Métricas objetivas para avaliação da qualidade de vídeo percebida [1]

C. M. de Farias
Programa de Pós-Graduação em Informática
Avenida Brigadeiro Tropowsky, Prédio CCMN, Bloco C
Rio de Janeiro, Brasil
claudiofarias@nce.ufrj.br

P.H. de A. Rodrigues
Programa de Pós-Graduação em Informática
Avenida Brigadeiro Tropowsky, Prédio CCMN, Bloco C
Rio de Janeiro, Brasil
aguiar@nce.ufrj.br

ABSTRACT
Video Quality assessment is one of the most challenging problems in real time communications. Most of the quality assessment algorithms is based on pixels, so they are dependant of the original video as reference. This dependancy make complicated to apply these algorithms in a real time scenario. Besides, these algorithms do not consider Human Visual System characteristics. This paper introduces 3 algorithms that use HVS characteristics. The first two algorithms still use the original video as reference and the third uses only the received video as parameter.

RESUMO
A avaliação da qualidade de vídeo um dos problemas mais desafiadores para a comunicação de tempo real. A maioria dos algoritmos de avaliação de qualidade de vídeo são baseados em pixels, ou seja dependem do vídeo original como referência. Essa dependência torna a aplicação desses algoritmos complicada no cenário de tempo real. Além disso esses algoritmos não levam em consideração informações do sistema visual humano. O trabalho apresenta 3 algoritmos que utilizam informações HVS. Os dois primeiros usam o vídeo original como referência e o terceiro usa somente o vídeo recebido como parâmetro.

Categories and subjective descriptors
C.2.3 [**Computer Systems and organization**]: Computer Communication Networks—*Network Operations*

General Terms
Algorithms, performance, management

Keywords
Video quality evaluation, HVS characteristics, FR and NR algorithms, resource consumption

[*]Objective metrics for perceived video quality assessment

1. INTRODUÇÃO
Com o advento de redes de altas velocidades, serviços de comunicação de vídeo em tempo real, tais como videoconferências, passam a se tornar lugar comum na sociedade da informação. Apesar de crescente popularidade e demanda, o tráfego de vídeo em tempo real ainda apresenta diversos desafios. A maioria das redes atuais ainda opera no esquema de melhor esforço, sem garantias para o tráfego em tempo real. Uso de QoS com reserva de banda ou priorização de filas ainda é incipiente.

Na Internet existe o problema de congestionamento nos roteadores, que pode gerar atraso na entrega de pacotes ou mesmo o seu descarte. O fato de um pacote na Internet passar por vários roteadores e, por conseguinte, por várias filas gera o problema da variação do atraso (*jitter*), que é extremamente prejudicial para as aplicações de tempo real, que são pouco tolerantes a grandes atrasos e *jitter*.

Aplicações como videoconferências e videochamadas ainda enfrentam o problema de garantir a comunicação entre as partes envolvidas. Os protocolos de comunicação mais comumente usados são o SIP [10] e o H.323 [9], sendo este último o mais utilizado ainda para conferências com vídeo.

As chamadas redes NGN (*Next Generation Networks*), ou redes de próxima geração, prometem integrar serviços multimídia entre os mais diversos dispositivos (computadores pessoais, telefones celulares e qualquer dispositivo capaz de acesso à rede). Nesse cenário os serviços de vídeo em tempo real e vídeo sob demanda se apresentam como serviços centrais a serem prestados. Nesse contexto, a discussão sobre a garantia de qualidade de serviço, já tão consolidada para voz, se torna um assunto cada vez mais comum quando se trata de vídeo.

Os parâmetros de QoS (*Quality of Service*) utilizados em aplicações de vídeo são diferentes dos utilizados no tráfego de voz, pela natureza intrinsecamente distinta das duas aplicações. Vídeo apresenta um volume de tráfego muito superior a voz. Além disso, vídeo dificilmente está desassociado da voz, o que leva ao problema de sincronização entre essas duas mídias.

Dada a capacidade visual humana, muitas vezes os erros introduzidos numa imagem passam desapercebidos. Assim, erros de transmissão, perdas de pacote e atrasos podem não

afetar a qualidade percebida pelo usuário. Seres humanos são mais tolerantes a distorções visuais do que a distorções no áudio. Portanto, o conceito de QoE (*Quality of Experience*) é essencial. Os modelos de avaliação de qualidade de vídeo devem levar em consideração certa tolerância aos erros introduzidos.

Em vista dos novos padrões de compressão de vídeo e proliferação de produtos para codificação de vídeo em tempo real, os algoritmos de avaliação de qualidade de imagem e vídeo se tornaram importantes. Estes algoritmos são utilizados pelas indústrias de telecomunicações, computação e mídia em geral em aplicações como HDTV, IPTV, serviços de vídeo na *web*, telefonia digital, entre outras. O sistema de medição baseado em MOS (*Mean Opinion Score*) é subjetivo e foi amplamente usado em medições de qualidade de vídeo durante a década passada. Porém existem diversos problemas associados a esta abordagem: há falta de um modelo consistente que realize a medição de forma simples; os paradigmas atuais são tediosos e caros de serem implementados; e uma avaliação subjetiva é muito difícil de ser inserida no processamento de vídeo em tempo real, pois as medições não podem ser implementadas de forma automatizada. Em contraposição, métricas de vídeo objetivas podem gerar uma medição de qualidade de vídeo por um período relativamente curto de tempo, o que é importante para as aplicações de tempo real.

