AVALIAÇÃO AUTOMÁTICA DE QUALIDADE DE VIDEOCONFERÊNCIAS DE ALTA DEFINIÇÃO

Cássius Rodrigo Duque Estrada

Dissertação de Mestrado apresentada ao Programa de Pós-graduação em Engenharia Elétrica, COPPE, da Universidade Federal do Rio de Janeiro, como parte dos requisitos necessários à obtenção do título de Mestre em Engenharia Elétrica.

Orientadores: Eduardo Antônio Barros da Silva
Alexandre Gomes Ciancio

Rio de Janeiro
Junho de 2011

AVALIAÇÃO AUTOMÁTICA DE QUALIDADE DE VIDEOCONFERÊNCIAS DE ALTA DEFINIÇÃO

Cássius Rodrigo Duque Estrada

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO INSTITUTO ALBERTO LUIZ COIMBRA DE PÓS-GRADUAÇÃO E PESQUISA DE ENGENHARIA (COPPE) DA UNIVERSIDADE FEDERAL DO RIO DE JANEIRO COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM CIÊNCIAS EM ENGENHARIA ELÉTRICA.

Examinada por:

Prof. Eduardo Antonio Barros da Silva, Ph.D.

Prof. Alexandre Gomes Ciancio, Ph.D.

Prof. Sergio Lima Netto, Ph.D.

Prof. Ronaldo de Freitas Zampolo, D.Sc.

RIO DE JANEIRO, RJ – BRASIL
JUNHO DE 2011

Estrada, Cássius Rodrigo Duque

Avaliação Automática de Qualidade de Videoconferências de Alta Definição/Cássius Rodrigo Duque Estrada. – Rio de Janeiro: UFRJ/COPPE, 2011.

XVI, 82 p.: il.; 29, 7cm.

Orientadores: Eduardo Antônio Barros da Silva
Alexandre Gomes Ciancio

Dissertação (mestrado) – UFRJ/COPPE/Programa de Engenharia Elétrica, 2011.

Referências Bibliográficas: p. 66 – 68.

1. Processamento de Imagem. 2. Avaliação de Qualidade de Vídeo. 3. Vídeoconferência. I. Silva, Eduardo Antônio Barros da *et al.* II. Universidade Federal do Rio de Janeiro, COPPE, Programa de Engenharia Elétrica. III. Título.

À minha querida família

Agradecimentos

Em primeiro lugar, gostaria de agradecer à minha família, pelo apoio incondicional em todos os momentos.

Aos colegas do LPS por estarem sempre dispostos em ajudar. Também a todos colegas que participaram e contribuíram com a realização dos testes de avaliação tão importantes para este trabalho.

Aos meus orientadores, Eduardo Antônio Barros da Silva e Alexandre Gomes Ciancio, pela compreensão, incentivo e orientação.

Resumo da Dissertação apresentada à COPPE/UFRJ como parte dos requisitos necessários para a obtenção do grau de Mestre em Ciências (M.Sc.)

AVALIAÇÃO AUTOMÁTICA DE QUALIDADE DE VIDEOCONFERÊNCIAS DE ALTA DEFINIÇÃO

Cássius Rodrigo Duque Estrada

Junho/2011

Orientadores: Eduardo Antônio Barros da Silva
Alexandre Gomes Ciancio

Programa: Engenharia Elétrica

Neste trabalho é abordado o tema da avaliação objetiva de qualidade em sinais de vídeo, especificamente para sistemas de videoconferência em alta definição simulando a presença de codificação e erros de canal. Começamos por desenvolver um banco de sequências em alta definição com cenas típicas de videoconferência contento uma série de degradações, incluindo as variações de percepção (como diferenças no brilho e contraste) e erros de canal (como perda de macrobloco seguida de métodos de ocultamento de erro). Investigamos o desempenho do estado-da-arte de métricas de qualidade de vídeo com referência completa e propomos uma nova métrica de qualidade com base na abordagem em região de interesse, onde determinadas regiões do vídeo, como faces, são consideradas mais importantes na avaliação da qualidade. Experiências subjetivas foram realizadas para obter o *ground-truth* associado ao banco de sequências a fim de permitir a avaliação de desempenho dos indicadores propostos. Investigações da influência da taxa de quadros, taxa de bits e resolução do vídeo codificado na percepção de qualidade também foram realizadas. Por fim, mostramos que os resultados indicam que a abordagem de região de interesse proposta tem o potencial de ajudar a melhorar a avaliação objetiva de qualidade de sequências de videoconferência.

Abstract of Dissertation presented to COPPE/UFRJ as a partial fulfillment of the requirements for the degree of Master of Science (M.Sc.)

AUTOMATIC HD VIDEO CONFERENCING QUALITY EVALUATION

Cássius Rodrigo Duque Estrada

June/2011

Advisors: Eduardo Antônio Barros da Silva
Alexandre Gomes Ciancio

Department: Electrical Engineering

This work addresses the issue of objective assessment of video signals quality, specifically for video conferencing systems in High Definition in the presence of coding and channel errors. We start by developing a large database with typical high definition sequences containing a variety of degradations, including perceptual variations (such as differences in brightness and contrast) and channel errors (such as macroblock loss followed by error concealment methods). We have extensively investigated the performance of state-of-art video quality full-reference metrics and proposed a novel quality metric based on a region-of-interest (ROI) approach, where certain regions of the video, such as faces, are given more relevance in the assessment of quality. Extensive subjective experiments were conducted to obtain the ground-truth associated with the database in order to allow the performance evaluation of the proposed metrics. Investigation of the influence of the data rate, frame rate and resolution of the video stream in the video's perceived quality was also performed. Finally, we show that our results indicate that proposed ROI approach improves the objective quality assessment of video conferencing sequences.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

ACR	Absolute Category Rating, p. 16
BTFR	British Telecom Full-Reference, p. 27
CSF	Constrast Sensitivity Function, p. 9
DCR	Degradtion Category Rating, p. 15
DSCQS	Double Stimulus Continuous Quality Scale, p. 14
DSIS	Double Stimulus Impairment Scale, p. 15
EPSNR	Edge Peak Signal to Noise Ratio, p. 28
FR	Full-Reference, p. 11
IES	Image Evaluation based on Segmentation, p. 28
ITU	International Telecommunications Union, p. 5
LGN	Lateral Geniculate Nucleos, p. 8
MOS	Mean Opinion Score, p. 5
MSE	Median Squared Error, p. 12
NR	No-Reference, p. 11
NTIA	National Telecommunications and Information Administration, p. 29
PC	Pair Comparison, p. 16
PSF	Point Spread Function, p. 7
PSNR	Peak Signal-to-Noise Ratio, p. 2, 12
RMS	Root mean square, p. 27
ROI	Region of Interest, p. 31

RR	Reduced Reference, p. 11
SSCQE	Single Stimulus Continuous Quality Evaluation, p. 15
SVH	Sistema Visual Humano, p. 6
VQEG	Video Quality Experts Group, p. 3, 25
VQM	Video Quality Model, p. 29

Capítulo 1

Introdução

Vídeos digitais, armazenados em bases de dados de vídeo e distribuídos através de redes de comunicação, estão sujeitos a vários tipos de distorções durante a aquisição, compressão, processamento, transmissão e reprodução. Por exemplo, técnicas de compressão de vídeo com perdas, que são quase sempre usadas para reduzir a largura de banda necessária para armazenar ou transmitir dados de vídeo, podem degradar o sinal durante o processo de quantização. Como outro exemplo, os fluxos de vídeo digital transmitidos através de canais sujeitos a erros, como os canais sem fios, podem ser recebidos de forma imperfeita, devido a danos ocorridos durante a transmissão. Redes de comunicação baseadas em troca de pacotes, tais como a Internet, podem provocar perda ou atraso considerável dos pacotes de dados recebidos, dependendo das condições da rede e da qualidade dos serviços. Todos estes erros de transmissão podem resultar em distorções no vídeo recebido. Como na maioria dessas aplicações o vídeo processado é destinado ao consumo humano, estes irão, em última instância, decidir se a operação foi bem sucedida ou não. Portanto, é essencial para um sistema de serviço de vídeo ser capaz de reconhecer e quantificar a degradação na qualidade do vídeo que ocorre no sistema, para que ele possa manter, controlar e, eventualmente, melhorar a qualidade deste sinal. Uma métrica efetiva de qualidade de vídeo é fundamental para essa finalidade.

Medir a qualidade do vídeo implica em uma comparação direta ou indireta do vídeo de teste com o vídeo original. A maneira mais confiável de avaliar a qualidade de uma imagem ou vídeo é a avaliação subjetiva. O método de avaliação subjetiva consiste na realização de experimentos psico-visuais com indivíduos não-especialistas, a fim de avaliar a qualidade de uma sequência de vídeo. A pontuação média de opinião, que é uma medida subjetiva de qualidade obtida a partir de um certo número de observadores humanos, tem sido considerada por muitos anos como a forma mais confiável de medição de qualidade de vídeo [5]. Porém experimentos de avaliação subjetiva são complicados por vários aspectos tais como: as condições de visualização, habilidade de observação, tradução da percepção de qualidade em uma

escala de pontuação, preferência por um conteúdo específico, adaptação, dispositivos de visualização, níveis de luz ambiente, etc. Também são necessários ambientes e equipamentos adequados, preparação de formulários de avaliação, elaboração das seções de testes e captação de um número razoável de participantes com disponibilidade de tempo para a realização das seções de avaliação e compilação da grande quantidade de dados gerados.

Assim, o objetivo das pesquisas em avaliação objetiva de qualidade de imagem e vídeo é a concepção de métricas de qualidade que possam predizer a qualidade percebida da imagem e do vídeo automaticamente. De modo geral, uma métrica objetiva de qualidade de imagem e vídeo pode ser utilizada com três finalidades:

- Monitorar a qualidade da imagem para os sistemas de controle de qualidade. Por exemplo, um sistema de aquisição de imagem e vídeo pode utilizar métricas de qualidade para monitorar e se ajustar automaticamente para obter a melhor qualidade em dados de imagem e vídeo. Um servidor de vídeo em rede pode analisar a qualidade dos vídeos digitais transmitidos na rede e controlar o fluxo de vídeo.

- Referenciar sistemas e algoritmos de processamento de imagem e vídeo. Se múltiplos sistemas de processamento de vídeo são disponíveis para uma tarefa específica, uma métrica de qualidade pode ajudar a determinar qual deles proporciona os melhores resultados em relação à qualidade.

- Ser incorporada em um sistema de processamento de imagem e vídeo para otimizar os algoritmos e os ajustes de parâmetros. Por exemplo, em um sistema de comunicação visual, uma métrica de qualidade pode ajudar a obter melhores algoritmos de pré-filtragem e de atribuição de bits no codificador ou algoritmos de ocultação de erro e pós-filtragem no decodificador.

A escolha de qual tipo de métrica deve considerar a aplicação e os seus requisitos e limitações. Historicamente, a PSNR (*Peak Signal-to-Noise Ratio*) e outras medidas similares têm sido os critérios de qualidade mais utilizados popularmente por engenheiros e pesquisadores para avaliar e otimizar o desempenho de sistemas de processamento de vídeo e imagem digital. No entanto, estas medidas não possuem alta correlação com o julgamento de qualidade do ser humano [1]. Portanto, há uma necessidade por criar métodos efetivos de avaliação objetiva de qualidade de vídeo.

Durante as últimas duas décadas, métodos de avaliação objetiva de qualidade de imagem e vídeo têm sido extensivamente estudados, e vários critérios têm sido propostos. Dependendo da aplicação e do tipo de informação acessível, a avaliação de qualidade de vídeo pode ser classificada em três tipos: métricas de referência

completa, métricas de referência reduzida, e métricas sem referência [6]. Todos estes tipos de métricas de qualidade visual são agora consideradas por vários grupos de padronização, incluindo o *Video Quality Experts Group* (VQEG) para vídeo e *Advanced Image Coding* (AIC) para imagens. Apesar de melhorias tendo sido alcançadas em relação à PSNR, ainda há uma necessidade de algoritmos mais efetivos e genéricos de avaliação de qualidade de vídeo.

Este trabalho tem como objetivo o estudo de métodos automáticos eficazes de avaliação de qualidade de vídeo para a aplicação específica de videoconferência em alta resolução. Estes métodos podem ajudar a melhorar significativamente sistemas de telepresença avançados. Métricas de avaliação de qualidade objetiva em tempo real podem ser usadas, por exemplo, para modificar a codificação, compressão, ou parâmetros de resolução, considerando, por exemplo, as especificações de *hardware* do usuário final ou a distorção do canal, melhorando significativamente seu desempenho. Outra aplicação de métricas de avaliação de qualidade de vídeo é nos casos em que os erros de transmissão estão presentes e é preciso avaliar a eficácia de técnicas de ocultamento de erro.

Fizeram parte deste estudo a análise e implementação de métodos de avaliação de qualidade considerados estado-da-arte e novas abordagens são propostas para a melhoria do desempenho de métricas de avaliação objetiva de qualidade de vídeos para videoconferência.

Uma das abordagens propostas como contribuição deste trabalho está relaciona à suposição que cada região em uma imagem não pode ter a mesma importância que outras. No entanto, a avaliação de qualidade de imagem baseada em região de interesse permanece relativamente inexplorada. É o avanço nesta área de avaliação de qualidade uma das motivações deste trabalho. Algoritmos de detecção de faces podem ser incorporados ao processo de avaliação, uma vez que as faces são geralmente o principal objeto de interesse em uma videoconferência. Sistemas desenvolvidos tendo em conta métricas de qualidade específicas para a região em torno da face, por exemplo, podem aumentar a qualidade percebida para o usuário final.

