qual delas apresenta maior nível de degradação. A questão aqui é que uma métrica derivada e avaliada em determinado contexto (considerando uma dada base de dados experimentais) pode apresentar níveis de desempenho diferentes quando o contexto se modifica.

A segunda fase, concernente ao desenvolvimento de uma nova métrica, depende da aplicação pretendida (que naturalmente também influencia a definição do experimento de coleta de dados). Por exemplo, em codificação de imagem ou vídeo, procura-se geralmente um adequado compromisso entre compressão e qualidade. Nesse caso, é interessante se dispor de uma métrica de qualidade cujos valores possam representar de maneira aceitável a provável qualidade percebida do sinal descodificado. Obter esse tipo de indicação é também importante na garantia de níveis de qualidade na transferência de imagem ou vídeo em sistema de transmissão *wireless*. Assim, a avaliação da imagem ou vídeo recebido pode servir como elemento de suporte à decisão em uma política de mudança ou seleção de canais. Outro exemplo em uma aplicação similar consiste no ajuste de parâmetros de sistemas de transmissão de banda larga em aplicações como IPTV (*Internet Protocol Television*). Por outro lado, em sistemas de marca d'água digital e codificadores de imagem para aplicações médicas, o maior interesse talvez seja dispor de uma métrica que verifique se as imagens produzidas são visualmente distintas das imagens originais (caracterização de limiares de detecção). Pode-se ainda apontar um terceiro objetivo na utilização de métricas perceptuais como aquele obtido em sistemas de restauração de imagens em que se pretende que a otimização paramétrica, guiada por algum critério objetivo, maximize a qualidade percebida da imagem restaurada ou a aproxime da imagem original em bases perceptuais.

Na terceira fase, comumente considera-se um conjunto de avaliações estatísticas visando quantificar a consistência da métrica proposta em relação aos dados experimentais. Coeficientes de correlação linear (Pearson) e de Spearman [3], taxa de *outliers* e valores de RMSE (*root-mean-squared error*) são medidas-padrão nesta etapa. Mais recentemente, contudo, face aos resultados obtidos pelo VQEG na avaliação de métricas de qualidade em seqüências de vídeo (quando se verificou que as métricas perceptuais avaliadas não eram distintas estatisticamente do PSNR), análises de significância vêm sendo incluídas com o objetivo principal de qualificar os resultados obtidos.

III. Métricas de qualidade visual

Nesta seção, as métricas de qualidade visual utilizadas neste trabalho são brevemente discutidas. Todas as métricas abordadas pertencem à classe das métricas de referência completa, ou seja, necessitam de uma imagem de referência, a partir da qual a qualidade das imagens-teste pode ser estimada. Algumas dessas métricas já foram objeto de extensiva comparação [4], entretanto, o contexto considerado era outro, diferente deste de caracterização do limiar de percepção.

A. Razão Sinal-Ruído (SNR) e Razão Sinal-Ruído de Pico (PSNR)

A razão sinal-ruído (*signal-to-noise ratio* – SNR) é uma métrica bem conhecida e bastante utilizada. Tal métrica simplesmente relaciona a energia do sinal de referência à do sinal-erro (diferença entre o sinal de teste e o de referência). A SNR (expressa em dB) é definida por

$$SNR = 10 \log \left\{ \frac{\sum_n x(n)^2}{\sum_n [x(n) - \hat{x}(n)]^2} \right\} \qquad (1)$$

onde $x(n)$ e $\hat{x}(n)$ representam, respectivamente, as imagens de referência e de teste.

A razão sinal-ruído de pico (*peak signal-to-noise* – PSNR), por sua vez, é também amplamente usada na área de codificação de imagem e vídeo; é a métrica padrão considerada para avaliar a qualidade dos diversos codificadores encontrados

nas mais diferentes aplicações. Sua expressão (dada em dB) é definida como

$$PSNR = 10 \log \left\{ \frac{Nk^2}{\sum_n [x(n) - \hat{x}(n)]^2} \right\} \quad (2)$$

onde N e k denotam, respectivamente, o número total de *pixels* e o valor máximo que um *pixel* pode assumir em uma dada imagem. Comumente, para imagem em tons de cinza com 8 bits/*pixel*, adota-se $k = 255$.

Tanto a SNR quanto a PSNR são métricas convencionais sem inspiração perceptual e possuem como vantagem uma baixa complexidade computacional.