Nos últimos anos, houve grande desenvolvimento nas técnicas de medição de qualidade de vídeo. Apesar do uso das métricas HVS (*Human Visual System*) resultar em resultados mais acurados pela perspectiva humana, a maioria desses algoritmos são algoritmos de Referência Completa (*Full Reference*) ou de Referência Reduzida (*Reduced Reference*), e alguns poucos são de Referência Nula (*Null Reference*) [26].

O método de referência completa computa a qualidade através das comparações de todos os pixels em cada imagem do vídeo distorcido com seu pixel correspondente no vídeo original. O método de referência reduzida extrai algumas características de ambos os vídeos (o original e o distorcido), e compara para gerar um valor de qualidade. Esse método é utilizado, em geral, quando o vídeo original não está disponível. O método de referência nula (ou sem referência) tenta avaliar a qualidade do vídeo distorcido baseando-se apenas no vídeo recebido, sem as informações do vídeo original. Esse método é utilizado quando o método de codificação do vídeo é conhecido.

Em cenários de tempo real é extremamente difícil ou mesmo impossível possuir o vídeo original como parâmetro para a medição da qualidade. Nestes casos, é de extrema importância usar métrica não referenciadas (*No Reference*).Trabalhos recentes propuseram diversas métricas não referenciadas [13], [27], [20], [21]. Além disso, nos cenários IMS (*IP Multimedia Subsystem*) muitos dispositivos são extremamente limitados em poder computacional e os algoritmos propostos podem não ser aplicáveis.

Este artigo apresenta 3 algoritmos, todos usando características HVS. Os dois primeiros são algoritmos de referência completa. O terceiro é um algoritmo de referência nula. O principal cenário para a aplicação dos algoritmos é o de vídeochamadas onde existem poucas mudanças de contexto das imagens. O diferencial do trabalho está em apresentar algoritmos para a avalição da qualidade de vídeo com baixo custo computacional. Por custo computacional entenda-se uso de memória e processador.

O artigo está organizado conforme apresentado a seguir. A seção 2 faz uma discussão sobre as métricas de vídeo e seu cenário atual. Na seção 3 os algoritmos são descritos. A seção 4 apresenta os resultados dos experimentos realizados. Finalmente, a seção 5 apresenta as conclusões do trabalho.

2. MÉTRICAS

O olho humano é sensível a luminância mais que ao valor absoluto de luminosidade [26]. De acordo com a Lei de Weber [2], se a luminância de um estímulo de teste é apenas notável a partir de uma luminância circundante, então a razão da diferença das luminâncias perceptíveis (a de estímulo e a circundante) e da luminância de estímulo é conhecida como Fração de Weber, em um valor aproximadamente constante [3]. Na prática, devido à iluminação ambiente cercando o display, o ruído em áreas escuras tende a ser menor do que o ruído em áreas de alta luminância. Entretanto, se a luminância de fundo está baixa, a Fração de Weber aumenta conforme a luminância de fundo diminui [15]. Por outro lado, se a luminância de fundo é alta, a fração de Weber permanece constante enquanto a luminância de fundo permanecer constante.

A sensibilidade ao contraste também varia conforme a frequência espacial, o que leva à função de sensibilidade de contraste (CSF - *Contrast Sensitivity Function*) [24]. É um fato conhecido que o olho humano é mais sensível a frequências espaciais menores do que as maiores. Essa propriedade tem sido amplamente explorada no desenvolvimento de TV's e câmeras. De fato a CSF é uma função da frequência espacial, frequência temporal, orientação, distância de visão e cor.

Considerando dois estímulos diferentes em uma mesma imagem, a presença e as características de um dos estímulos vai afetar como o outro é percebido. Isso é que chamamos de efeito de mascaramento. O efeito de mascaramento é complicado o suficiente para que nenhuma formulação teórica tenha sido capaz de justificar todas as suas nuânces [3].

Nas últimas três décadas várias métricas objetivas foram propostas [4], [11] e [20] para avaliar a qualidade de vídeo. A maneira mais fácil de se atribuir um valor de qualidade é utilizar simples diferenças numéricas entre valores da imagem distorcida e uma imagem de referência. Os métodos mais adotados utilizando características estatísticas da imagem são o MSE (*Mean Square Error*) e suas variantes. Dentre elas cabe destacar o PSNR (*Peak Signal Noise Ratio*), como métrica mais utilizada.

Entretanto, o MSE e suas variantes não estão bem relacionados com medidas de qualidade subjetivas, pois a percepção humana das distorções de imagem/vídeo e artefatos não é levada em consideração. A Figura 1 mostra um exemplo desse tipo de problema. Nota-se que nas imagens b e c, apesar de possuírem o mesmo valor de PSNR, possuem

Figura 1: Exemplo de distorção

percepções visuais muito distintas. Então percebe-se que a utilização do PSNR (e de mtricas baseadas em erros) por si só não implica em verificação consistente da qualidade da imagem. Faz necessário utilização de métricas auxiliares. Uma discussão detalhada do MSE é dada em [5].