Ao final deste trabalho, experimentos de avaliação subjetiva e objetiva foram realizados com o propósito de obter o *ground-truth* associado a sequências de vídeo e confirmação da eficácia dos métodos propostos. Para a realização destes experimentos foram criados dois bancos de sequências de vídeo. Um de sequências de videoconferência de alta definição para investigar a influência de parâmetros perceptuais (brilho e contraste) e erros de transmissão em conjunto com técnicas de ocultamento de erro na qualidade percebida. E um segundo de sequências de alta qualidade para analisar a influência de diferentes configurações de codificação tais como taxa de compressão, taxa de quadros e resolução do vídeo codificado na percepção de qualidade. Os conjuntos de sequências e os seus respectivos conjuntos de

notas subjetivas também são considerados importantes contribuições deste trabalho.

1.1 Organização da Dissertação

Esta dissertação está organizada como segue. No Capítulo 2 são apresentadas algumas características básicas do Sistema Visual Humano que nos ajudam a compreender melhor a percepção de qualidade do ser humano e também por que métodos de avaliação objetiva que consideram tais características possuem bom desempenho. São descritos também os métodos tipicamente usados na avaliação subjetiva de qualidade de vídeo, necessários para a validação de métricas objetivas. São apresentados ainda alguns artefatos comuns introduzidos por sistemas de vídeo digital.

O Capítulo 3 descrevemos as três categorias de avaliação objetiva e as métricas mais utilizadas na área. Também são discutidos os estudos realizados pelo VQEG na avaliação de desempenho de métricas de qualidade de vídeo e apresentamos os métodos do estado-da-arte recomendados pelo grupo na ITU J.144 [2]. Uma destas técnicas apresentadas foi utilizada como base do método proposto e para comparar os resultados obtidos.

No Capítulo 4 são propostas duas abordagens, uma que considera regiões de interesse e outra que considera o efeito da persistência retiniana, a serem incorporadas as métricas do estado-da-arte de avaliação objetiva de qualidade de vídeo.

No Capítulo 5 são descritas as bases de sequências de vídeo criadas e utilizadas neste trabalho, bem como o processo de avaliação subjetiva realizado e o desempenho dos método propostos em relação aos resultados subjetivos. Por fim, no Capítulo 6 são feitas as conclusões finais sobre o trabalho realizado além de serem apresentadas perspectivas para trabalhos futuros.

Capítulo 2

Avaliação de Qualidade de Vídeo

Este trabalho tem como objetivo o estudo sobre o processo de desenvolvimento de um método objetivo de avaliação de qualidade de vídeo para aplicações de videoconferência de alta definição. A primeira etapa para o desenvolvimento de uma métrica objetiva de avaliação de qualidade de vídeo é compreender como o ser humano avalia a qualidade percebida. A qualidade de uma sequência de vídeo é uma característica que mede a degradação percebida, normalmente em comparação a uma sequência ideal ou perfeita. Porém o termo qualidade é definido como um conceito subjetivo que está relacionado diretamente às percepções de cada indivíduo. Portanto os métodos de avaliação subjetiva de sequências de vídeo serão sempre os melhores critérios de medida de qualidade. A *International Telecommunications Union* (ITU) padronizou alguns desses métodos [3, 4] e de acordo com suas recomendações, 15 espectadores devem ser convidados a avaliar um conjunto de sequências de vídeo e os resultados de todos os espectadores são computados como uma pontuação média de opinião (MOS). No entanto, esta abordagem é muito demorada. Em sistemas de transmissão de vídeo, por exemplo, onde a qualidade do sinal recebida pelo consumidor é um ponto crítido no serviço oferecido, a realização desses testes subjetivos é impraticável em determinadas situações, por exemplo quando se deseja ter um controle automático da qualidade do vídeo em relação à taxa de transmissão disponível.

É por este e outros motivos que estudos acerca da automatização deste processo vem sendo realizados nos últimos anos. Medidas objetivas de qualidade são modelos matemáticos que se aproximam destes resultados subjetivos e são baseados em critérios e métricas que podem ser obtidos e avaliados automaticamente por programas de computador. Compreender como o Sistema Visual Humano processa e interpreta os sinais de visuais é um ponto essencial na elaboração de medidas objetivas de qualidade bem correlacionadas com os critérios subjetivos. A seguir, será feita uma breve introdução aos componentes fisiológicos e psicofísicos relevantes do Sistema Visual Humano (SVH). Este estudo irá auxiliar na compreensão melhor dos

critérios utilizados por estes algoritmos de avaliação.

2.1 O Sistema Visual Humano

Não está claramente compreendido como o cérebro humano extrai informações de alto nível cognitivo a partir de estímulos visuais nas fases posteriores à visão, mas os componentes do SVH são bem compreendidos e aceitos pela comunidade da ciência da visão [5]. Nesta seção, apresentamos os aspectos básicos da anatomia e as características psicofísicas do SVH que são considerados relevantes para os algoritmos de processamento de vídeo e, mais especificamente, ao projeto de métricas de qualidade de vídeo.

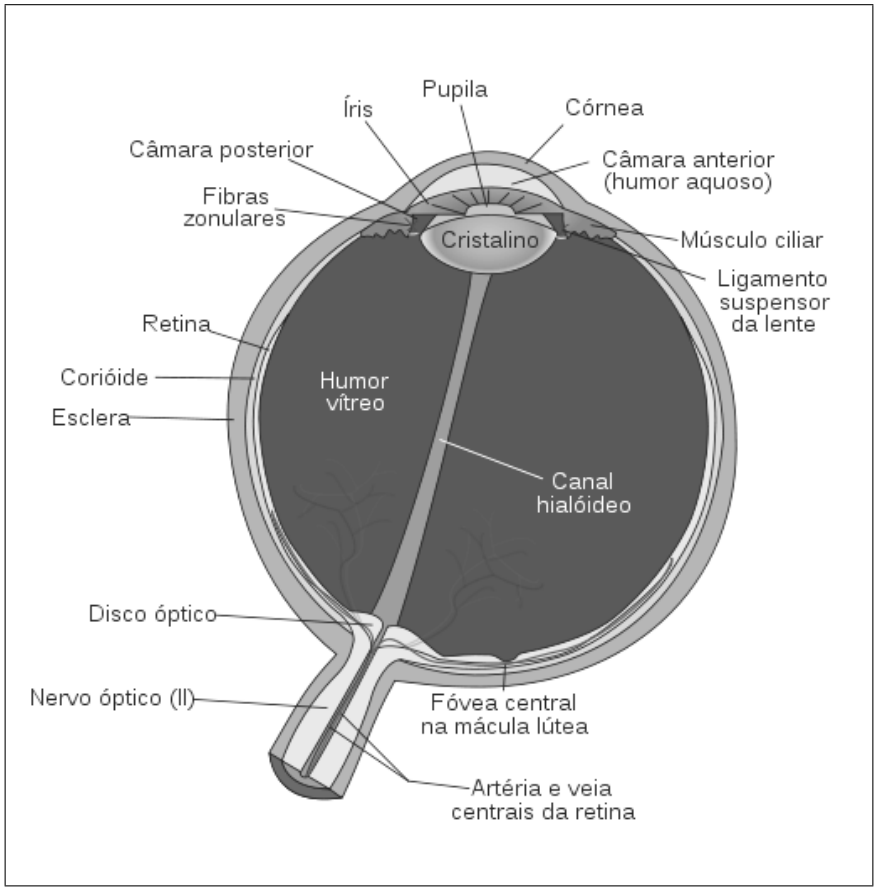


Figura 2.1: O olho humano (seção transversal do olho esquerdo).

2.1.1 A Anatomia do SVH

O estímulo visual em forma de luz proveniente de objetos no ambiente é focado pelos componentes óticos na retina. A retina é uma membrana no fundo dos olhos que contém várias camadas de neurônios, incluindo células fotorreceptoras. A ótica consiste da córnea, da pupila (a abertura que controla a quantidade de luz entrando

no olho), da lente e dos fluidos que preenchem o olho. Uma seção transversal do olho humano é vista na Figura 2.1. O sistema ótico focaliza os estímulos visuais na retina, entretanto, o foco da imagem é imperfeito devido às limitações e imperfeições inerentes a seus componentes. O borrão causado pela perda de foco empresta uma característica passa-baixas ao SVH, tipicamente modelada como um sistema linear invariante no espaço caracterizado por uma função de espalhamento do ponto (*point spread function* - PSF) [5]. Uma aproximação simples da PSF do olho humano, de acordo com Westheimer [5], pode ser vista na Figura 2.2.

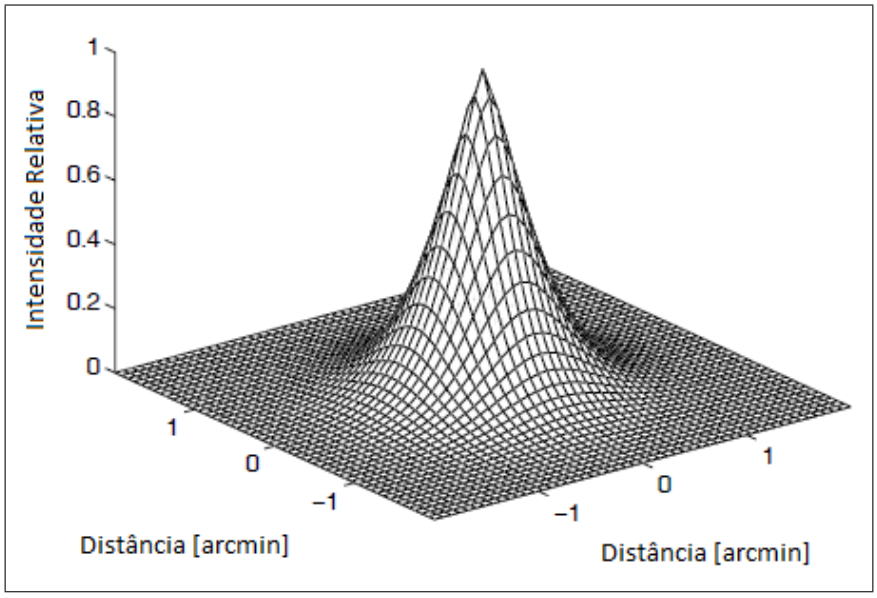


Figura 2.2: Função de espalhamento do ponto do olho humano em função do ângulo visual.

As células fotorreceptoras realizam uma amostragem da imagem projetada na retina. Existem dois tipos de células fotorreceptoras na retina: os cones e bastonetes. Os cones são responsáveis pela visão em condições normais de luz, enquanto os bastonetes são responsáveis pela visão em condições de baixa luminosidade e, portanto, são geralmente ignorados na modelagem do SVH [6]. Existem três diferentes tipos de células cones, correspondendo a três diferentes comprimentos de onda de luz os quais são mais sensíveis. Os L-cones, M-cones e S-cones (com pico de sensibilidade para comprimentos de onda longo, médio e curto, respectivamente) dividem a imagem projetada sobre a retina em três fluxos visuais. Estes fluxos visuais podem ser entendidos como as componentes de cor vermelha, verde e azul do estímulo visual, em uma aproximação rude. Os sinais dos fotorreceptores passam por várias camadas de interconexão neurais na retina antes de serem levados ao cérebro pelo nervo ótico.

As células fotorrecptoras não são uniformimente distribuídas pela superfície da retina. O ponto da retina que se encontra sobre o eixo visual é chamado de fóvea, e possui a maior densidade das células cone. Esta densidade diminui rapidamente a

medida que se afasta da fóvea. A distribuição das células ganglionares, os neurônios que levam o sinal elétrico do olho ao cérebro através do nervo ótico, também é altamente não uniforme, e decai mais rápido do que a densidade dos cones. O efeito resultante é que o SVH não consegue perceber o estímulo visual por completo com uma resolução uniforme.

Os fluxos visuais provenientes do olho são reorganizados no quiasma ótico e no núcleo geniculado lateral (*lateral geniculate nucleos* - LGN), antes de serem retransmitidos para o córtex visual primário. Os neurônios no córtex visual são conhecidos por serem ajustados para diferentes aspectos dos fluxos de entrada, como as frequências espaciais e temporais, orientações e direções de movimento. Normalmente, apenas a frequência espacial e a orientação seletiva são consideradas por métricas de avaliação de qualidade. Os neurônios no córtex têm campos receptivos que são bem aproximados por funções de Gabor bidimensionais. O conjunto destes neurônios é modelado como um banco de filtros de Gabor [7] em oitavas, para os quais, o espectro de frequência espacial (em representação polar) é amostrado em intervalos de oitava ao longo das frequências radiais, e em intervalos uniformes ao longo da orientação [8]. Outro aspecto dos neurônios no córtex visual é a sua saturação em resposta a estímulos de contraste, onde a saída de um neurônio tende a saturar a medida que o contraste da entrada aumenta.

2.1.2 Características Psicofísicas do SVH

Uma série de fenômenos da percepção visual são uma consequência das características do sistema ótico humano. Os fenômenos descritos nesta seção são de especial interesse para a área de processamento de imagem e, mais especificamente para a avaliação de qualidade do vídeo.

Visão Central e Periférica

Como foi dito acima, as densidades de células cone e das células ganglionares da retina não são uniformes, culminando na fóvea e diminuindo rapidamente com a distância em relação a ela. A consequência natural é que sempre que um observador humano se fixa em um ponto no seu meio ambiente, a região em torno do ponto de fixação é visível com a maior resolução espacial, enquanto a resolução diminui com a distância ao ponto de fixação. A visão de alta resolução, devido à fixação do observador a uma região, é chamada de visão central, enquanto a visão progressivamente de menor resolução é chamada de visão periférica. A maioria dos modelos de avaliação de qualidade de imagem trabalham com a visão central e poucos incorporam a visão periférica [6]. Os modelos também podem reamostrar a imagem a fim de reproduzir a mesma densidade de amostragem dos receptores na fóvea, para proporcionar uma

melhor aproximação do SVH, bem como proporcionar uma calibração mais robusta do modelo.