B. Similaridade Estrutural Média (MSSIM)

A métrica similaridade estrutural média (*mean structural similarity* – MSSIM) [5], desenvolvida e validada através da base LIVE [6], é considerada uma métrica de baixa complexidade computacional, apesar de pertencer à classe das métricas perceptuais. A MSSIM mede o quanto a estrutura da imagem de teste está distante da estrutura da imagem de referência, associando então a similaridade estrutural avaliada com a qualidade percebida. A métrica em questão vem despertando a atenção da comunidade de pesquisadores devido aos bons resultados obtidos na representação da qualidade percebida como também pela sua simplicidade quando comparada com outras métricas psicovisuais.

C. Critério de Fidelidade de Informação (IFC)

A abordagem do critério de fidelidade de informação (*information fidelity criterion* – IFC) parte do pressuposto que o sistema visual humano evoluiu e adaptou-se a partir de estímulos provenientes de cenas naturais, sendo, portanto, sensível às variações das estatísticas que caracterizam esse tipo de imagem [7]. O IFC, então, assume que as estatísticas de imagens naturais podem ser modeladas no domínio das *wavelets* por um modelo tipo GSM (*Gaussian scale mixtures*) e define a qualidade visual como sendo a informação mútua entre uma imagem-teste e uma imagem de referência, considerando todas as sub-bandas envolvidas no processo de análise.

D. Fidelidade de Informação Visual (VIF)

A métrica fidelidade de informação visual (*visual information fidelity* – VIF) [8] consiste em uma extensão do IFC. As mesmas hipóteses e modelos são considerados, contudo, nesse caso, a qualidade visual é definida a partir da medida do quanto de informação a imagem de referência possui e o quanto dela pode ser extraída da imagem-teste. As análises deste trabalho apresentam resultados para a VIF tal qual definida em [8], e para uma versão simplificada, de menor complexidade computacional, denominada VIFp. Essa simplificação é proposta pelos mesmos autores da VIF original.

E. C4

Esta métrica é proposta em [9], onde é validada experimentalmente. As etapas para determinar a qualidade visual, segundo a métrica C4, consistem em: (a) normalização dinâmica, (b) correção de gama, (c) conversão para um espaço de cores perceptual (ACr1Cr2), (d) filtragem usando função de sensibilidade ao contraste (*contrast sensitivity function* – CSF), (e) decomposição em canais perceptuais, (f) extração de segmentos orientados, (g) transformação de coordenadas, e (h) medida de similaridade. A implementação da métrica C4 usada neste trabalho pode ser obtida em `http://membres.lycos.fr/dcapplications/`. Essa métrica apresenta elevada complexidade computacional, como o número de etapas necessárias para o seu cálculo já deixa entrever. Dois tipos de resultados podem ser obtidos pela implementação usada da C4. O primeiro consiste em uma medida de similaridade estrutural (denominada C4sim) e, o segundo, em um mapeamento não-linear dessa para uma medida de qualidade (*quality score*, chamada C4qs). A variação de C4sim é de 0 (pior similaridade estrutural possível) a 1 (identidade estrutural), enquanto a C4qs varia de 0 (qualidade mais baixa) a 5 (excelente qualidade).

IV. Procedimento experimental adotado

O tipo de experimento utilizado neste trabalho procura caracterizar limiares de percepção. O conjunto de teste é definido a partir de cinco imagens em tons-de-cinza encontradas na base IVC [10]: *Barbara*, *Clown*, *Fruit*, *Isabel* e *Mandril* (Fig. 1). Para cada uma dessas 5 imagens (aqui chamadas de imagens originais), são obtidas mais 11 imagens (denominadas imagens degradadas), geradas a partir da adição de ruído branco Gaussiano de diferentes intensidades (Fig. 2). Assim, o conjunto de teste totaliza 60 elementos, que podem ser repartidos em 5 grupos (um para cada imagem original) de 12 elementos cada (11 imagens degradadas + 1 imagem original).

Para a realização do experimento, são utilizados 3 computadores de mesa com uma configuração similar, com monitores convencionais de 17 polegadas e resolução de 1152×864 *pixels*. As condições de iluminação adotadas são as de um ambiente de trabalho típico em uma sala de aula ou escritório e a distância média de visualização é de 50 cm. A avaliação consistiu na apresentação de 60 pares de imagens (12 pares de cada um dos grupos citados) ao avaliador, selecionadas aleatoriamente entre imagens originais e imagens degradadas. O avaliador deveria indicar a imagem que apresentasse maior nível de ruído.