A análise de características do sistema visual humano (*Human Visual System* – HVS)[4] tem gerado grandes avanços no campo da avaliação da qualidade de vídeo. Existem muitas características do HVS [24] que podem influenciar a percepção visual humana da qualidade da imagem. Ainda que o HVS seja extremamente complexo para ser completamente compreendido com os presentes meios da psicofisiologia, a incorporação de um modelo, mesmo que simplificado, em medição objetiva leva a uma melhor correlação com a opinião de observadores humanos [4]. Muitos algoritmos tem empregado modelos HVS de forma bem-sucedida [22, 27, 13, 25].

Apesar dos grandes avanços trazidos pelos métodos HVS, o alto custo computacional torna sua utilização muito complicada para dispositivos de baixo poder computacional. Atualmente a abordagem adotada é o uso de simplificação de modelos [22], retornando muitas vezes ao uso de MSE e PSNR para as análises. Basicamente, no trabalho de Apostolopoulos [22] é investigada a relação entre a perda de pacotes e distorções no vídeo decodificado. O argumento utilizado é que o uso de técnicas MSE torna mais simples a estimativa da qualidade em tempo real.

Segundo o trabalho proposto por Zhang [28], a medição da qualidade deve ser realizada sob dois pontos de vista: baseado no QoS da rede e dos pontos finais. Segundo os autores, as medições nos pontos finais podem ser vistas por agentes inteligentes medindo as condições da rede. Os métodos de medição nos pontos finais tem focado na diferenciação da perda por congestionamento da perda por erros, adotando alguns métodos heurísticos a partir dos intervalos de chegada dos pacotes ou dos intervalos entre pacotes consecutivos.

O trabalho de Suresh [20] apresenta uma métrica sem referência que busca de vários artefatos observados em sistemas de transmissão de vídeo. Esse algoritmo busca combinar linearmente diversas características da imagem para chegar um resultado mais relevante do ponto de vista subjetivo.

A maior vantagem das métricas baseadas em pixels (ou nos erros dos pixels) é a sua simplicidade. Elas podem ser facilmente adaptadas a um sistema de processamento de imagem e vídeo. Entretanto a falta de parâmetros HVS não as torna boas para a avaliação da distorção de perceptual de imagens. É apresentado em [4] e em [6] que um grande desenvolvimento pode ser obtido através da combinação de um modelo baseado em pixels com um modelo HVS muito simples.

Métodos baseados em sensibilidade aos erros tem sido os principais métodos de avaliação, entretanto alguns autores [25] seguem a filosofia de trabalhar com as distorções da estrutura da imagem como uma estimativa da distorção visual percebida. Inclusive, neste mesmo trabalho, uma abordagem para avaliação da qualidade é desenvolvida. Algumas métricas consideram apenas alguns tipos de distorções [11] ou métodos de codificação especiais [12]. Para imagens coloridas é desejado encontrar um bom espaço de cor onde cada canal de cor pode ser considerado independentemente [17].

3. ALGORITMOS PROPOSTOS

Avaliar a qualidade de vídeos transmitidos em tempo real através de algoritmos de referência completa é extremamente difícil visto que é necessário obter uma amostra confiável do vídeo original para ser comparado com o vídeo transmitido. Atualmente têm sido vistos muitos esforços para o desenvolvimento de métricas não referenciadas. [21] Apesar dessas métricas haverem obtido bons resultados, elas requerem recursos que não estão disponíveis em dispositivos muito restritos (em termos de poder de processamento e uso de memória). Por dispositivos restritos entenda-se dispositivos que usem processadores até a família ARM 920 e 64 Mb de RAM. Nessa categoria encontram-se os celulares mais comuns disponíveis atualmente.

Nota-se que em diversos trabalhos apresenta-se a luminância como a principal característica do sistema visual humano. Os algoritmos propostos tem como alvo cenários de tempo real como os vistos em IMS. O objetivo é construir uma métrica objetiva que use características HVS para melhorar o desempenho da métrica. Luminância é utilizada ao invés dos canais de cores pois, conforme dito anteriormente, o olho humano é mais sensível a luminância do que para as cores e também porque com essa estratégia haverá um volume menor de dados a ser avaliado (1 canal ao invés de 3 canais).

Os algoritmos apresentados trabalham com vídeo decodificado. Os erros introduzidos pela decodificação, perdas de pacotes, atrasos e *jitter* não são considerados de forma diferenciada. Estes são transmitidos como distorções no vídeo decodificado.

3.1 Primeiro algoritmo

A primeira tentativa para introduzir características HVS ao algoritmo é dividir a imagem em regiões geométricas. Esse algoritmo é uma variação do PSNR que usa os conceitos de mascaramento e a influência da luminância.

O algoritmo simplesmente divide cada *frame* em regiões iguais e calcula o PSNR sobre cada uma dessas regiões. Mas, como

visto na seção anterior, a luminância afeta a percepção das falhas, pesa-se cada região baseado em sua luminância média. E assim obtêm-se um valor de qualidade para o *frame*. Faz-se isto para cada *frame* do vídeo. Soma-se todos os valores obtidos e divide-se pelo número de *frames*.

ALGORITMO 1:

Primeiro passo: Divida a imagem em N regiões

Segundo passo: Calcule o PSNR em cada região

Terceiro passo: Para cada região calcule a luminância média

Quarto passo: A distorção de uma região será o produto do PSNR de cada região e da média da luminância. Faça isso para cada frame.