Adaptação à Luz

O SVH opera sobre uma vasta gama de níveis de intensidade luminosa, abrangendo várias ordens de magnitude, indo de uma noite enluarada a um dia ensolarado. Ele é capaz de lidar com uma faixa tão vasta graças a um fenômeno conhecido como adaptação à luz, que ocorre devido ao controle da quantidade de luz que entra no olho através da pupila, assim como devido a mecanismos de adaptação nas células da retina que ajustam o ganho dos neurônios pós-receptores. O resultado é que a retina codifica o contraste relativo do estímulo visual em vez de codificar a intensidade absoluta da luz. O fenômeno que mantém a sensibilidade ao contraste do SVH sobre uma ampla faixa de intensidade de luz de fundo é conhecido como a Lei de Weber [9].

Funções de Sensibilidade ao Contraste

A função de sensibilidade ao contraste (*constrast sensitivity function* - CSF) modela a variação na sensibilidade do SVH a diferentes frequências espacias e temporais que estão presentes no estímulo visual. Esta variação pode ser explicada pelas características dos campos receptivos das células ganglionares e das células do LGN. Consequentemente, alguns modelos do SVH optam por implementar a CSF por meio de fatores de ponderação para sub-bandas de frequências, após uma decomposição. A CSF também varia com a distância à fovea, mas para visão central a CSF espacial é tipicamente modelada como uma função passa-banda invariante ao deslocamento (Figura 2.3). Embora a CSF seja modelável como um filtro passa-banda, a maioria dos algoritmos de avaliação de qualidade implementam uma versão passa-baixas. A sensibilidade ao contraste também é uma função da frequência temporal. Este fato entretanto possui pequena relevância para a avaliação de qualidade de imagens, mas costuma ser usado em modelos para avaliação da qualidade de vídeos usando filtros temporais simples [6].

Mascaramento

Mascaramento é um aspecto importante do SVH na modelagem de interações entre os diferentes componentes da imagem presentes na mesma localização espacial. Mascaramento refere-se ao fato de que a presença de um componente de imagem (chamado de máscara) pode aumentar/diminuir a visibilidade de outro componente da imagem (o chamado sinal de teste). A máscara geralmente reduz a visibilidade do sinal de teste em comparação com o caso que a máscara está ausente. No entanto, a

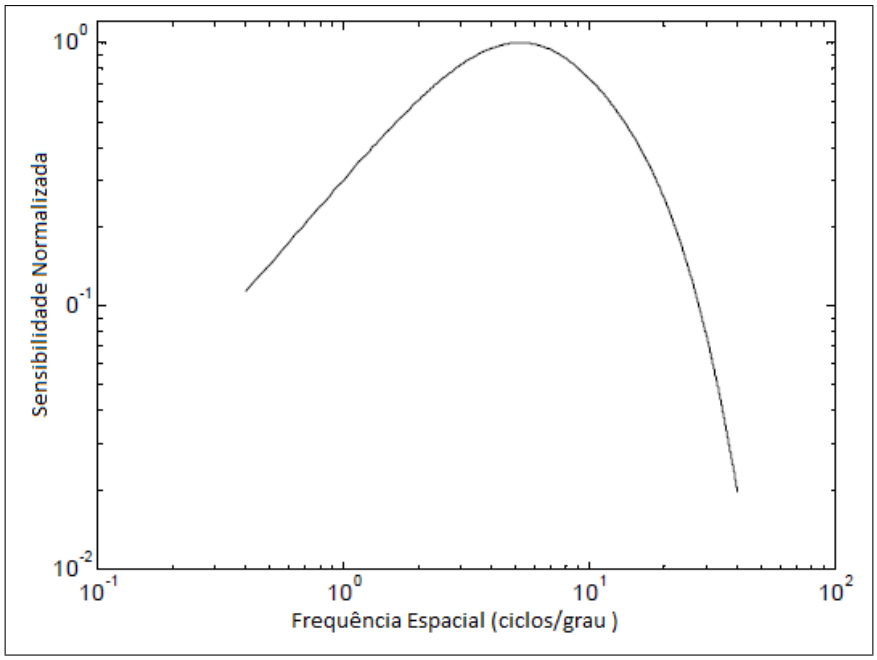


Figura 2.3: Função de sensibilidade ao contraste normalizada.

máscara pode às vezes facilitar também a detecção do sinal de teste. Normalmente, o efeito de mascaramento é mais forte quando a máscara e o sinal de teste têm conteúdo de frequência e orientação similares.

Agrupamento

Agrupamento refere-se à tarefa de chegar a uma única medição de qualidade, ou uma decisão sobre a visibilidade dos artefatos, a partir das saídas dos fluxos visuais. Não é bem entendida a forma como o SVH faz o agrupamento desses fluxos. É bastante óbvio que o agrupamento envolve cognição, onde uma distorção perceptível pode ser mais irritante em algumas áreas da cena (como rostos humanos) do que em outras. No entanto, a maioria das métricas de avaliação de qualidade utilizam a abordagem de Minkowski [10] para agrupar o sinal de erro de frequência e orientação seletiva dos diferentes fluxos para chegar a uma medida de fidelidade. O agrupamento de Minkowski é definido como:

$$M = \frac{1}{N}\sum_{i=1}^{N}(m_i)^p \tag{2.1}$$

onde N é o número de amostras do mapa de qualidade/distorção, m_i é o valor do índice de qualidade na localização espacial $i-$ésima do mapa de qualidade/distorção, e p é a potência de Minkowski. Quando a medida m_i representa a diferença absoluta entre pixels e $p = 1$, então a equação 2.1 se reduz ao erro absoluto médio (MAE). Quando $p = 2$, se torna o erro quadrático médio (MSE). À medida que p aumenta, mais ênfase será dada às regiões de imagem de alta distorção.

Modelos Baseados nas Características do SVH

No último século, o conhecimento sobre o sistema visual humano tem aumentado. Embora muito mais precise ser aprendido antes de podermos afirmar compreendê-lo completamente, o estado-da-arte atual dos mecanismos de processamento e informação visual já são suficientes para fornecer informações importantes que podem ser efetivamente utilizadas no desenvolvimento de métricas de qualidade de vídeo [6]. Na verdade, resultados na literatura demostram que métricas de qualidade de vídeo que usam modelos baseados nas características do SVH possuem melhor desempenho, isto é, fazem previsões da qualidade de vídeo que são mais correlacionadas com os valores dados por observadores humanos [2].

2.2 Avaliação Objetiva

Métricas objetivas de qualidade de imagem e de vídeo podem ser classificadas de acordo com a disponibilidade ou não da imagem e do sinal de vídeo originais, que são considerados de qualidade perfeita e sem distorção, e podem ser usados como referência na comparação com uma imagem ou sinal de vídeo distorcido. A maioria das métricas objetivas de qualidade propostas na literatura assumem que o sinal de referência não distorcido está totalmente disponível. Embora o termo qualidade de imagem e vídeo seja frequentemente usado por razões históricas, o termo mais preciso para este tipo de métrica seria medida de semelhança ou fidelidade da imagem e vídeo, ou avaliação da qualidade de imagem e vídeo com referência completa (*Full-Reference* - FR). É interessante notar que em muitas aplicações práticas de serviço de vídeo, as sequências de vídeo de referência muitas vezes não são acessíveis. Portanto, é altamente desejável desenvolver abordagens de medição que possam avaliar a qualidade de imagem e vídeo cegamente. A avaliação cega ou sem referência (*No-Reference* - NR) da qualidade de vídeo passa a ser uma tarefa muito difícil, apesar de observadores humanos geralmente serem capazes de avaliar de maneira eficaz e fiel a qualidade da imagem ou vídeo distorcido sem a utilização de uma referência. Existe um terceiro tipo de método de avaliação da qualidade de imagem em que a imagem ou sinal de vídeo original não são totalmente disponíveis. Em vez disso, certas características são extraídas do sinal original e transmitidas ao sistema de avaliação de qualidade como informação lateral para ajudar na avaliação de qualidade da imagem ou do vídeo distorcido. Esse método é chamado como avaliação com referência reduzida (*Reduced Reference* - RR). No Capítulo 3 veremos com mais detalhes esses três tipos de métricas e como elas abordam o problema da avaliação objetiva. Isto será complementado com exemplos da literatura.

Atualmente, as métricas de distorção/qualidade de imagem e vídeo mais utiliza-

das do tipo FR são o erro quadrático médio (MSE) e a razão pico de sinal-ruído ($PSNR$), que são definidas como:

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)^2 \tag{2.2}$$

$$PSNR = 10\log_{10}\frac{L^2}{MSE} \tag{2.3}$$

onde N é o número de pixels na imagem ou sinal de vídeo, e x_i e y_i são os i-ésimos pixels nos sinais original e distorcido, respectivamente. L é a faixa dinâmica dos valores de pixel. Para um sinal de 8 bits/pixel, L é igual a 255. Mesmo sendo métricas que não se baseiam nas características do SVH, o MSE e $PSNR$ são amplamente usados por serem simples de calcular, terem claro significado físico, e serem matematicamente fáceis de lidar para fins de otimização (MSE é diferenciável, por exemplo). No entanto na Figura 2.4 podemos ver como diferentes tipos de distorções causam uma diferente percepção de qualidade, porém possuem o mesmo valor de MSE [11]. Alguns dos motivos pelo qual o MSE não ser considerado uma boa medida de qualidade de vídeo serão vistos no Capítulo 3.

Visando solucionar este problema e propor métricas mais correlacionadas com avaliações subjetivas um grande esforço tem sido feito nas últimas três ou quatro décadas para desenvolver métodos objetivos de avaliação de qualidade de imagem e vídeo que incorporam medidas de percepção de qualidade, considerando características do SVH. Iremos descrever brevemente algumas destas abordagens no Capítulo 3.

2.3 Validação de Indicadores de Avaliação de Qualidade

A validação é um passo importante para o desenvolvimento bem sucedido de sistemas de medição de qualidade de vídeo. Desde que o objetivo desses sistemas é prever a qualidade percebida do vídeo, é essencial construir um banco de vídeos com notas de avaliação subjetiva associadas a cada uma das sequências de vídeo no banco de dados. Essa base de dados pode então ser utilizada para avaliar o desempenho de predição dos algoritmos de medida objetiva de qualidade.

2.3.1 Avaliação Subjetiva de Qualidade de Vídeo

Para ser capaz de projetar métricas de qualidade visual confiáveis, é necessário compreender o que qualidade significa para o espectador. A satisfação de um espectador

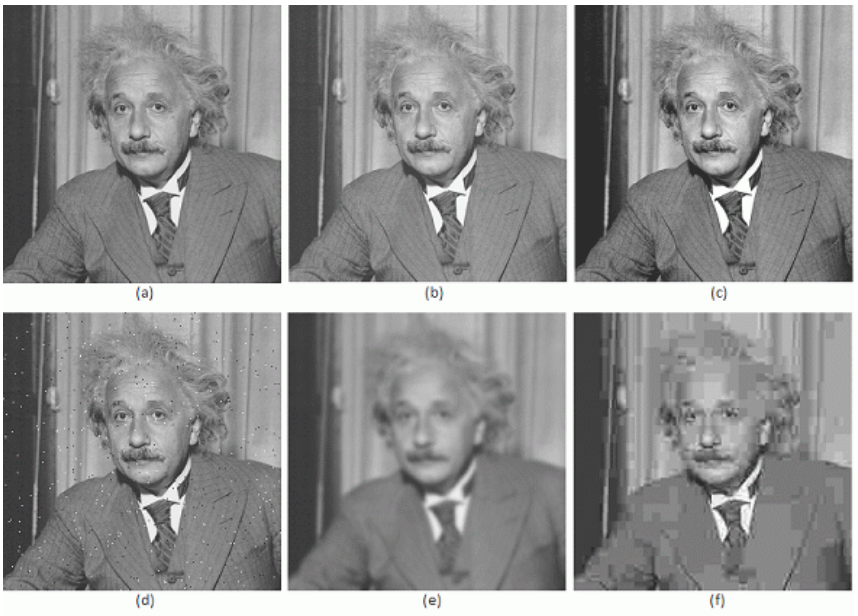


Figura 2.4: Comparação de imagens com diferentes tipos de distorção e todos com MSE = 144 [11]. (a) Imagem original. (b) Imagem com o nível médio de luminância deslocado. (c) Imagem com o contraste ampliado. (d) Imagem contamidada com ruído impulsivo. (e) Imagem desfocada. (f) Imagem comprimida com JPEG.

ao assistir vídeos depende de muitos fatores. Um dos mais importantes é o conteúdo do programa. Por melhor que seja a qualidade visual de um vídeo, este fato não fará diferença caso o conteúdo não seja considerado "assistível "pelos espectadores. Fornecendo um conteúdo por si só "assistível ", a qualidade do vídeo e som desempenham um papel essencial. Pesquisas mostram que a qualidade do vídeo percebido depende da distância de visualização, tamanho da tela, resolução, brilho, contraste, nitidez, cor e outros fatores [5]. Além desses fatores visuais, o som que acompanha também tem grande influência sobre a qualidade do vídeo percebido: classificações subjetivas de qualidade são geralmente mais elevadas quando as cenas de teste são acompanhadas por som de boa qualidade, que aparentemente, reduz a capacidade do espectador para detectar deficiências de vídeo [5].

Também tendemos a esperar qualidades diferentes em situações diferentes. Um exemplo disso é a diferença de espectativa ao assistir um filme no cinema versus um pequeno vídeo em um telefone móvel. Ao mesmo tempo, avanços na tecnologia, como o *Blu-ray* aumentaram as espectativas quanto ao nível da qualidade - um DVD ao qual ninguém teria objeções hà alguns anos atrás, agora é considerado de qualidade inferior por todos que têm um leitor de *Blu-ray* em casa.

Outro ponto importante a ser considerado é que muitas vezes há uma diferença entre fidelidade (a reprodução fiel do original no visor) e qualidade percebida. Imagens nítidas com alto contraste são geralmente mais atraentes para o telespectador médio. Da mesma forma, alguns indivíduos preferem imagens um pouco mais colo-

ridas e saturadas, apesar de perceberem que elas parecem um tanto artificiais.