Aproximadamente 120 avaliadores participaram do experimento, dentre os quais tem-se indivíduos de ambos os sexos, professores e alunos da UFPA, com faixa etária entre 16 e 40 anos. A grande maioria não possuía qualquer especialidade na área de processamento de imagens. O tempo de duração de um experimento foi de aproximadamente 15 minutos, estando bem aquém dos 30 minutos recomendados como limite máximo para se evitar que os dados obtidos sejam comprometidos pela fadiga dos avaliadores [1].

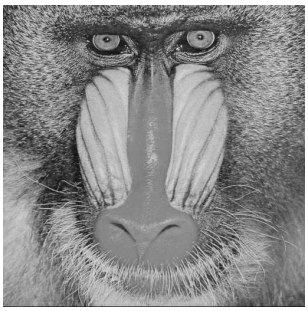
O procedimento de coleta de dados experimentais é automatizado através de um *software* escrito em linguagem Java, desenvolvido no Laboratório de Processamento de Sinais (LaPS) da UFPA. Além da referida automatização, o *software* realiza também algumas análises dos dados e possui suporte para apresentação gráfica dessas mesmas análises. Tanto os

(a) Barbara

(b) Clown

(c) Fruit

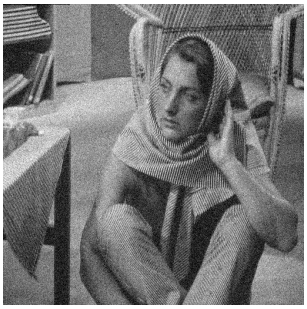
(d) Isabel



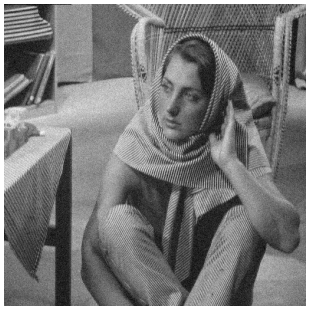
(e) Mandril

Fig. 1. Imagens originais utilizadas para definição do conjunto de teste.

(a) PSNR = $16,5763$ dB



(b) PSNR = $21,3787$ dB



(c) PSNR = $26,3409$ dB

(d) PSNR = $41,2575$ dB

Fig. 2. Exemplos de imagens-teste do grupo *Barbara*.

dados experimentais quanto os dados obtidos dos observadores são armazenados em um banco de dados MySQL remoto. O *software* ainda implementa rotinas para a realização automática de cópias de segurança dos resultados.

A. Interpretação dos Dados Experimentais

Neste experimento, são possíveis apenas dois tipos de respostas: respostas corretas (quando o avaliador aponta a imagem que realmente tem o maior nível de ruído); e respostas incorretas (quando o observador indica como mais ruidosa a imagem que na verdade apresenta um nível de ruído menor). Tais respostas resultam de duas situações: (a) a diferença entre as duas imagens apresentadas é efetivamente detectada, o que sempre leva a uma resposta correta; (b) a diferença não é percebida, resultando na seleção aleatória de uma das imagens pelo usuário, o que leva a um igual número de respostas corretas e incorretas por simetria [11]. Dessa forma, a probabilidade de respostas corretas pc (probabilidade do observador indicar a imagem que realmente tem mais ruído) pode ser dada por

$$pc = pd + \frac{1 - pd}{2} \tag{3}$$

$$= \frac{pd + 1}{2} \tag{4}$$

onde pd corresponde à probabilidade de que a diferença entre as imagens de um determinado par seja efetivamente detectada.

Quando metade dos observadores de fato percebe a diferença entre duas imagens ($pd = 0,5$), a probabilidade de resposta correta (pc) será igual a $0,75$. Tal valor é tomado como base para definir uma JND (*just noticeable difference*) [11], parâmetro que mede a percepção de diferenças entre imagens. Quanto maior for o valor de JND, maior será a percepção das diferenças entre as imagens. Dependendo da necessidade, pode-se definir JND para valores diferentes de pd, os quais então devem ser indicados explicitamente. Por exemplo, adotando $pd = 0,60$, referencia-se a JND associada por JND60 (60%).