Pode-se analisar o algoritmo pelo seguinte prisma: regiões com baixa luminância possuem uma percepção de distorção mais tolerante do que regiões de luminância alta [26]. Nota-se, porém, que a distorção condicionada à luminância de determinada região altera positivamente ou negativamente a percepção de uma imagem. No exemplo da Figura 2, enquanto em b a distorção se deu em uma área de luminância alta e ficou extremamente perceptível ao usuário, na imagem c a distorção se deu em uma parte da imagem com a luminância mais baixa (a parte de baixo da imagem). A qualidade percebida pela imagem c é bem próxima da conseguida pela imagem A, que é a original.

3.2 Segundo algoritmo

A segunda tentativa é dividir a imagem utilizando suas bordas (através da aplicação de um filtro como Sobel [18], em nosso caso). É seguida a mesma premissa do algoritmo anterior, porém ao invés de dividir o *frame* geometricamente, são utilizadas as bordas das figuras. Neste algoritmo, parte-se do pressuposto de que seres humanos tendem a focar sua atenção em determinados objetos na cena e que distorções são percebidas a partir desses objetos [21]. Quanto maior for a distorção em uma determinada superfície de um objeto maior será a perda de qualidade sentida.

A aplicação do filtro gera ruídos. Esses ruídos são não podem ser processados como regiões da imagem. Para o agrupamento das regiões utiliza-se o algoritmo de Máxima Verossimilhança entre os tamanhos das áreas encontradas [1]. É esperado que o ruído gerado seja substancialmente menor que as informações da imagem. O tamanho de uma região ï¿½ dado unicamente pelo número de *pixels* que a compõe. Portanto pode-se agrupar esses conjunto de ruídos gerados e excluí-los da avaliação.

Esse procedimento é importante pois o ruído tende a ocorrer na borda das imagens. Então considera-se o objeto sempre com uma margem além da borda encontrada pelo filtro. Dessa forma não se considera o ruído da imagem, para que essa não seja dividida em duas regiões distintas.

Na figura 1, pode-se notar como o ruído se encontra próximo ao peixe. Esse ruído não deve ser considerado como região, mas como parte do objeto que compõe o fundo da imagem.

O algoritmo 2 é similar ao primeiro, também é uma variação do PSNR. Utiliza-se dos conceitos de mascaramento e influência da luminância. Entretanto o algoritmo 2 se utiliza de mais uma característica do sistema visual humano: foco em objetos.

O algoritmo divide a imagem usando um filtro de bordas (Sobel). A partir das bordas determinam-se regiões. Cada objeto descoberto pelo filtro é considerado uma região. Aplica-se o PSNR a essas regiões. Cada região é pesada de acordo com sua a luminância média (média ponderada). E assim obtêm-se um valor de qualidade para o *frame*. Faz-se isto para cada *frame* do vídeo. Soma-se todos os valores obtidos e divide-se pelo número de *frames*.

ALGORITMO 2:

Primeiro passo: Aplique o filtro de Sobel sobre a imagem e selecione as regiões que foram delimitadas. Para evitar o ruído gerado pelo filtro, apenas regiões com tamanhos dentro de uma mesma ordem de grandeza serão consideradas.

Segundo passo: calcule o PSNR de cada região delimitada

Terceiro passo: Para cada região calcule a luminância média.

Quarto passo: A distorção de uma região será o o produto do PSNR de cada região e da Luminância média da mesma. Faça isso para cada frame.

O problema com os dois algoritmos apresentados é que esses ainda são algoritmos de referência completa. Em alguns sistemas de vídeo conferência, o vídeo gerado localmente serve de parâmetro para a avaliação do vídeo enviado. Mas em uma chamada de vídeo normal, não há vídeo de origem para ser tomado como referência.

3.3 Terceiro Algoritmo

O terceiro algoritmo é não referenciado. O algoritmo se utiliza da variação da luminância dos pixels entre os diferentes frames como fator de avaliação. Se a variação da luminância de um pixel é similar à variação da região a sua volta (sua vizinhança [18]), então essa variação é considerada movimento; se a variação for diferente em relação à variação da regiãoa sua volta, ela é considerada distorção. O algoritmo também considera a influência de regiões de grande movimento. Seres humanos podem tolerar distorções em uma região de grande movimento. [27]

Neste algoritmo também há a preocupação com o ruído gerado pelo filtro. O ruído também deve ser ignorado a critério de comparação do algoritmo.

O algoritmo 3 não é variação do PSNR. Utiliza-se dos conceitos de mascaramento, influência da luminânci e foco em objetos. O algoritmo 3 se utiliza de mais características do sistema visual humano: influência do movimento e sequência de cenas.

O algoritmo divide a imagem usando um filtro de bordas (Sobel). A partir das bordas determinam-se regiões. Cada objeto descoberto pelo filtro é considerado uma região. Para

cada regiÃ£o verifica-se a luminância média de sua vizinhança de 4 [18]. Como a vizinhança é fundamental para a conectividade dos *pixels* faz sentido usá-la como um dos avaliadores de distorção. Isso significa que sendo os *pixels* conexos [18], caso um *pixel* varie é esperado que todos os *pixels* que formam sua vizinhança variem de forma uniforme. Caso isso não aconteça há distorção. A partir disso, verifica-se uma determinada vizinhança em dois frames consecutivos.