Todas essas considerações servem para ilustrar a tafera complexa que é medir a qualidade de vídeo percebida. Infelizmente, o nível de qualidade julgado pelo ser humano não pode ser representada por um número exato, devido a sua subjetividade inerente, que só pode ser descrita estatisticamente. Mesmo em experiências de limiar psicofísico, onde a tarefa do observador é apenas dar uma resposta sim ou não, existe uma variação significativa entre as funções de sensibilidade ao contraste dos observadores e outros parâmetros visuais críticos de baixo nível. Além disso, as imagens são utilizadas para uma variedade de propósitos, de modo que o incômodo provocado por determinados artefatos também depende das expectativas do espectador e presunções quanto à aplicação pretendida. Tendo em conta estas advertências, os procedimentos de testes para avaliação da qualidade subjetiva serão discutidos a seguir.

Testes subjetivos para avaliação da qualidade visual foram formalizados nas Recomendações ITU-R BT.500-10 [3] e ITU-T P.910 [4], que sugerem condições padrão de visualização, os critérios de seleção dos observadores e de materiais de ensaio, procedimentos de avaliação e os métodos de análise de dados. Os procedimentos mais utilizados são os seguintes:

- *Double Stimulus Continuous Quality Scale* (DSCQS) [3]. A sequência de apresentação de sinais de vídeo para um teste DSCQS é ilustrada na Figura 2.5. Aos espectadores são apresentados múltiplos pares de sequências que consistem em uma sequência de referência e uma de teste, que são bem curtas (normalmente 10 segundos). A sequência de referência e a de teste são apresentadas duas vezes cada uma de maneira alternada, com a ordem das duas escolhidas aleatoriamente a cada rodada. Os indivíduos não são informados qual é a sequência de referência e qual é a de teste. Eles classificam cada uma das duas separadamente em uma escala contínua de qualidade que varia desde ruim a excelente como mostra a Figura 2.6. A análise qualitativa das avaliações baseia-se na diferença de classificação entre cada par, que é calculada a partir de uma escala numérica equivalente de 0 a 100. Esta diferenciação ajuda a reduzir a subjetividade em relação ao conteúdo da cena e da experiência. DSCQS é o método preferido quando a qualidade da sequência de teste e de referência são semelhantes, por ser mais sensível a pequenas diferenças na qualidade.

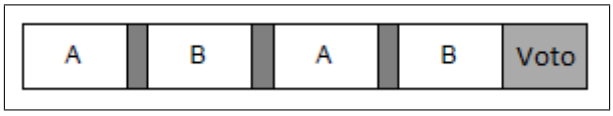


Figura 2.5: Sequência de apresentação para o método DSCQS.

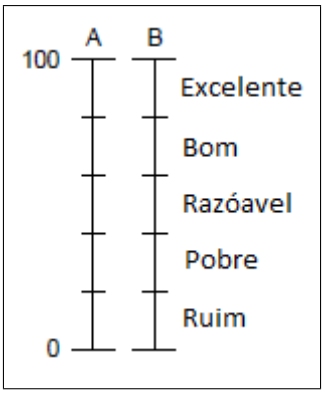


Figura 2.6: Escala de classificação para o método DSCQS.

- *Double Stimulus Impairment Scale* (DSIS) [3] ou *Degradtion Category Rating* [4]. A sequência de apresentação para um julgamento DSIS é ilustrada na Figura 2.5. Ao contrário do método DSCQS, a referência é sempre mostrada antes da sequência de teste, e não se repete. Os observadores classificam a quantidade de distorção na sequência de teste em uma escala de cinco níveis distintos variando de "muito incômodo "a "imperceptível ", como ilustrado na Figura 2.8. O método DSIS é adequado para avaliar deficiências claramente visíveis, tais como artefatos causados por erros de transmissão.

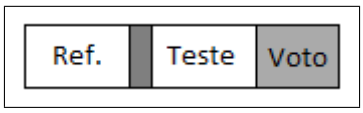


Figura 2.7: Sequência de apresentação para o método DSIS.

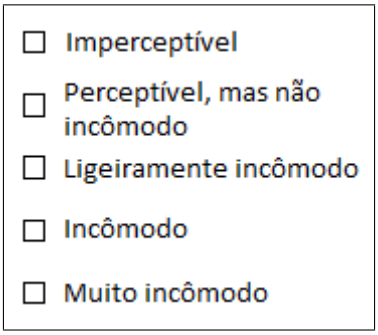


Figura 2.8: Escala de classificação para o método DSIS.

- *Single Stimulus Continuous Quality Evaluation* (SSCQE) [3]. Em vez de ver separadamente pares curtos de sequências, os espectadores assistem a um programa tipicamente de 20 a 30 minutos de duração que foi processado pelo sistema em teste. A referência não é mostrada. Usando uma barra deslizante, indivíduos avaliam continuamente a qualidade percebida instantaneamente na escala DSCQS de ruim a excelente.

- *Absolute Category Rating* (ACR) [4]. Este é um método de estímulo único, onde os telespectadores só observam o vídeo em teste, sem a referência. Eles atribuem uma classificação para a sua qualidade global utilizando uma escala de cinco níveis distintos de ruim a excelente. O fato de que a referência não é mostrada com as sequências de teste, faz do ACR um método mais rápido em comparação com DSIS ou DSCQS, que levam cerca de 2 ou 4 vezes mais tempo, respectivamente.

- *Pair Comparison* (PC) [4]. Para este método, sequências de teste de uma mesma cena, mas em diferentes condições de degradação são exibidos em pares em todas as combinações possíveis, e os observadores fazem um julgamento de preferência para cada par. Isso permite uma discriminação de qualidade muito fina entre as sequências.

Apesar de cada método de avaliação ter suas próprias exigências, as seguintes recomendações são válidas na maioria dos casos:

- A escolha de sequências de teste deve levar em conta o objetivo do experimento. O conteúdo espacial e temporal das cenas, por exemplo, são parâmetros críticos. Estes parâmetros determinam o tipo e a gravidade das deficiências presentes nas sequências de teste.

- É importante que o conjunto de cenas de teste se estenda pela gama de qualidade geralmente encontrada para as condições específicas do ensaio.

- Quando uma comparação entre resultados de diferentes laboratórios é a intenção, é obrigatório o uso de um conjunto de sequências de origem comum para eliminar outras fontes de variação.

- As sequências de teste devem ser apresentadas em uma ordem pseudo-aleatória e, preferencialmente, o pesquisador deve evitar que sequências geradas a partir da mesma referência sejam mostradas em ordem subsequente.

- As condições de visualização, que incluem a distância do olho do observador com o monitor e a luz do ambiente, devem ser definidos de acordo com as normas.

- O tamanho e tipo de monitor ou tela usada no experimento devem ser adequados para a aplicação sob teste. Calibração do monitor pode ser necessária.

- É melhor usar a tela inteira para exibir as sequências de prova. Caso isso não seja possível, as sequências devem ser exibidas em uma janela, com um fundo 50% cinza (Y=U=V=128) ao seu redor.

- Antes de iniciar o experimento, os indivíduos devem ser testados para acuidade visual. Depois disso, instruções escritas e orais devem ser dadas a eles, descrevendo a aplicação pretendida para o sistema, o tipo de avaliação, a escala de opinião e a metodologia de apresentação.

- Pelo menos 15 pessoas devem ser utilizadas no experimento. De preferência, os indivíduos não devem ser considerados "especialistas", isto é, possuir grande conhecimento na área de processamento de imagem e vídeo.

Para conduzir adequadamente os experimentos de avaliações subjetivas é necessário primeiro selecionar, a partir dos diferentes métodos disponíveis, aquele que melhor se adequam aos objetivos e circunstâncias do problema de avaliação em foco. Para isso devemos entender as principais causas de degração em sinais de vídeo digitais e os efeitos na percepção de qualidade que eles produzem.

2.4 Artefatos Comuns em Sistemas de Vídeo Digital

Fluxos de vídeo comprimidos são destinados principalmente para a transmissão por redes de comunicação. Mas existem diversos tipos de meios de comunicações por vídeo. Cada um tem condições e propriedades de funcionamento específicos. Para o caso da comunicação por videoconferência abordada neste trabalho, os canais utilizados podem suportar uma taxa de transmissão constante ou variável de bits e devem fornecer alguma forma de qualidade de serviço (QoS). Por último, a transmissão pode ser ponto a ponto, *multicast* ou por difusão. Na maioria dos casos, depois que o vídeo foi digitalmente comprimido, o fluxo contínuo de dados resultante é segmentado em pacotes fixos ou variáveis e multiplexados com outros tipos de dados, como áudio. A próxima etapa é a codificação de canal, que irá adicionar códigos corretores de erro aos dados. A principal característica de uma aplicação de videoconferência, é claro, ter um grande impacto sobre a qualidade do vídeo exibido no receptor.

A degradação é uma propriedade do vídeo que é considerada indesejável, quer seja no original ou não. Distorções podem ser introduzidas durante a captura, transmissão, armazenamento e/ou exibição, bem como por qualquer algoritmo de processamento (compressão, por exemplo) que pode ser aplicado ao longo do processo. Degradações podem ser muito complexas em suas descrições físicas e também nas suas descrições perceptivas. A maioria delas tem mais de uma característica perceptual, mas é possível ter distorções que são relativamente simples. Para diferenciar as distorções de suas características perceptivas vamos usar os termos "artefato "para

nos referirmos às características de percepção das distorções e “sinal de degradação ”para se referir ao sinal físico que produz o artefato. Os artefatos mais comuns presentes em vídeos digitais são:

- Efeito de bloco - Artefato caracterizado por um padrão visível de blocos na imagem. Deve-se à quantização independente de blocos individuais (geralmente de tamanho 8x8 pixels) em esquemas de codificação baseadas em blocos DCT, levando a descontinuidades nas fronteiras dos blocos adjacentes. O efeito de bloco geralmente é o artefato mais visível em um vídeo comprimido, dada a sua periodicidade e a extensão do padrão. Codificadores mais modernos, como o H.264, usam um filtro de *deblocking* para reduzir o incômodo causado por este artefato.

- *Blur* ou embaçamento - É caracterizado por uma perda de detalhe espacial e uma redução da nitidez das bordas. Uma das causas de embaçamento está relacionado à fase de compressão, onde embaçamento é introduzido pela supressão dos coeficientes de alta frequência no processo de quantização. Em sequências de vídeo, alterações devidas à movimentação rápida durante a gravação de um único quadro também causam o efeito de embaçamento (borão de movimento).

- *Color Bleeding* - Caracteriza-se por uma mancha de cores entre diferentes áreas de forte crominância. É o resultado da supressão de coeficientes de alta frequência das componentes de croma. Devido à subamostragem da croma, o *color bleeding* se estende sobre um macrobloco inteiro.

- Efeito escadas - Esse artefato ocorre como consequência do fato de que codificações baseadas em DCT são mais adequadas para a representação de linhas horizontais e verticais. A representação de linhas com outras orientações exigem maiores coeficientes da DCT para a reconstrução exata. Portanto, quando as frequências mais altas são perdidas, linhas inclinadas (como escadas) podem aparecer como escadas.

- *Ringing* - Associado ao fenômeno de Gibbs. É mais evidente ao longo das bordas de alto contraste do que em áreas de superfície lisa. É uma consequência direta da quantização levando a irregularidades de alta frequência na reconstrução. O efeito de *ringing* ocorre com as componentes de iluminação e cor.

- Ruído tipo Mosquito - Artefato temporal que é visto principalmente nas regiões com textura como flutuações (mosquitos) de luminância/crominância em torno das bordas de alto contraste ou objetos em movimento. É uma consequência das diferenças de codificação para a mesma área de uma cena em quadros consecutivos de uma sequência.

- Cintilação - Ocorre quando uma cena tem um conteúdo de textura elevado. Blocos com textura são compactados com fatores de quantização variados ao longo do tempo, o que resulta em um efeito de cintilação visível.

- Perda de pacotes - Ela ocorre quando os pacotes que fazem parte do vídeo são perdidos na transmissão digital. Como consequência, as partes (blocos) do vídeo ficam faltando para vários quadros.

- *Aliasing* - Pode ser notado quando a frequência de amostragem, espacial ou temporal, é inferior à taxa de Nyquist do conteúdo da cena.

O desempenho de um sistema de vídeo digital em particular pode ser melhorado se o tipo de artefato que afeta a qualidade do vídeo é conhecido. Este tipo de informação também pode ser usada para melhorar o vídeo, reduzindo ou eliminando os artefatos identificados. Em resumo, este conhecimento torna possível a implementação de um sistema completo para detectar, estimar e corrigir artefatos em sequências de vídeo. O objetivo desta dissertação é exatamente buscar desenvolver mecanismos que auxiliem no aperfeiçoamento de sistemas de transmissão de vídeo digital, mais especificamente sistemas de videoconferência.

Neste capítulo, foram introduzidos diversos aspectos da qualidade de vídeo. A anatomia do SVH foi descrita e uma série de fenômenos da percepção visual que são de particular relevância para a qualidade do vídeo foram discutidas. Vimos porque métodos objetivos que não consideram o SVH não possuem bom desempenho em avaliação de qualidade de vídeo. Descrevemos métodos subjetivos para avaliar a qualidade de vídeo e as técnicas mais comuns padronizadas pela ITU. E por último apresentamos uma breve introdução às principais características dos sistemas de vídeo digital, com foco nos erros (artefatos) comumente presentes em aplicações de vídeo digital. No Capítulo 3 apresentaremos com mais detalhes alguns algoritmos de avaliação objetiva de qualidade de vídeo, assim como as principais ideias utilizadas no desenvolvimento destes métodos, exemplificando um conjunto representativo dos tipos de métricas FR, RR e NR.

Capítulo 3

Algoritmos de Avaliação de Qualidade de Vídeo

Para o desenvolvimeto de uma métrica objetiva de avaliação de qualidade é importante pesquisar e compreender os principais conceitos, idéias e abordagens utilizados em métodos de avaliação objetivas. Neste capítulo descrevemos as três categorias de métricas objetivas (FR, RR e NR) já mencionadas no Capítulo 2, além de citar exemplos conhecidos e utilizados pela comunidade científica.