Uma definição formal da JND pressupõe que a avaliação de diferenças entre pares de imagens obedece a uma dada distribuição de probabilidade. Para se evitar os erros ocasionados pelos prolongamentos infinitos das extremidades da distribuição normal, adota-se a função de distribuição de probabilidade angular, definida por

$$p_a(z_a) = \begin{cases} sen^2\left(\frac{z_a}{\sqrt{2\pi}} + \frac{\pi}{4}\right), & |z_a| \leq \sqrt{\frac{\pi^3}{8}} \\ 0, & z_a < -\sqrt{\frac{\pi^3}{8}} \\ 1, & z_a > \sqrt{\frac{\pi^3}{8}} \end{cases} \tag{5}$$

onde z_a caracteriza o desvio angular.

Invertendo a equação (5), obtém-se

$$z_a(p_a) = \sqrt{2\pi}\left[sen^{-1}(\sqrt{p_a}) - \frac{\pi}{4}\right]. \quad (6)$$

A JND de uma dada imagem é definida matematicamente (para o experimento particular de indicação de qual dentre duas imagens é a mais ruidosa) como a razão entre o desvio angular associado à probabilidade de uma imagem ser indicada como tendo mais ruído e o desvio angular associado à probabilidade da resposta correta (7). Assim,

$$JND = \frac{sen^{-1}(\sqrt{pp}) - \frac{\pi}{4}}{sen^{-1}(\sqrt{pc}) - \frac{\pi}{4}} \quad (7)$$

onde pp é a probabilidade da imagem ser indicada como tendo um maior nível de ruído.

No contexto experimental, a probabilidade pp é aproximada pela freqüência relativa de uma determinada imagem ter sido escolhida pelo usuário como tendo maior nível de ruído. Os resultados apresentados neste trabalho consideram $pd = 0,50$.

V. Análise de resultados

Esta seção é dividida em duas partes. Na primeira, é apresentada a metodologia utilizada na avaliação das métricas na caracterização do limiar de percepção. Na segunda, os resultados propriamente ditos são mostrados e discutidos.

A. Metodologia de Avaliação de Métricas para Caracterização de Limiar de Percepção

A metodologia proposta neste trabalho visa avaliar métricas de qualidade visual na caracterização objetiva de limiares de percepção da diferença entre imagens. Tal metodologia é aplicada à base de dados experimentais obtida segundo procedimentos e ferramentas descritos na Seção IV.

As imagens originais (Fig. 1) são sempre consideradas como imagens de referência nos experimentos pareados de comparação entre imagens. Verifica-se, em cada grupo, para quais valores das métricas avaliadas a deteção de diferenças ocorre. Em seguida, estima-se a média e o desvio-padrão dos valores de detecção com o objetivo de caracterizar, do ponto de vista estatístico, o desempenho de cada métrica. Assim, a média associa, em cada métrica, um valor numérico ao limiar de detecção, enquanto o desvio-padrão, por sua vez, avalia a robustez do referido valor de limiar para diferentes imagens e, conseqüentemente, a exeqüibilidade da utilização de uma determinada métrica em situações práticas.

B. Apresentação de Resultados e Comentários

Nas Figs. 3 e 4, são mostrados resultados típicos em que os dados experimentais para um determinado grupo de imagens são normalizados em termos de JND. Por razões de espaço, não são mostrados os gráficos de todos os grupos e de todas as métricas avaliadas. As Figs. 3 e 4 apresentam, respectivamente, a JND das imagens do grupo *Barbara* em função da PSNR e da métrica C4sim.

As linhas tracejadas nas Figs. 3 e 4 definem o que se pode denominar de área de não-detecção (entre -1 JND e 1 JND). As imagens cujos pontos situam-se nessa área não são percebidas como distintas da imagem de referência (nesse caso, a imagem original de cada grupo).