Como o movimento influencia na percepção de distorção, pesa-se a influencia da distorção pela velocidade do movimento. Considera-se movimento a variação da luminância de uma vizinhança entre dois *frames* consecutivos. Para cada objeto na cena é realizado esse procedimento e para cada *frame* no vídeo. De maneira semelhante ao que acontece no algoritmo 2, cada objeto também é pesado pelo seu valor de luminância média.

ALGORITMO 3:

Primeiro passo: Tome o frame n e aplique um filtro de Sobel sobre a imagem e selecione as regiões delimitadas. Para evitar o ruído gerado pelo filtro, apenas regiões com tamanhos dentro de uma mesma ordem de grandeza. Dentro de uma região determine (x,y) para cada pixel.

Segundo passo: Para cada pixel (x,y) dentro de uma região, determine os pixels (x+1,y), (x-1,y), (x,y+1), (x,y-1).

Terceiro passo: Tome o frame n+1, o próximo frame.

Quarto passo: Tome luminância média dos pixels (x+1,y), (x-1,y), (x,y+1), (x, y-1) e (x,y) do frame n (chame esse valor de Lum(n)).

Quinto passo: Calcule Lum(n) e Lum(n+1).

Sexto passo: Se

$$|(X_n, Y_n) - (X_{n+1}, Y_{n+1})| \leq 2 * |Lum(n) - Lum(n+1)| \quad (1)$$

, considere que há distorção nesse pixel

Sétimo passo: Com o objetivo de pesar a influência do movimento, defina distorção como

$$\frac{|(X_n, Y_n) - (X_{n+1}, Y_{n+1})|}{max(Lum(n), Lum(n+1))} \quad (2)$$

Oitavo passo: Repita isso para cada região do frame e some essas distorções. Faça isso para todos os frames.

4. RESULTADOS

Os experimentos para a avaliação da qualidade consistiram na apresentação de um conjunto de 100 vídeos. Todas as sequências de vídeo estavam no formato CIF (352X288), com um *frame rate* de 25 Hz. A métrica foi comparada com o PSNR e resultados subjetivos. Foram seguidas as especificações dadas pelo VQEG [19] para testes de qualidade e os resultados foram apresentados para 25 espectadores não especialistas. Os espectadores avaliaram o vídeo em uma escala contínua marcada com "Muito bom", "Bom", "Médio", "Ruim"e "Muito ruim". Esses resultados estão quantificados em uma escala de 0 a 100.

Os vídeos estavam divididos em quatro categorias:"movimento", "conferï¿½ncia", "Imagem estática"e "cenas desconexas". Vídeos do tipo "movimento"são aqueles nos quais os objetos de cena estão em movimento, como por exemplo: esportes, movimento de veículos e cenas de ação. Vídeos do tipo conferência são vídeos similares a vídeochamadas e vídeosaulas. Vídeos do tipo imagem estática são a repetição de uma mesma imagem. Vídeos do tipo cena disconexa são aqueles que mudam de cenário repetidas vezes durante a exibição.

Para realizar a comparação, todos os resultados foram normalizados. Os resultados subjetivos foram normalizados usando:

$$R = \frac{|r_i - r_{melhor}|}{|r_{pior} - r_{melhor}|} * 100 \quad (3)$$

onde R é o resultado final indo entre 0 a 100, r_i é a média dos resultados obtidos através dos espectadores, r_{pior} é a pior avaliação do vídeo dada pelos espectadores e r_{melhor} é a melhor avaliação do vídeo dada pelos espectadores.

Para o terceiro algoritmo a operação de normalização é semelhante ao resultado subjetivo. Entretanto, como a avaliação dada por um algoritmo é única, usa-se como melhor caso para avaliaćão o vídeo orginal e como pior caso o vídeo original com todos os *pixels* invertidos.

$$R = \frac{|r_i - r_{original}|}{|r_{invertido} - r_{original}|} * 100 \quad (4)$$

Analogamente para as avaliações feitas em escala logarítimica (como as dadas pelos algortimos 1,2 e o PSNR) a normalização foi:

$$L = \frac{|l_i - l_{original}|}{|l_{invertido} - l_{original}|} * 100 \quad (5)$$

onde L é o resultado final indo entre 0 a 100, l_i é antilog do resultado obtido através da aplicação do algoritmo ao vídeo i, $l_{invertido}$ é o antilog da inversão de todos os pixels do vídeo e $l_{original}$ é antilog do vídeo original.

Em termos de qualidade o objetivo era se aproximar o máximo possível das medidas subjetivas e era esperado que o algoritmo apresentasse melhor desempenho que o PSNR em termos de qualidade e se aproximasse em termos de consumo de recursos.

A comparação dos algoritmos propostos com o PSNR [7] tem como objetivo prover uma base de comparação confiável com outras métricas criadas. Apesar de existirem outras métricas com melhor desempenho do que o PSNR (em termos de correlação com a percepção humana) como no trabalho de Suresh [21], o uso do PSNR (até hoje a técnica mais popular e que reconhecidamente não considera características físicas e psicológicas dos seres humanos) também nos dá base para comparar as vantagens obtidas através de uma métrica com uso do HVS .