Uma imagem ou sinal de vídeo cuja qualidade está sendo avaliada pode ser interpretada como uma soma de um sinal de referência perfeito e um sinal de erro. Podemos supor que a perda de qualidade está diretamente relacionada com a intensidade do sinal de erro. Portanto, um caminho natural para avaliar a qualidade da imagem é quantificar o erro entre o sinal distorcido e o sinal de referência, que está totalmente disponível na avaliação de qualidade com referência completa. A implementação mais simples desse conceito é o MSE como consta na equação (2.2). No entanto, há uma série de razões pelas quais o MSE pode não se correlacionar tão bem com a percepção humana da qualidade:

- A escala de valores de um pixel digital, em que se baseia o cálculo do MSE normalmente, pode não representar exatamente o estímulo de luz que é recebido pelo olho humano.
- A sensibilidade do SVH a erros pode ser diferente para variados tipos de sinais de erro, e também pode variar com o contexto visual. Essa diferença pode não ser capturada adequadamente pelo MSE.
- Dois sinais de imagem distorcida com a mesma quantidade de energia de erro podem ter diferentes tipos de sinais de erro, conforme visto no Capítulo 2, Figura 2.4.

- O somatório simples de sinais de erro, como é implementada na formulação do MSE, pode ser muito diferente da maneira como o SVH e o cérebro chega a uma avaliação de distorção perceptível [6].

Nas últimas três décadas, a maioria das propostas de métricas de qualidade de imagem e vídeo tem tentado aperfeiçoar o MSE, abordando as questões acima. Tais propostas seguiram o paradigma baseado na sensibilidade ao erro, que tenta analisar e quantificar o sinal de erro de uma maneira que simule as características da percepção visual humana de erro.

Parte dos esforços da comunidade científica tem se centrado no problema da avaliação da qualidade de imagem, e só recentemente a avaliação da qualidade de vídeo tem recebido mais atenção [7]. As atuais métricas de avaliação de qualidade de vídeo utilizam modelos do SVH semelhantes aos utilizados em muitas métricas de avaliação de qualidade de imagem, com extensões necessárias para incorporar os aspectos temporais do SVH [6].

Como mencionado no Capítulo 2, métricas de qualidade de vídeo podem ser divididas em três categorias diferentes de acordo com a disponibilidade do sinal de vídeo original (de referência):

- Métricas de Referência Completa (*Full Reference metric* - FR) – os sinais de vídeo original e distorcido estão disponíveis.

- Métricas de Referência Reduzida (*Reduced Reference metric* - RR) – Além do sinal de vídeo distorcido, uma descrição do sinal original e alguns parâmetros estão disponíveis.

- Métricas Sem Referência (*No-reference metric* – NR) – Apenas o sinal de vídeo distorcido está disponível.

As Figuras 3.1, 3.2 e 3.3 retratam os diagramas de blocos correspondentes às métricas de qualidade de vídeo de referência completa, referência reduzida e sem referência, respectivamente. Observa-se que na abordagem FR o sinal completo da referência está disponível no ponto de medição. Na abordagem RR apenas uma parte do sinal referência está disponível através de um canal auxiliar. Neste caso, as informações disponíveis no ponto de medição geralmente consistem de um conjunto de características extraídas da referência. Para a abordagem NR nenhuma informação sobre a referência está disponível no ponto de medição.

Estas três classes de métricas são direcionadas para diferentes aplicações. Métricas FR são mais adequadas para as medições de qualidade *off-line*, para a qual uma medição detalhada e precisa da qualidade do vídeo é de maior prioridade do que ter resultados imediatos. Métricas RR e NR são direcionadas a aplicações em

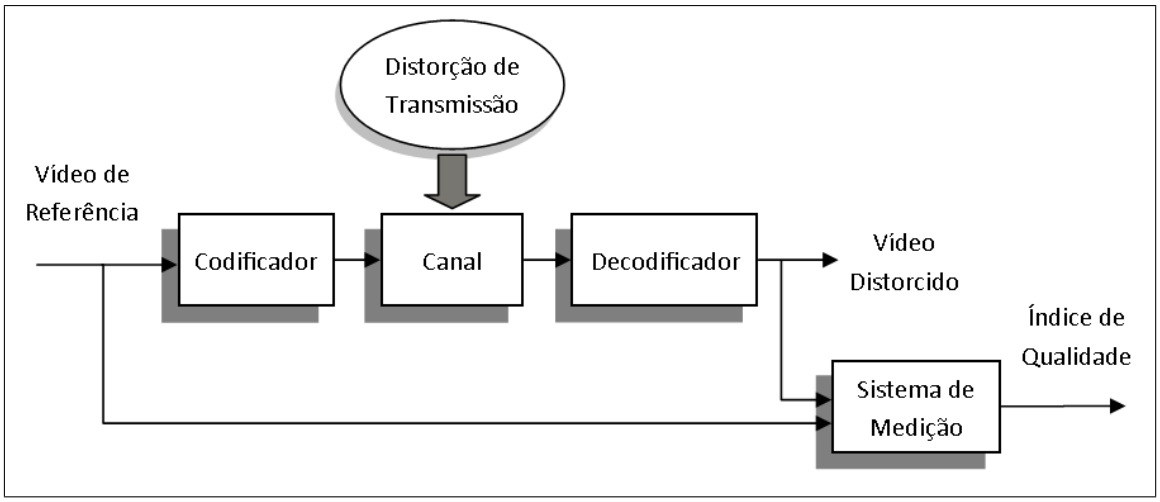


Figura 3.1: Diagrama de blocos de um sistema de medida de qualidade de vídeo com referência completa.

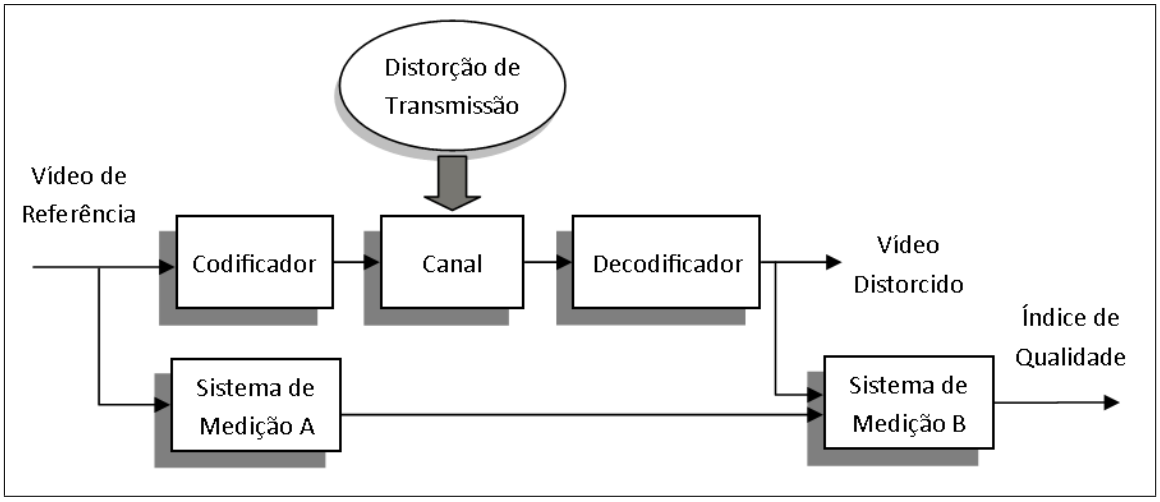


Figura 3.2: Diagrama de blocos de um sistema de medida de qualidade de vídeo com referência reduzida.

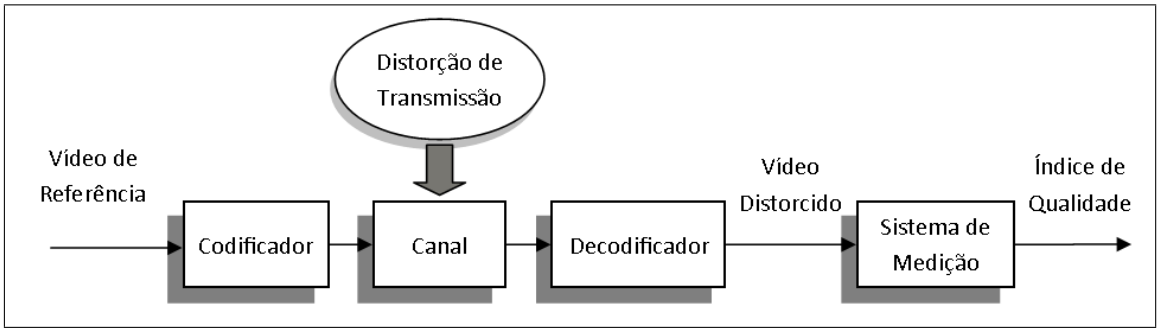


Figura 3.3: Diagrama de blocos de um sistema de medida de qualidade de vídeo sem referência.

tempo real, onde as limitações de complexidade computacional e da falta de acesso a referência são as principais restrições.

As métricas de qualidade também podem ser classificadas de acordo com a abor-

dagem que elas realizam para estimar a quantidade de distorção em um vídeo. Existem basicamente duas abordagens principais. A primeira é a abordagem de sensibilidade ao erro que tenta analisar as diferenças visíveis entre os vídeos de teste e de referência. Esta abordagem é mais usada para métricas de referência completa, já que este é o único tipo de métrica onde a diferença pixel a pixel entre os sinais originais e de teste podem ser geradas.

A segunda abordagem é a técnica de extração de características que compara características de nível superior que não pertencem ao vídeo original para obter uma estimativa da qualidade do vídeo. Métricas sem referência ou de referência reduzida frequentemente utilizam a abordagem por extração de características fazendo uso de algum conhecimento a priori das características do vídeo original.

Finalmente, métricas de qualidade podem ainda ser classificadas de acordo com o tipo de informação que consideram quando processam o vídeo. Métricas que levam em conta caractéristicas do SVH normalmente são chamadas de métricas perceptuais. Métricas mais simples que apenas medem a fidelidade do sinal de vídeo sem considerar seu conteúdo são chamadas de métricas de dados O MSE e o PSNR são exemplos de métricas de dados.

3.1 Métricas de Referência Completa de Qualidade de Vídeo

Em geral, métricas de referência completa têm o melhor desempenho entre os três tipos de métricas. Isto é devido principalmente à disponibilidade do vídeo de referência. Além disso, como métricas FR são destinadas a aplicações *off-line*, elas podem ser mais complexas computacionalmente e incorporar vários aspectos do SVH. A principal desvantagem da abordagem com referência completa é o fato que um alinhamento espacial e temporal muito preciso entre os vídeos de referência e distorcidos é necessário para garantir a precisão da métrica. Um grande número de métricas FR são baseadas na sensibilidade ao erro, que tenta analisar e quantificar o sinal de erro de uma maneira que simula o julgamento de qualidade do ser humano. Alguns exemplos são as métricas conhecidas como a *Visible Differences Predictor* (VDP) de Daly [12], o modelo *Sarnoff Just Noticeable Difference* (Sarnoff JND) de Lubin [13], o *Structural Similarity* (SSIM) de Wang *et al* [11] e o *Perceptual Distortion Metric* (PDM) de Winkler [14].

3.1.1 *Structural Similarity* - SSIM

A métrica *Structural Similarity* (SSIM) se baseia na idéia de que imagens naturais são altamente estruturadas. Em outras palavras, seus pixels apresentam forte

dependência, especialmente quando eles estão espacialmente próximos, e essas dependências carregam informações importantes sobre a estrutura dos objetos na cena [11].

Para estimar a similaridade entre uma imagem de teste e de referência, o algoritmo SSIM compara a luminosidade $l(x, y)$, contraste $c(x, y)$, e estrutura $s(x, y)$ da imagem de teste y e de referência x, usando as seguintes expressões:

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

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

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

onde C_1, C_2 e C_3 são constantes dadas por $C_1 = (K_1 \cdot L)^2$, $C_2 = (K_2 \cdot L)^2$, $C_3 = C_2/2$. L é a faixa dinâmica dos valores de pixel (para imagens com 8 bits/pixel, $L = 255$), $K_1 << 1$, e $K_2 << 1$. Comumente, $K_1 = 0.01$ e $K_2 = 0.03$. As constantes C_1, C_2 e C_3 evitam o surgimento de instabilidades quando o denominador tende a zero. A fórmula geral da métrica SSIM é dada por:

$$SSIM(x, y) = [l(x, y)]^\alpha \cdot [c(x, y)]^\beta \cdot [s(x, y)]^\gamma$$

onde, α, β e γ são parâmetros que definem a importância relativa das componentes luminância, contraste e estrutura, respectivamente. Se $\alpha = \beta = \gamma = 1$, a equação acima se reduz a:

$$SSIM(x, y) = \frac{(2\mu_x\mu_y + C_1)(2\sigma_{xy} + C_2)}{(\mu_x^2 + \mu_y^2 + C1)(\sigma_x^2 + \sigma_y^2 + C2)}$$

O SSIM tem uma faixa de valores variando de 0 a 1, sendo 1 o melhor valor possível. Uma implementação em linguagem C da métrica SSIM está disponível em [15].

3.2 Métricas de Referência Reduzida de Qualidade de Vídeo

Métricas de referência reduzida de qualidade de vídeo exigem apenas informação parcial sobre o vídeo de referência. Para auxiliar a avaliar a qualidade do vídeo, alguns recursos ou medidas físicas são extraídas da referência e transmitidas para o receptor como informação lateral. Uma das características interessantes de métricas do tipo RR é a possibilidade de escolher a quantidade de informação secundária.

Na prática, a quantidade exata de informação será limitada pelas características do canal lateral, que é utilizado para transmitir esses dados auxiliares. Tipicamente as taxas de bits do canal de referência reduzida podem ir de zero (métricas sem referência) a 15 kbps, 80 kbps ou 256 kbps [16]. Métricas desta classe podem ser menos precisas do que as métricas de referência completa, mas elas também são menos complexas, e fazem as implementações em tempo real mais acessíveis. No entanto, a sincronização entre os dados originais e distorcidos ainda é necessária. Trabalhos desta área incluem o trabalho de Webster [17] e Gunawan e Ghanbari [18].