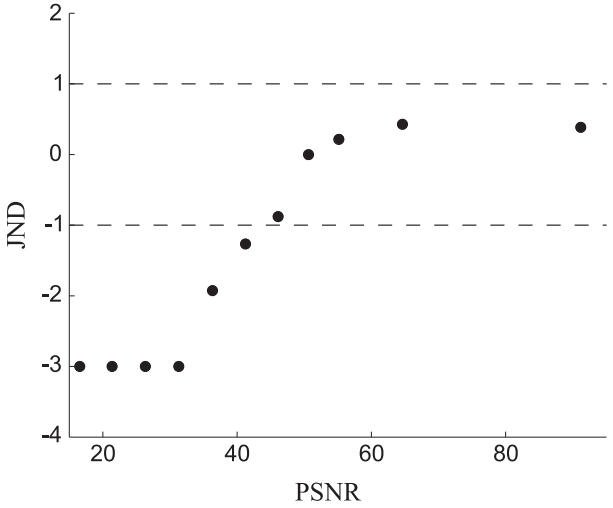


Fig. 3. Gráfico PSNR × JND para o grupo *Barbara*.

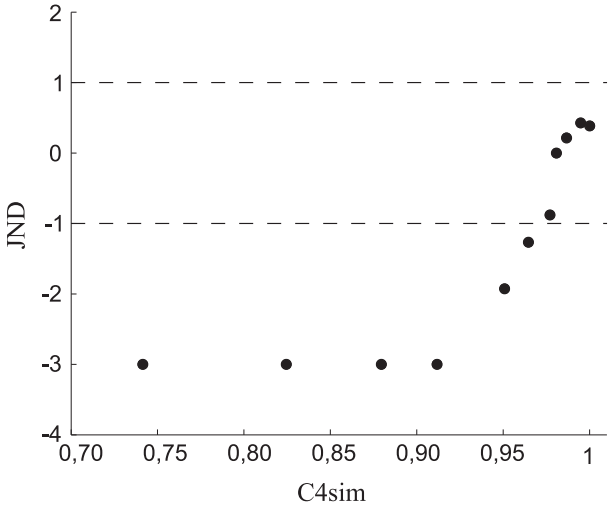


Fig. 4. Gráfico C4sim × JND para o grupo *Barbara*.

Por exemplo, na Fig. 3, é mostrado que para imagens do grupo *Barbara* cujos respectivos PSNRs estejam abaixo de 41 dB aproximadamente, mais de 50% dos observadores detectam a diferença entre essas imagens e a imagem *Barbara* original. Avaliação semelhante pode ser realizada para a Fig. 4, na qual observa-se o mesmo efeito para as imagens cujos valores de C4sim estejam abaixo de $0,9646$ aproximadamente.

As Tabelas I e II são obtidas através da análise dos limiares de detecção de todas as métricas para cada um dos grupos do conjunto de teste.

A Tabela I mostra os limiares referentes às métricas perceptuais, enquanto a Tabela II apresenta os limiares de percepção para as métricas convencionais. Na parte inferior de cada tabela, estão os valores médios e desvios-padrão para os limiares de percepção de cada métrica, considerando todos os grupos do conjunto de teste.

A fim de avaliar o desempenho das métricas testadas na caracterização do limiar de detecção, adotou-se como critério

de desempenho a análise dos valores de desvio-padrão. Em princípio, as melhores métricas na caracterização de limiar apresentam menores desvios-padrão.

Observando as Tabelas I e II, verifica-se que a métrica C4qs possui o menor desvio-padrão. Entretanto, nesse caso, a análise isolada desse parâmetro é inadequada. Nota-se que dos cinco grupos analisados, quatro apresentam o valor de detecção igual a $5,00$. Contudo, o valor máximo possível para C4qs é justamente $5,00$. Significa dizer que os usuários detectam as diferenças entre pares de imagens, mas a métrica C4qs não. Portanto, se não há discriminação numérica entre imagens percebidas como diferentes, a métrica não mostra utilidade na caracterização do limiar de percepção.

Em seguida, o melhor desvio-padrão é o da métrica C4sim, igual a $0,0125$. Diferente do que ocorre com a C4qs, para todos os cinco grupos testados, há discriminação numérica do limiar de percepção. Dessa forma, as imagens percebidas como diferentes da imagem original são também consideradas diferentes da imagem original pela métrica. Os resultados referentes à SNR e PSNR são merecedores de comentários particulares. Em primeiro lugar, deve-se notar que ambas as métricas apresentam limiar de percepção com desvios-padrão altos, em torno de 4 dB. Supondo que a estimativa de limiar possua uma distribuição Gaussiana, a região de $\pm 3\sigma$ em torno da média corresponde a aproximadamente 24 dB, o que é uma variação muito grande para pretender caracterizar um limiar de percepção usando qualquer uma das duas métricas citadas. Deve-se ressaltar que, em [4], as avaliações mostram que a PSNR apresentou desempenho compatível com as melhores métricas perceptuais testadas na caracterização de qualidade de imagens degradadas por injeção de ruído branco Gaussiano. Os resultados do presente trabalho mostram que, apesar da classe de imagens ser exatamente a mesma (imagens degradadas por adição de ruído branco Gaussiano), o desempenho da PSNR se altera pelo fato de se ter modificado o experimento de avaliação e conseqüentemente os dados obtidos.