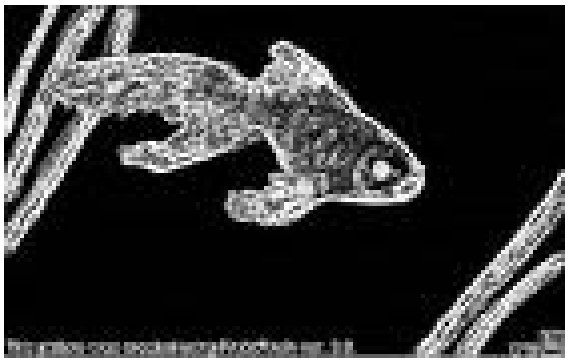

Figura 2: Aplicação do Filtro de Sobel em uma imagem com inúmeros detalhes

O uso da avaliação subjetiva no trabalho, apesar de impraticável em ambientes de tempo real, é importante como objetivo a ser atingido pelos algoritmos. Como visto na seção anterior a distorção e a sensação de erro nas imagens não estão necessariamente associados.

Todos os algoritmos foram programados na linguagem Java. O uso do processador foi medido em MIPS. A implementação do PSNR obteve um valor de 50,17 MIPS.

O primeiro algoritmo teve resultados similares ao PSNR em termos de consumo de recursos (51,07 MIPS), mas se mostrou ligeiramente melhor em termos de qualidade do que o PSNR (cerca de 5 % de todas as amostras estiveram melhor relacionadas com os resultados subjetivos do que o PSNR).

O segundo algoritmo teve maior consumo de recursos (aproximadamente 10 % para cada processador, 54,73 MIPS) e um melhor desempenho de qualidade do que o PSNR (9 % de todas as amostras estão mais próximas dos resultados subjetivos). Por outro lado quando o vídeo possuía um grande número de pequenos detalhes, o algoritmo confundia o ruído com regiões delimitadas pelo filtro, o que gerou resultados muito fracos (aproximadamente 30% piores do que o PSNR). Um exemplo poder ser visto na figura a Figura 2, nas bordas do peixe existem muitas regiões de falha.

O terceiro algoritmo teve uma consumo de recursos ligeiramente maior do que o segundo algoritmo, mas teve resultados excelentes no desempenho da avaliação de qualidade (quase 20% das amostras possuíram resultados mais próximos dos resultados subjetivos, 61,17 MIPS). Como o PSNR possui consome poucos recursos dos dispositivos, o algoritmo 3 ao possuir um custo aproximadamente 20% maior ainda se encaixa em nosso cenário.

A tabela 1 mostra os resultados com as maiores diferenças na avaliação subjetiva (visto através do intervalo de confianą). A primeira coluna mostra quais foram as amostras com que tiveram resultados menos uniformes. A segunda coluna apresenta os resultados subjetivos. A terceira coluna apresenta o intervalo de confiaça de 90 % para esses vídeos. A quarta coluna apresenta o PSNR. A quinta coluna apresenta os resultados do algoritmo 3.

Vídeo	Subjetivo	I.C.	PSNR	Algoritmo 3
31	17	26,91	38	23
32	8	12,43	5	12
40	67	16,52	50	72
43	90	13,88	79	88

Tabela 1: Vídeos com menor uniformidade de resultados

Algoritmo	Mdia	Desvio Padrão	MIPS
PSNR	7,04	5,69	50,17
Primeiro algoritmo	6,95	10,65	51,07
Segundo algoritmo	6,09	5,81	54,73
Terceiro Algoritmo	3,48	2,19	61,17

Tabela 2: comparaão de resultados dos algoritmos

Essa tabela mostra que mesmo em vídeos em que o resultado subjetivo não era uniforme o algoritmo 3 obteve bons resultados. Isso mostra que o algoritmo consegue representar a opinão de um grupo de espectadores sobre um vídeo com certo grau de segurança.

A tabela 2 resume os resultados obtidos por nossos algoritmos. A primeira coluna mostra a mdia das diferenas do algoritmo para o resultado subjetivo. A segunda coluna mostra o desvio padrão dessas diferenas. A terceira coluna mostra o uso do processador (avaliado em MIPS) para cada algoritmo. O consumo de memória não foi considerado porque 64 Mb se provaram suficientes para todos os algoritmos e o uso de telefone.

Como pode ser visto, o algoritmo 3 possui um resultado muito mais próximo do subjetivo do que os demais algoritmos incluindo o PSNR, que teve o pior desempenho. O terceiro algoritmo apresenta resultados piores do que o PSNR e o outros dois algoritmos somente quando há mudança frequente de cenas. Isso ocorre porque o algoritmo perde a referência do último frame. Visto que o algoritmo 3 usa as diferenças prováveis entre 2 frames consecutivos como fator de avaliação da qualidade, a perda do referencial faz com que o algoritmo perca precisão. Dentro do cenário de vídeochamadas essa mudança de cenário praticamente não ocorre. Na tabela 3 podemos ver em que vídeos isso ocorre de forma mais explícita. Observa-se que em alguns deles o algortimo 3 se comporta pior do que o PSNR.

Os algoritmos 2 e 3 possuem valores de desvio padrão maiores do que o PSNR. Para o algoritmo 1, isso aconteceu porque a distorção não se concentra em uma única região, na maior parte dos vídeos avaliados. No algoritmo 2, o algoritmo misturou regiões com ruídos em vídeos com grande número de pequenos detalhes.