3.3 Métricas Sem Referência de Qualidade de Vídeo

Exigir o vídeo de referência ou mesmo uma pequena parcela se torna um sério obstáculo em muitas aplicações de transmissão em tempo real. Neste caso, torna-se fundamental desenvolver formas de estimar a qualidade de um vídeo cegamente com uma métrica sem referência. Ocorre que, apesar de observadores humanos geralmente poderem avaliar a qualidade de um vídeo sem usar uma referência, criar uma métrica sem referência é uma tarefa muito difícil. Com excessão da métrica de Gastaldo *et al* [**?**] que utiliza uma rede neural, a maioria das métricas propostas são baseadas em extração de características que estimam aspectos do vídeo. Devido às dificuldades encontradas na elaboração de métricas sem referência, várias métricas dependem de uma ou duas características para estimar a qualidade. Na maioria dos casos, os recursos utilizados nos algoritmos são medidas de artefatos, sendo os mais populares o efeito de bloco, o embaçamento e o *ringing*. Um exemplo é a métrica elaborada por Farias e Mitra [19] que utiliza as medidas de quatro artefatos na estimação de qualidade perceptiva. Também vale citar o trabalho realizado em [20], onde uma rede neural é utilizada para avaliar embaçamento em imagens estáticas.

Neste trabalho concentramos os estudos em métricas FR, como base para nossa proposta. Na próxima seção veremos as principais pesquisas nesta área assim como métodos estado-da-arte em avaliação objetiva do tipo FR para sequências de vídeo.

3.4 *Video Quality Experts Group* (VQEG)

O VQEG foi formado em 1997 com o objetivo de coletar conjuntos de avaliações subjetivas confiáveis para um conjunto bem definido de sequências de teste e avaliar o desempenho de diferentes sistemas de avaliação de qualidade de vídeo objetivos em relação a essas sequências. O grupo é composto por peritos em matéria de avaliação de qualidade de vídeo da indústria, universidades e organizações internacionais.

A ênfase da Fase I [1] de estudos do VQEG foi baseada em testes de referência completa (a sequência de referência está disponível para a métrica) de vídeos de produção e distribuição para a televisão. Portanto, as condições de teste consistiram, principalmente em, sequências codificadas com MPEG-2 com diferentes perfis, níveis e variações de outros parâmetros, incluindo a concatenação de codificadores, conversões entre vídeos analógicos e digitais, e erros de transmissão. Um conjunto de cenas de 8 segundos com características diferentes (por exemplo, detalhes do espaço, movimento, cor) foram selecionadas por laboratórios independentes. As cenas foram divulgadas aos proponentes apenas após a apresentação de suas métricas. No total, 20 cenas foram codificados para 16 condições de ensaio cada. Avaliações subjetivas para essas sequências foram obtidas em experimentos de grande escala, utilizando o método DSCQS da Recomendação ITU-R BT.500-10 [3].

Os proponentes de métricas de qualidade de vídeo para o teste VQEG foram, entre outros, o CPqD (Brasil), EPFL (Suíça), KDD (Japão), a KPN *Research / Swisscom* (Holanda / Suíça), a NASA (EUA), NHK / Mitsubishi (Japão), NTIA / ITS (EUA), TAPETES (UE) *Technische Universitat Braunschweig* (Alemanha), e *Tektronix / Sarnoff* (EUA).

Três atributos da capacidade dessas métricas em predizer avaliações subjetivas foram avaliados: precisão, monotonicidade e consistência das previsões da métrica. Os métodos estatísticos utilizados para a análise desses atributos foram regressão com variância ponderada, regressão linear, correlação com a ordem de classificação Spearman, e o *outlier ratio*. Os resultados dos dados analisados mostraram que o desempenho da maioria dos modelos, bem como PSNR, são estatisticamente equivalentes para todos os três critérios, levando à conclusão de que nenhum modelo supera os demais em todos os casos e para toda a gama de sequências de teste. Os resultados são descritos em detalhe no relatório final [1].

Após esta primeira fase, o VQEG realizou uma segunda rodada, Fase II [21], de testes de métricas de referência completa concluida em 2003. A fim de obter resultados mais exigentes, esta segunda fase foi projetada com um forte foco em distribuição secundária de vídeo de qualidade de televisão digitalmente codificados e uma vasta gama de distorções. Novas sequências fonte e condições de teste foram definidos, e um total de 128 sequências de teste foram produzidas. Avaliações subjetivas para essas sequências foram novamente coletadas utilizando o método DSCQS [3].

Os proponentes dessa Fase II foram a *British Telecom* (Reino Unido), *Chiba University* (Japão), o CPqD (Brasil), a NASA (EUA), NTIA / ITS (EUA) e *Yonsei University* (Coréia). Sete critérios estatísticos foram definidos para analisar o desempenho da previsão das métricas. Todos os critérios produziram a mesma classificação de métricas, portanto apenas as correlações são citadas aqui. As melhores métricas nos testes alcançaram correlação tão alta quanto 94% com MOS melho-

rando assim significativamente o desempenho em relação ao PSNR, que apresentou uma correlação de cerca de 70% com as notas subjetivas. Os resultados deste teste são a base da recomendação ITU-T Rec. J.144 [2]. A seguir fazemos uma breve descrição de cada uma dessas métricas.

3.4.1 *British Telecom Full-Reference* - BTFR

O modelo de referência completa da *British Telecom*, BTFR, tenta simular digitalmente características do sistema visual humano (SVH) para dar previsões exatas de qualidade de vídeo [2]. O modelo compara somente aspectos dos sinais que são perceptualmente relevantes ao usuário. Consiste basicamente em duas etapas primárias, detecção e integração. A fase de detecção busca nas regiões do sinal de vídeo degradado identificar a região de melhor casamento no sinal de referência. Propriedades como o PSNR, PSNR das componetes de cor e diferenças de complexidade espacial são extraídas. O estágio final do modelo, a fase de integração, regride os parâmetros extraídos da detecção com uma função linear e forma uma predição da avaliação subjetiva para cada par de sinais de vídeo de referência e distorcido.

O modelo BTFR apresentado pela *British Telecom* para os testes do VQEG produziu para imagens no padrão NTSC correlações com os valores subjetivos de 0,937 com os dados da avaliação subjetiva e um erro RMS de 0,075 [2]. Além de alcançar uma correlação maior que o PSNR (0,804), o modelo apresentou um dos melhores resultados entre os modelos avaliados.

3.4.2 *Edge Peak Signal to Noise Ratio* - EPSNR

Este modelo foi proposto pela Universidade de Yonsei e a *South Korean Telecom*. É uma tentativa de analisar as sequências de vídeo, aproveitando a sensibilidade do sistema visual humano a degradações de borda, o que dá aos quadros de imagem uma aparência de desfocada [2].

As sequências de vídeo, após a compressão, tendem a ter uma grande quantidade de ruído introduzidas nelas. O modelo é baseado na comparação entre os pixels de borda dos quadros da sequência de vídeo processada ao de referência. Algoritmos de detecção de borda são aplicados às sequências para localizar as áreas de borda. Um limiar de borda é estabelecido como critério de escolha de quais pixels da detecção fazem parte das bordas da imagem. A degradação em torno das bordas é medida pelo cálculo do erro médio quadrático ao longo das bordas, *Edge Peak Signal to Noise Ratio* (EPSNR). Este EPSNR é calculado e utilizado como métrica de qualidade de vídeo após o pós-processamento. A fase de pós-processamento do modelo leva em conta imagens desfocadas seriamente. Estas tendem a ser imagens que tenhem uma maior quantidade de arestas removidas ou deslocadas devido à compressão [2]. O

modelo reduz o seu limiar de borda para compensar essas imagens.

O modelo EPSNR obteve um coeficiente de 0,857 de correlação com a avaliação subjetiva nos testes do VQEG [2]. Mesmo não sendo um dos mais altos entre os modelos avaliados, é o de processamento mais rápido e de menor complexidade computacional.

3.4.3 *Image Evaluation based on Segmentation* - IES

O modelo do CPqD apresentado para a Fase II de esperimentos do VQEG foi nomeado de CPqD-IES [2]. O CPqD-IES implementa a avaliação da qualidade de vídeo usando parâmetros objetivos com base na segmentação de imagens. Cenas naturais são segmentadas em regiões planas, bordas e texturas e um conjunto de parâmetros objetivos é atribuído a cada um destes contextos.

O algoritmo CPqD-IES utiliza um banco de dados de modelos de distorção para cenas diferentes da cena original, a fim de estimar o índice de qualidade de vídeo. Esta base de dados consiste em resultados de testes de avaliação subjetiva sobre doze cenas que apresentam diferentes graus de movimento (cenas estáticas e dinâmicas), natureza (cenas reais e sintéticos), e de contexto (quantidade de pixels de textura, de plano e de borda).

O modelo estima o índice de avaliação subjetiva calculando a relação entre as medidas objetivas e os resultados presentes no banco de dados. A relação entre cada um dos parâmetros objetivos e o nível de distorção subjetivo é aproximada por uma curva logística, resultando em um nível de distorção estimado para cada parâmetro. O resultado final é obtido através de uma combinação dos níveis de distorção estimados, com base em suas confiabilidades estatísticas. Um classificador de cena é utilizado para obter um sistema de avaliação independente da cena. Essa classificação utiliza a informação espacial (com base na análise DCT) e informações temporais (com base em alterações na segmentação) da sequência de entrada para obter os parâmetros do modelo a partir das informações do seu banco de dados.

O CPqD-IES é um modelo bastante complexo e de processamento pesado. Nos testes do VQEG o modelo alcançou uma correlação 0,835 com as notas subjetivas.

3.4.4 *Video Quality Model* - VQM

Outra abordagem é a proposta pela *National Telecommunications and Information Administration* (NTIA), que desenvolveu o modelo de qualidade de vídeo (VQM). Durante 2000 e 2001, a NTIA desenvolveu quatro modelos totalmente automatizados de avaliação objetiva de qualidade de vídeo, um modelo geral, um para televisão, um para videoconferência, e um para desenvolvedor [22]. O modelo geral foi projetado para ser uma métrica de propósito geral para sistemas de vídeo que abrangem uma

vasta gama de qualidade e taxas de bits. O modelo de televisão foi otimizado especificamente para distorções tipicamente encontradas em sistemas de televisão (por exemplo, vídeo codificados em MPEG-2), enquanto o modelo de video conferência foi otimizado especificamente para distorções comuns de aplicações de videoconferência (por exemplo, vídeos codificados em H.263, MPEG-4). O modelo para desenvolvedor foi otimizado para a mesma gama de qualidade de vídeo e taxas de bits que o modelo geral, mas com a restrição adicional de computação rápida. Estes quatro modelos em conjunto com o modelo PSNR e técnicas de calibração automática (por exemplo, registro espacial, registro temporal, estimativa e correção de ganho e deslocamento) foram totalmente implementadas em software. Este software, além de manuais de usuário e uma divulgação técnica completa dos algoritmos, está disponível em [23].

O modelo geral foi selecionado para apresentação ao teste de referência completa Fase II do VQEG, uma vez que proporciona a métrica mais robusta e de proposta geral que pode ser aplicado à mais ampla gama de sistemas de vídeo. Enquanto a Fase II de testes do VQEG avaliou somente o desempenho do modelo geral para os sistemas de televisão, o modelo geral foi projetado e testado para trabalhar para vários tipos de codificação e sistemas de transmissão (por exemplo, taxas de bits entre 10 kbits e 45 Mbit/s , MPEG-1/2/4, sistemas de transmissão digital com erros, sistemas de transmissão analógica e os sistemas baseados em fita). O modelo geral utiliza tecnologia patenteada de referência reduzida e produz os resultados de estimativa de qualidade que melhor simulam a percepção human [22]. Nos resultados das avaliações da Fase II, o modelo obteve uma correlação de 0,938, uma das mais altas, e um erro RMS de 0,074.

As quatro métricas citadas acima consideradas como estado-da-arte foram estudas e implementadas com os objetivos de se adquirir conhecimento dos métodos aplicados e se ter ferramentas de avaliação que gerassem resultados a serem comparados aos nossos. Todos os algoritmos foram implementados baseados nas descrições contidas no recomendação J.144, na linguagem C++ e com auxilio da biblioteca de computação visual OpenCV [24], também em linguagem C++, multi-plataforma e de código livre. A biblioteca OpenCV possui diversas funções e ferramentas de processamento de imagens que facilitaram o desenvolvimento desses algoritmos.

Da recomendação J.144 constam os valores gerados por cada métrica ao avaliar um conjunto de sequências de testes. Tais resultados serviriam como um critério de validação para as métricas implementadas para este trabalho, uma vez que obtivessem os mesmo valores para o mesmo conjunto de sequências. Infelizmente o conjunto de sequências utilizadas na Fase II e publicadas na J.144 não foram disponibilizados pelo grupo ainda, tornando duvidosos os resultados obtidos. Outro meio de validar as métricas implementadas seria comparando nossos resultados com os gerados pelos *softwares* de referência enviados por cada proponente ao VQEG du-

rante a realização das Fases I e II. Das quatro métricas, apenas o VQM possui uma implementação disponibilizada pelo NTIA. O software bvqm.exe [23] serviu como referência na validação da nossa versão do VQM.

Embora a métrica SSIM não faça parte dos testes do VQEG por ter sido desenvolvida para avaliar a percepção de qualidade de imagens estáticas, ela também foi considerada neste trabalho por ser amplamente utilizada.

Neste capítulo, descrevemos alguns métodos objetivos para avaliar a qualidade de vídeo. Discutimos também as principais idéias utilizadas no projeto de algoritmos de métricas objetivas, enumerando um conjunto representativo de métricas de qualidade de vídeo (FR, RR, e NR). Finalmente, discutimos os trabalhos do VQEG e a elaboração de métricas que representam hoje o estado-da-arte.