TABELA I

MÉTRICAS PERCEPTUAIS – LIMIARES DE DETECÇÃO (μ: LIMIAR DE DETECÇÃO MÉDIO; σ: DESVIO-PADRÃO DOS LIMIARES DE DETECÇÃO)

Grupo	MSSIM	IFC	VIFp	VIF	C4qs	C4sim
Barbara	0,9734	9,0253	0,8014	0,9177	5,0000	0,9646
Clown	0,9289	6,3242	0,6654	0,8215	5,0000	0,9595
Fruits	0,9878	10,8936	0,8935	0,9666	5,0000	0,9817
Isabel	0,9524	7,3921	0,7178	0,8714	5,0000	0,9613
Mandril	0,9502	8,6836	0,6487	0,7857	4,9955	0,9470
μ	0,9585	8,4638	0,7454	0,8726	4,9991	0,9628
σ	0,0227	1,7316	0,1020	0,0725	0,0020	0,0125

VI. CONCLUSÕES

Neste trabalho, foi proposta uma metodologia para avaliação de métricas de qualidade visual aplicadas em detecção de diferenças entre imagens. Essa metodologia tem como ponto de partida a disponibilidade de uma base de dados obtida em experimentos subjetivos de detecção de artefatos entre duas imagens.

TABELA II

MÉTRICAS CONVENCIONAIS – LIMIARES DE DETECÇÃO (μ: LIMIAR DE DETECÇÃO MÉDIO; σ: DESVIO-PADRÃO DOS LIMIARES DE DETECÇÃO)

Grupo	SNR	PSNR
Barbara	34,9190	41,2575
Clown	29,9902	36,8731
Fruits	39,8047	45,5447
Isabel	34,9560	39,3025
Mandril	29,9743	35,2027
μ	33,9288	39,6361
σ	4,1144	4,0293

Em particular, o conjunto de teste utilizado continha imagens degradadas unicamente por ruído branco aditivo Gaussiano. Dentre as métricas candidatas encontravam-se: MSSIM, IFC, VIF, VIFp, C4qs, C4sim, SNR e PSNR. Dessas, apenas as duas últimas são não-perceptuais. Os resultados mostram que a métrica mais consistente e de maior potencial de utilização em sistemas práticos é a C4sim, pois apresenta menor desvio-padrão do seu valor de limiar quando comparado com os correspondentes valores das outras métricas candidatas. A PSNR, apesar do tipo de degradação do conjunto de teste, apresenta limiar de detecção com variância muito grande, sendo portanto inadequada.

Como proposta de continuidade deste trabalho, pretende-se investigar o desempenho de outras métricas de referência completa e também o desempenho de métricas de referência reduzida na caracterização de limiares de detecção. Dessa forma, visando ampliar o alcance dos resultados obtidos, novos conjuntos de teste estão em desenvolvimento. Nesses conjuntos, as imagens poderão apresentar outros tipos de degradação, tais como ruído impulsivo, quantização de coeficientes DCT e coeficientes *wavelet*.

REFERÊNCIAS

[1] ITU, "Itu-r recommendation bt.500-11 methodology for the subjective assessment for the television pictures," 2002.

[2] VQEG, "Final report from the video quality experts group on the validation of objective models of video quality assessment vqeg," on-line, mar 2002. [Online]. Available: www.vqeg.org

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

[4] 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.

[5] 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, 2004.

[6] H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, "Live image quality assessment database," 2006, http://live.ece.utexas.edu/research/quality.

[7] H. Sheikh, A. Bovik, and G. de Veciana, "An information fidelity criterion for image quality assessment using natural scene statistics," *IEEE Trans. Image Process.*, vol. 14, no. 12, pp. 2117–2128, Dec. 2005.

[8] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[9] 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 Processing*, Sep. 2003.

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

[11] B. W. Keelan, *Handbook of Image Quality*. Marcel Dekker, 2002.