Vídeo	Subjetivo	PSNR	Algoritmo 3
10	19	14	12
85	83	69	78
86	83	71	76
99	45	56	57

Tabela 3: Vídeos com cenas desconexas

Para generalizar os algoritmos propostos para a avaliação da qualidade de imagens coloridas, o caminho mais direto é aplicar o mesmo modelo com parâmetros diferentes aos diferentes canais de cores (RGB) respectivamente e então avaliar e combinar os erros nos diferentes canais. Entretanto essa solução não é boa pois os canais RGB são correlacionados. Nos trabalhos Wandell e Poirson [17] sugeriram um novo espaço de cor, chamado de espaço de cor oponente, nas quais as coordenadas seriam perceptualmente ortogonais. As três coordenadas desse sistema corresponderiam a luminescência, vermelho-verde (R/G) e azul-amarelo (B/Y) respectivamente. A mudança do espaço de cor (do RGB, YUV ou YCrCb no espaço oponente) torna mais razoável lidar com cada caminho separadamente. C.J. Van den Branden adotou esse novo espaço de cor na sua métrica de qualidade de imagem [23].

5. CONCLUSÃO

Foi observado que métricas de Referência Completa não são adequadas para a avaliação de mídia de tempo real, visto que não há amostras de referência confiáveis para a avaliação do fluxo. Métricas não referenciadas usam apenas o fluxo recebido como parâmetro para a avaliação e por isso são mais adequadas para cenários de tempo real.

Foram propostos 3 algoritmos que utilizam algumas características HVS de forma a melhorar a avaliação da qualidade. Os dois primeiros algoritmos são algoritmos de Referência Completa e usam características da imagem para a avaliação da qualidade. O terceiro algoritmo é um algoritmo não referenciado e usa a variação da luminância entre frames para avaliar qualidade.

É notório que esses 3 algoritmos possuem complexidade crescente. A maior complexidade indica uma melhor avaliação da qualidade, mas indica também um aumento no uso de recursos computacionais. Também é possível notar que o algoritmo 3 possui a melhor relação entre custo e desempenho dentre os algoritmos apresentados, mesmo sendo um algoritmo não referenciado. Isso ocorre pelo fato do algoritmo 3 usar uma série de características HVS, como um modelo simples de mascaramento, o uso da luminância, movimento e continuidade do vídeo.

Como pôde ser observado, a qualidade de vídeo perceptual não pode ser medida somente em termos dos fatores técnicos como atraso e distorção. Experimentos realizados por mostram que uma perda significativa de pacotes não influencia proporcionalmente a compreensão e percepção de um usuário de um determinado evento apresentado [16]. Isso se deve em parte porque o usuário vai possuir mais tempo para assimilar as informações contidas na imagem. Através dos mesmos experimentos também pode ser observado que cenas extremamente dinâmicas, com grande número de elementos visuais em mudança constante, apesar de possuírem alto custo de representação, possuem um impacto negativo na absorção e compreensão da informação. Usuários tendem em focar sua atenção em um único tipo de mídia (áudio, vídeo ou informação textual), apesar de haverem constantes trocas de foco. Isso implica que se possível que informações muito relevantes para determinada apresentação deveriam ser transmitidas em um único tipo de mídia. Caso isto não seja possível, essas informações concorrentes deveriam ser apresentadas com o máximo de qualidade possível. Como a arquitetura de redes atual é *best-effort* isso pode se tornar bastante complicado.

Ainda pode-se observar que dada as necessidades ergonômicas humanas, existe uma tolerância maior à degradação devido à atrasos e ao *jitter* caso exista pouca dinamicidade no vídeo [14]. Isso significa que a qualidade perceptual estaria mais ligada à variação da imagem do que realmente à transmissão e atraso.

Através desse pensamento, um caminho a ser trilhado é a análise de limites de parâmetros de QoS mais comumente utilizados: *jitter* e atraso. Também é importante ressaltar que novos parâmetros que reflitam a percepção do usuário devam ser criados. O uso conjunto dessas métricas com métricas para a avaliação de qualidade de áudio poderia ser usado para a avaliação de multimídia. O problema de sincronização entre áudio e vídeo é complexo [6] e a falta do *lip sync* certamente impactaria na qualidade. Todas as presentes métricas de qualidade não levam em consideração a sincronização de áudio e vídeo e algumas contribuições podem ser obtidas nessa direção.

O uso de ontologias para a definição de um cenário rico em contextos é uma possibilidade. Dados os parâmetros desenvolvidos para a qualidade de vídeo e sincronia, pode-se vislumbrar que essas informações instanciem uma ontologia. Esse conjunto de informações dá base a um cenário semântico de gerência de rede, baseada nas condições temporais da arquitetura. Outra aplicação interessante seria adequar o vídeo dinamicamente às condições apresentadas pela rede. Através de algum instrumento de inteligência computacional (lógica *fuzzy* ou mesmo uma rede neural), avaliar as condições da rede e tentar antecipar congestionamentos pode ser uma solução bastante elegante para o problema de QoS. Combinar a adaptabilidade do vídeo com a instância das ontologias pode ser uma solução extremamente poderosa.

Existem modificações nos algoritmos que podem ser utilizadas para melhorar a avaliação ou reduzir o consumo de recursos. No segundo e no terceiro algoritmos poderiam ser usados filtros melhores de forma a reduzir o impacto do ruído (como o filtro de Canny [18]), ao custo de um aumento no consumo de recursos. Há ainda a possibilidade de no primeiro e segundo algoritmos poderia-se utilizar somente alguns frames, como os *key frames* por exemplo para a avaliação de qualidade. No terceiro algoritmo, poderia-se utilizar somente alguns pares de *frames*, como *key frames* consecutivos. Isso reduziria o desempenho da avaliação de qualidade, mas reduziria o consumo de recursos.