Capítulo 4

Método VCQM - *Video Conferencing Quality Model*

Neste capítulo propomos duas abordagens a serem incorporadas a métricas objetivas de avaliação. A primeira considera a característica humana de focar sua atenção visual a determinadas regiões da imagem, as chamadas regiões de interesse (*Region of Interest* - ROI). Distorções em tais regiões afetam mais a percepção de qualidade que distorções no restante da imagem. Um método que avalia separadamente tais regiões para depois combiná-las considerando a maior importância de uma região em relação à outra é desenvolvido para ser aplicado em métricas objetivas.

A segunda abordagem trata dos efeitos que a persistência retiniana produzem em casos como da presença de erro de canal. Distorções que ocorrem em um detereminado quadro da sequência parecem se prolongar por mais quadros devido esta característica. O método de "Janela Temporal de Mínimo "desenvolvido visa simular este efeito na avaliação de qualidade. Ambos métodos são combinados em uma métrica baseada no VQM para se obter um melhor desempenho na avaliação de qualidade de sequências de videoconferência.

4.1 O Algoritmo VQM

Por ter obtido os melhores resultados nos testes da Fase II do VQEG e ser a única métrica a ter um aplicativo de referência para validar nossa implementação, escolhemos o modelo geral do VQM [22] proposto pela NTIA (seção 3.4.4) para as alterações considerando Regiões de Interesse e Janela Temporal de Mínimos. Para entendermos melhor as modificações introduzidas como contribuição desta dissertação, antes é necessária uma descrição mais detalhada do algoritmo. O VQM pode ser dividido nas seguintes fases:

1. Calibração - Esta primeira etapa tem o objetivo de calibrar o vídeo em pre-

paração para a fase de extração de características. Com esta proposta, são estimados e corrigidos, juntamente com os deslocamentos espaciais e temporais, os deslocamentos de contraste e brilho, da sequência de vídeo processada em relação a sequência de vídeo original.

2. Extração de características de qualidade - Nesta fase, um conjunto de características de qualidade que descreve as mudanças na percepção espacial, temporal e de crominância é extraído de sub-regiões espaço-temporais da sequência de vídeo.

3. Estimativa de parâmetros de qualidade - Nesta fase, um conjunto de parâmetros de qualidade que descrevem as mudanças de percepção é calculado comparando características extraídas do vídeo processado com aquelas extraídas do vídeo de referência.

4. Estimativa da medida de qualidade - A etapa final consiste em calcular uma métrica de qualidade global através de uma combinação linear dos parâmetros calculados nas fases anteriores.

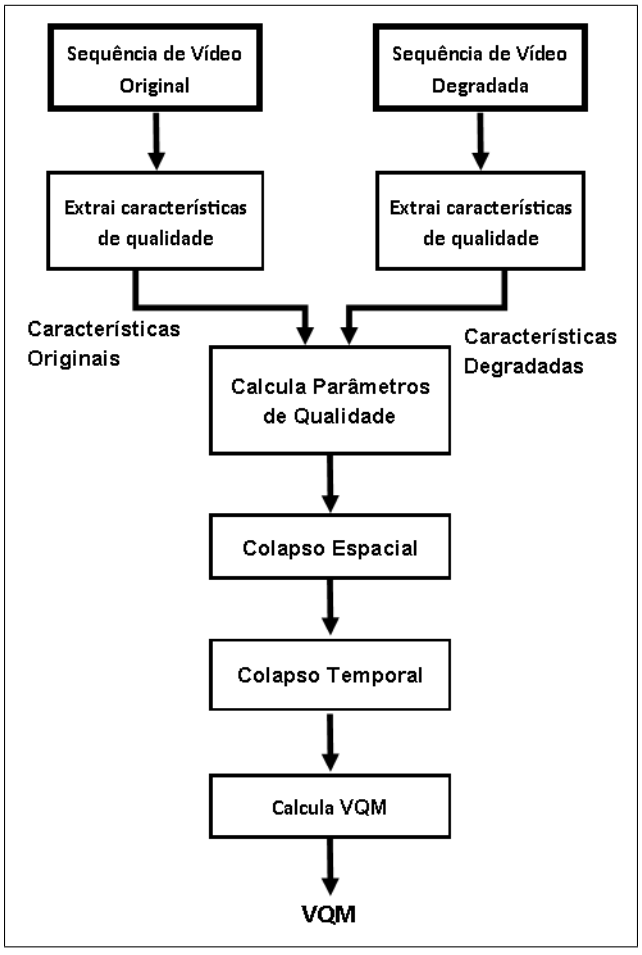


Figura 4.1: Diagrama de blocos do modelo VQM.

4.1.1 Extração de Características de Qualidade

Uma característica de qualidade no contexto do algoritmo VQM é definida como uma quantidade de informação associada, ou extraída de uma sub-região espaço-temporal de uma sequência de vídeo (originais ou processada). Em resumo, a fase de extração de características de qualidade segue os passos:

1. Aplicar um filtro perceptivo (opcional). Algumas características podem utilizar dois ou mais filtros perceptivos diferentes. Os chamados filtros perceptivos são funções que realçam alguns aspectos da qualidade do vídeo. A componente de luminância Y contém informações pertinentes às bordas. Uma versão de Y com realce das bordas identifica com maior precisão embaçamento, efeito de bloco e outros efeitos de borda. As componentes de cor, C_b e C_r, são úteis para identificar distorções de matiz e erros de transmissão. A diferenciação no tempo das componentes Y de quadros consecutivos destaca movimentos não naturais e bruscos.

2. Dividir a sequência de vídeo em sub-regiões espaço-temporais. Cada sub-região espaço-temporal descreve um bloco de pixels, $b(s,t)$. A dimensão dos blocos $b(s,t)$ é descrita pelo número de pixels na horizontal, o número de linhas de quadros na vertical, e o tempo de duração da região, dado em unidades de quadros de vídeo. A Figura 4.2 ilustra um bloco $b(s,t)$ de 8 pixels na horizontal x 8 linhas verticais x 6 quadros de vídeo, para um total de 384 pixels.

3. Extrair características de cada região (por exemplo, média, desvio padrão). Características $f_o(s,t)$ e $f_p(s,t)$, do sinal original e processado respectivamente, são extraídas de cada um destes blocos $b(s,t)$ utilizando uma função matemática simples. As duas funções mais utilizadas são média e desvio padrão. Após extração de características, o eixo temporal t já não corresponde aos quadros individuais. Pelo contrário, o eixo temporal contém um número de amostras igual ao número de quadros na sequência de vídeo dividido pela extensão temporal dos blocos.

4. Aplicar um limite de perceptibilidade (opcional). Finalmente, uma função de corte pode ser aplicada para reduzir a sensiblidade do parâmetro às distorções imperceptíveis [22]. A função de corte substitui os valores do parâmetro entre o limiar de corte e zero pelo limiar de corte. Esta função é representada matematicamente como:

$$f_{clip} = \begin{cases} max(f,T) & \text{se } f \text{ é posivito} \\ min(f,T) & \text{se } f \text{ é negativo} \end{cases}$$

onde T é o limiar de corte.

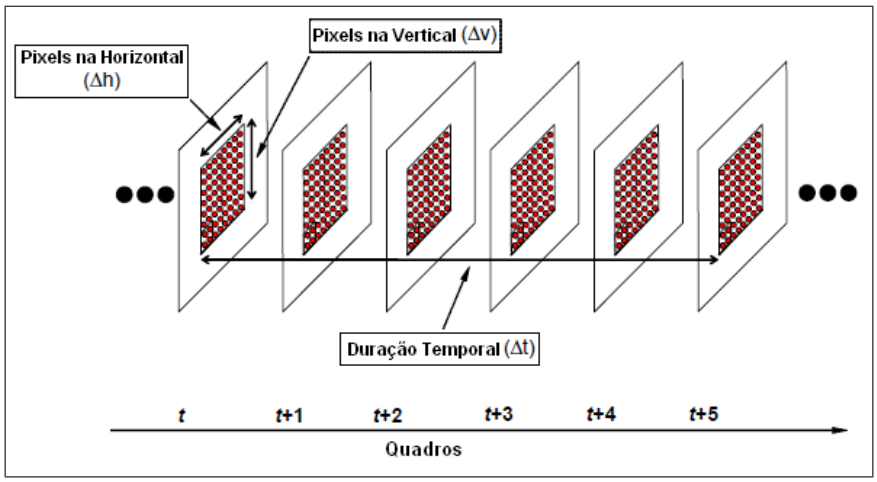


Figura 4.2: Exemplo de um bloco espaço-temporal $b(s,t)$ de extração das características de qualidade.

4.1.2 Extração de Parâmetros de Qualidade

Os parâmetros de qualidade $p(s,t)$ que medem as distorções na qualidade de vídeo são inicialmente calculados para cada bloco $b(s,t)$, comparando o valor da característica de qualidade da sequência original $f_o(s,t)$ com o valor $f_p(s,t)$ do bloco correspondente da sequência degradada. Em seguida, funções no espaço e no tempo de agrupamento de erro são utilizadas para simular como os seres humanos inferem avaliações da qualidade subjetiva. Agrupamentos de erro no espaço serão referidos como colapsos espaciais, e agrupamentos de erro ao longo do tempo são referidos como colapsos temporais.

Os parâmetros $p(s,t)$ dos blocos $b(s,t)$ formam matrizes tridimensionais abrangendo o eixo temporal e duas dimensões espaciais (ou seja, horizontal e vertical). Para a etapa de colapso espacial, distorções de blocos $b(s,t)$ com mesmo índice temporal t são combinados utilizando funções de colapso espacial (por exemplo, média, desvio padrão ou seleção percentual por limiar). As extensivas investigações realizadas pela NTIA revelaram que a função de colapso espacial ótima, muitas vezes envolve alguma forma de processamento do pior caso, como tomar a média dos 5% piores distorções observadas naquele momento. Isto ocorre porque distorções localizadas tendem a chamar a atenção do espectador, fazendo com que a pior região do quadro seja o fator predominante na decisão da qualidade subjetiva.

Os resultados $p(t)$ ao longo do eixo temporal da função de colapso espacial são em seguida combinadas utilizando uma função de colapso temporal (média, pior caso) produzindo um parâmetro objetivo p. Quando $p = 0$ não há degradação.

4.1.3 Estimativa da Medida de Qualidade

No modelo geral conforme descrito na seção 3.4.4, sete parâmetros de qualidade p independentes são combinados por meio de modelos lineares para obter o valor final de medida de qualidade VQM. Cada um dos sete parâmetros são obtidos através da comparação de características de qualidade específicas extraídas da sequência de referência e degradada. O modelo geral produz valores de saída que variam de zero (nenhuma alteração perceptível) a 1 (máxima distorção percebida).

4.2 Região de Interesse - ROI

Sob a hipótese de que certas regiões em uma imagem podem ser visualmente mais importantes que outras, métodos de avaliação utilizando esta informação no agrupamento espacial das medidas de qualidade possuem uma atraente possibilidade de melhorar a correlação da métrica a avaliação subjetiva.

Trabalhos recentes como em [25] e [26] propõem métricas de qualidade de vídeo que ponderem as medidas de distorção sobre a percepção da importância da região onde está localizado. O primeiro passo do algoritmo é encontrar quais são as áreas perceptualmente importantes do quadro de vídeo. Para isso, modelos estimam as principais características que atraem a atenção: o contraste da cor, tamanho de objetos, orientação e excentricidade. A medição dessas propriedades determina quais são as áreas de maior importância ou interesse. É importante ressaltar que extrair regiões de interesse de sequências de vídeo é uma tarefa complexa, pois tanto a extensão espacial como a evolução dinâmica das regiões devem ser consideradas.

No caso de aplicações de videoconferência, as sequências de vídeo possuem características específicas não necessariamente presentes em outros conteúdos. Uma das mais importantes é que a atenção do usuário geralmente está mais focada nas faces dos indivíduos do que nas outras regiões do vídeo. Portanto, degrações presentes em torno das faces podem ser mais facilmente notadas e incômodas do que quando ocorrem em regiões do fundo, por exemplo. Baseado nessa hipótese, a solução proposta modifica a métrica VQM para introduzir uma abordagem baseada em regiões de interesse (ROI) que segmenta os quadros da sequência em duas áreas, o interior da ROI, *In-ROI*, e o exterior da ROI, *Out-ROI* e calcula um valor objetivo de qualidade para cada. A nota final é obtida pela combinação das notas das duas regiões.

4.2.1 Detecção de Faces

Considerando o problema de detecção de faces, são inúmeras as metodologias computacionais existentes para a detecção e o reconhecimento de faces, a extração de informação facial, a análise de expressões faciais e a reconstrução de faces. Usu-

almente, essas metodologias consideram que a face humana pode ser interpretada globalmente, ou então, examinando algumas das suas características faciais mais representativas, como olhos, boca, sobrancelhas e nariz.

A escolha adequada da metodologia computacional que satisfaça as exigências de uma determinada aplicação depende muito da forma como a face se apresenta no cenário em estudo, sendo comum considerar-se como fatores de decisão variáveis como a luminosidade, escala, rotação, oclusão parcial e existência de óculos, bigode ou barba.

Existem diversas técnicas que podem ser empregadas para tal tarefa. Neste trabalho optamos por utilizar o algoritmo Viola-Jones [27] para detecção de faces por ser muito utilizado, ter processamento rápido e estar presente na biblioteca OpenCV. O algoritmo Viola-Jones de detecção de objetos foi o primeiro a oferecer taxas de detecção de objetos em tempo real. Embora possa ser treinado para detectar uma variedade de classes de objetos, ele foi motivado primeiramente pelo problema de detecção de faces.

O algoritmo é composto de três partes. A primeira delas é a representação da imagem em um espaço de características baseadas nos filtros de Haar. Isto é feito com auxílo da imagem integral [27]. A imagem integral é uma representação intermediária para a imagem usada para calcular rapidamente características contidas em retângulos. A segunda é a montagem de um classificador baseado em *Boosting* capaz de selecionar as características mais relevantes. Por fim é feita uma combinação em cascata destes classificadores de modo a garantir bom desempenho e velocidade de processamento. Na Figura 4.3 estão ilustrados exemplos das regiões de face detectadas pelo algoritmo.