O terceiro algoritmo em particular ainda apresenta uma característica interessante. Visto que as características de Luminância são contínuas [18], ou seja não haverão saltos de valores que não estejam dentro de um padrão consistente. Tendo esse fato em vista, pode-se supor que haja formas de compensar perdas através de alguma forma de interpolação dos valores dos pixels entre conjuntos de *frames*. Ou seja, dado determinada sequência de frames supõe-se ser possível recuperar informações de Luminância de um *frame* perdido e/ou distorcido através das informações de seus sucessores e antecessores.

Uma outra atividade primordial será verificar a possibilidade de realizar a criação de um modelo de qualidade para vídeo de forma análoga ao modelo E [8] para vídeo, criando as ferramentas necessárias para um modelo multimídia [6]. Um modelo no cenário NGN é um dos objeto de estudo para a área de qualidade de serviço, principalmente modelos que consigam trabalhar as questões de sincronização entre áudio e vídeo.

6. REFERÊNCIAS

[1] A. R. Alves, E. M. Lapolli, R. Bastos, and L. Bastos. Classificação de imagens digitais pelo mtodo da verossimilhana - uma nova abordagem. 7 Simpsio Brasileiro de Sensoriamento Remoto, 1993.

[2] E. C. Carterette and M. P. Friedman. Handbook of perception. volume 5. New York Academic, 1975.

[3] C. H. Chou and Y. C. Li. A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile. volume 5. IEEE Transactions on Circuits and Systems for Video Technology, 1995.

[4] A. M. Eskicioglu and P. S. Fisher. Image quality measures and their performance. volume 43. IEEE Transactions on Communications, 1995.

[5] B. Girod. Digital images and human vision. A. B. Watson, 1993.

[6] D. Hands. A basic multimedia quality model. volume 6. IEEE Transactions on Multimedia, 2004.

[7] Q. Huynh-Thu and M. Ghanbari. Scope of validity of psnr in image/video quality assessment. Arizona, 2008. Electronics Letters 44.

[8] ITU-T. Itu-t recommendation g.107; the e-model, a computational model for use in transmission planning, Mar. 2003.

[9] ITU-T. Recomendação H.323 V6, junho 2006.

[10] J. Rosenberg and J. Rosenberg and G. Camarillo and A. Johnston and J. Peterson and R. Sparks and M. Handley and E. Schooler. SIP: Session Initiation Protocol. RFC 3261, junho 2002.

[11] S. A. Karunasekera and N. G. Kingsbury. A distortion measure for blocking artifacts in images based on human visual sensitivity. volume 4. IEEE Transactions on Image Processing, 1995.

[12] S. A. Karunasekera and N. G. Kingsbury. Image quality measure based on a human visual system model. volume 28. Optical Engineering, 1995.

[13] H. Liu, N. Klomp, and I. Heynderickx. A no-reference metric for perceived ringing. Arizona, 2009. Fourth International Workshop On Video Processing and Quality Metrics.

[14] J. T. M. Claypool. The effects of jitter on the perceptual quality of video. Orlando, 1999. ACM Multimedia.

[15] A. N. Netravali and B. G. Haskell. *Digital Pictures: Representation and Compression*. New York: Plenum, 1988.

[16] E. Peli. Contrast in complex images. volume 7. Journal of Optical Society of America, 1990.

[17] A. B. Poirson and B. A. Wandell. Pattern-color separable pathways predict sensitivity to simple colored patterns. volume 36. Vision Research, 1996.

[18] R. E. W. R. C. Gonzalez. *Digital Image Processing*. Prentice Hall, 2008.

[19] V. Report. Final report from video quality experts group on the validation of objective models of video quality assessment, Oct. 2000. Disponvel em: http://www-ext.crc.ca/vqeg/frames.html.

[20] N. Suresh, P. Mane, and N. Jayant. Real-time prototype of a zero-reference video quality algorithm. Las Vegas, 2008. International Conference on Consumer Electronics.

[21] N. Suresh, P. Mane, and N. Jayant. Testing of a no-reference vq metric: Monitoring quality and detecting visible artifacts. Arizona, 2009. Fourth International Workshop On Video Processing and Quality Metrics.

[22] S. Tao, J. Apostolopoulos, and R. Gurin. Real time monitoring of video quality on ip networks. Washington, 2005. NOSSDAV.

[23] C. J. van den Branden Lambrecht. Color moving pictures quality metric. volume 1. IEEE International Conference on Image Processing, 1996.

[24] B. A. Wandell. Sinauer Associates, 1995.

[25] Z. Wang, L. Lu, and A. Bovik. Video quality assessment based on structural distortion measurement. volume 19. Signaling Processing: Image Communication, 2003.

[26] S. Winkler. *Digital Video Quality Visions, Models and Metrics*. John Wiley, 2005.

[27] F. Yang, S. Wan, Y. Chang, and H. Wu. A novel objective no-reference metric for digital video quality assessment. volume 12. IEEE Signal Processing Letters, 2005.

[28] Q. Zhang, Y. Zhang, and W. Zhu. End-to-end qos for video delivery over wireless internet. volume 93. Proceedings of the IEEE, 2005.