Para a manipulação e ajuste desse método a função responsável do OpenCV possui alguns parâmetros, entre eles: fator de escala para a janela, o número de mínimos de vizinhos, o menor tamanho possível para cada face e a opção de procurar faces em áreas que provavelmente elas não existem. A configuração desses parâmetros influenciam no tempo de processamento e a chance de não se detectar uma face corretamente.

- O fator de escala da janela é o valor pelo qual a janela de pesquisa é escalada a cada ciclo de varredura da imagem, assim, para um valor de 1,1, a janela é escalada dez porcento em cada ciclo. Para um valor de 1,2 escalaria vinte porcento o que resultaria em menos janelas a serem verificadas e que resulta em um menor tempo de processamento porém uma maior chance de se perderem faces.

- O número de mínimo de vizinhos serve para resolvermos o problema ilustrado pela Figura 4.4. Neste caso, o valor adotado zero, retornou todas as regiões

Figura 4.3: Regiões de face detectadas pelo Viola-Jones.

da imagem que o detector julgou serem faces. Normalmente, o local onde realmente existe uma face recebe diversas marcações que se sobrepõem. Então, para obtermos um resultado mais adequado, utiliza-se esse parâmetro para definir o número mínimo de regiões sobrepostas necessárias para que uma face seja retornada.

- O parâmetro do menor tamanho possível para cada face é utilizado para definirmos um tamanho mínimo para a face ser encontrada diferente do padrão pelo qual o filtro em cascata foi treinado, que em geral é de 20x20 ou 24x24 pixels. Assim, podemos determinar que apenas faces de tamanhos maiores sejam encontradas com a intenção de que não se perca tempo de processamento tentando detectar faces que não interessam para a aplicação que o sistema esteja sendo desenvolvido.

- O último parâmetro permite evitar a procura de faces em áreas que provavelmente elas não existem. Isso também é útil para evitar que se perca tempo de processamento em áreas que, após a passagem de um detetor de bordas na imagem, acredita-se que não existam faces.

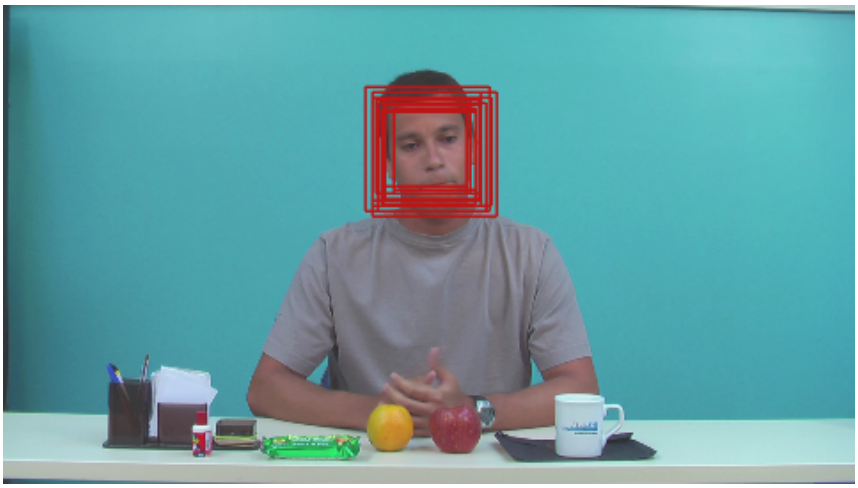


Figura 4.4: Ilustração da detecção com número mínimo de vizinhos igual a zero.

Após implementar o método, diversos ajustes foram necessários para maximizar seu desempenho. Os parâmetros da função do OpenCV do algoritmo de Viola-Jones foram variados e também combinou-se essa função com outras etapas de processamento das regiões após sua detecção. Na tentativa de aumentar o número de faces detectadas, o parâmetro de número de mínimos de vizinhos foi diminuído, entretanto, isso também resultou no aumento do número de falsos positivos, o que era esperado. Um das contribuições desta dissertação diz respeito à eliminação dos falsos positivos e à redução da variabilidade das regiões de interesse entre quadros vizinhos, para isto, foi acrescentado ao processo de detecção um pós-processamento das regiões encontradas. As etapas deste pós-processamento estão listadas a seguir:

1. Janelamento dos quadros: a cada quadro, uma janela com os quadros vizinhos é criada.

2. Agrupamento temporal de regiões de face sobrepostas: dentro da janela, cada quadro pode ter um número indeterminado de possíveis regiões de face detectadas pelo algoritmo. As regiões com sobreposição espacial são classificadas como uma possível face para esta posição do vídeo. Por exemplo, todas regiões localizadas no centro da imagem formam uma possível face para esta posição.

3. Eliminação de falsos positivos: dentro da janela, possíveis faces com poucas regiões detectadas nesta posição em relação ao tamanho da janela são descartadas.

4. Execução de um filtro de mediana vetorial [28] em cada janela para minimizar a variação da região entre quadros da sequência de vídeo. Definindo P como o ponto superior esquerdo da ROI e considerando uma janela de N pontos, uma amostra P_i é escolhida para substituir o ponto central se o somatório

de distâncias euclidianas entre o ponto P_i e as demais amostras da janela for mínimo. O coeficiente de escolha C_i de um ponto P_i é dado por:

$$C_i = \sum_{j=0}^{N} (D(P_i, P_j))$$

$$D(P_i, P_j) = |P_i - P_j|$$

onde i é o índice da amostra escolhida.

5. Ajuste automático das regiões: A regiões são ampliadas horizontalmente e verticalmente para incluir pontos relevantes da sua vizinhança (pescoço, cabelos, etc).

Em nossos testes foram utilizadas janelas de tamanho 10, 15, 30, 45 e 60 quadros para o filtro de mediana vetorial. A maior taxa de faces encontradas e falsos positivos eliminados foi obtido com janelas de 30 quadros. A Figura 4.5 ilustra o resultado do pós-processamento em um quadro. Os retângulos vermelhos, a região da face e um falso positivo, são as regiões detectadas pelo algoritmo Viola-Jones. O resultado do pós-processamento representado pelo retângulo verde elimina o falso positivo e amplia a região da face incluindo pescoço e cabelos. Como era esperado, faces frontais foram bem identificadas contanto que não estivessem rotacionadas, já que essa característica é uma limitação do método utilizado. Apenas faces levemente rotacionadas ainda eram identificadas. Com relação à invariância a escala, testes com sequências de resolução QCIF, CIF, SD e HD demonstraram que o método continua detectando as regiões de face corretamente. O único fato que pode ser percebido foi sua restrição ao tamanho mínimo da face presente no vídeo, que decorre da limitação existente no detector em cascata utilizada pelo método. Isso limita a identificação de faces muito pequenas presentes em vídeos, mas não teve impacto nos resultados, já que faces desse tamanho não são tão frequentes em videoconferências, e quando presentes, normalmente não fazem parte da ação principal da cena.

4.3 Métrica VQM baseada em Regiões de Interesse

Para aplicar o método de região de interesse proposto foram implementadas alterações no algoritmo do VQM modelo geral. Duas novas etapas anteriores a extração de características de qualidade 4.1, a extração das ROIs e segmentação dos quadros foram incluidas ao algoritmo. O diagrama de blocos do método é ilustrado na Figura 4.6

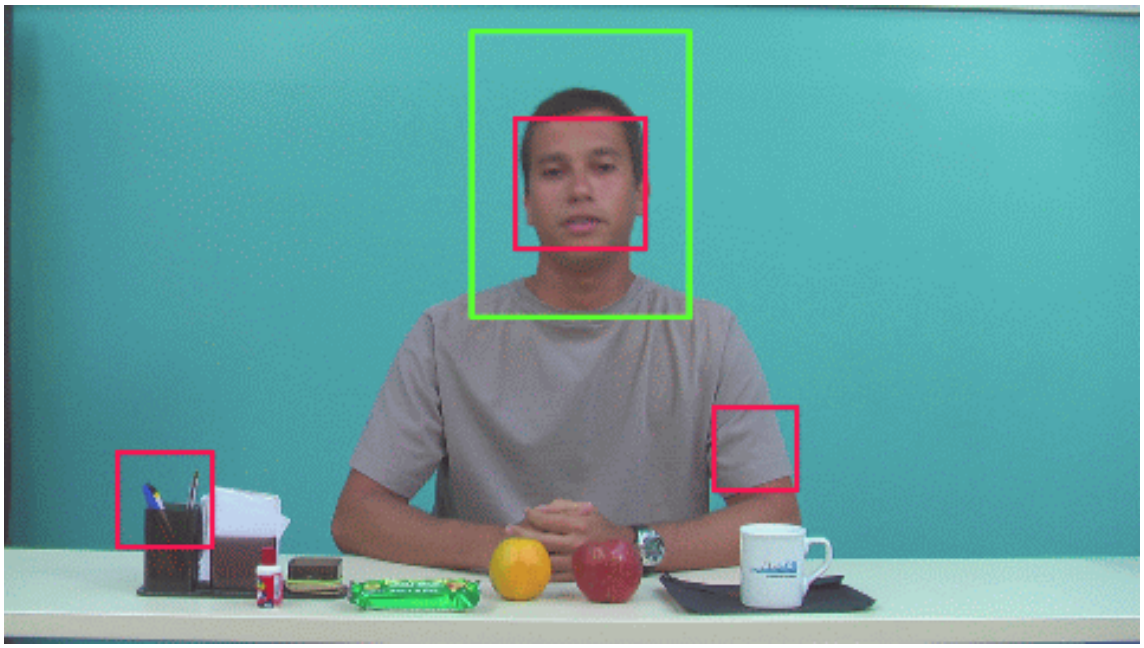


Figura 4.5: Em vermelho, as regiões detectadas pelo Viola-Jones. Em verde a região resultante do pós-processamento.

Na etapa de extração das ROIs, o método de detecção de faces processa a sequência de referência e gera as ROIs de cada quadro do vídeo. As regiões de alta frequências mencionadas ao final da seção 5.3 também foram consideradas como ROIs de cada quadro.

A partir de cada quadro dois novos são gerados, *In-ROI* e *Out-ROI*. A partir da sequência formada pelos quadros *In-ROI* estimamos o valor do VQM na região interna da ROI, VQM_{In-ROI}. A estimativa do VQM na região externa da ROI, $VQM_{Out-ROI}$, é obtida utilizando a sequência formada pelos quadros *Out-ROI*. A Figura 4.7 exemplifica a etapa de segmentação em quadros *In-ROI* e *Out-ROI*.

Para ambas sequências, *In-ROI* e *Out-ROI*, o método de estimativa do VQM é o mesmo. Primeiramente definimos como região de análise (ROA) a região da sequência onde queremos obter o VQM. Nas sequências *In-ROI* trata-se da região interna das ROIs. Nas sequências *Out-ROI* as ROAs é a região externa das ROIs.

Para estimar o VQM de uma ROA, apenas valores de $p(s,t)$ pertencentes a região ROA devem ser considerados no cálculo da métrica. Como o método VQM divide a sequência em blocos $b(s,t)$ a eliminação de blocos fora da ROA garantem esta necessidade. Porém uma abordagem mais simples e rápida, em que não é necessário verificar quais blocos pertencem a ROA, foi elaborada para obter os mesmos resultados.

Considerando-se que $p = 0$ quando não há distorções na imagem, é razoável considerar que valores de $p(s,t)$ também serão iguais a zero em regiões idênticas. No processo de montagem do quadro *In-ROI* e *Out-ROI* a ROA é retirada do

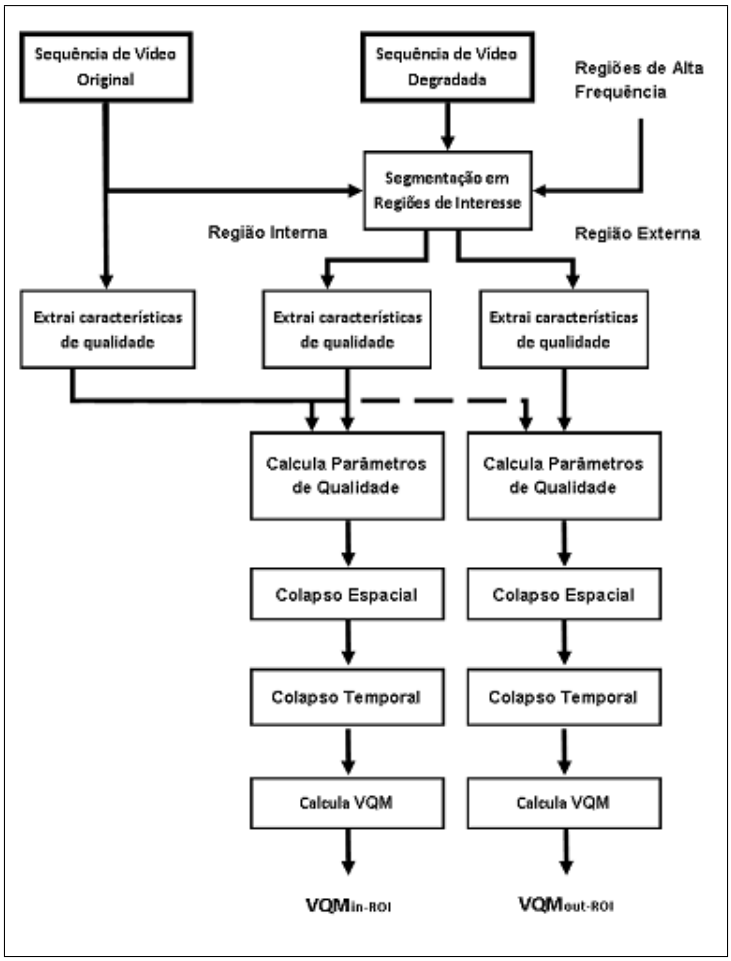


Figura 4.6: Diagrama de modificação do algoritmo VQM considenrado regiões de interesse.

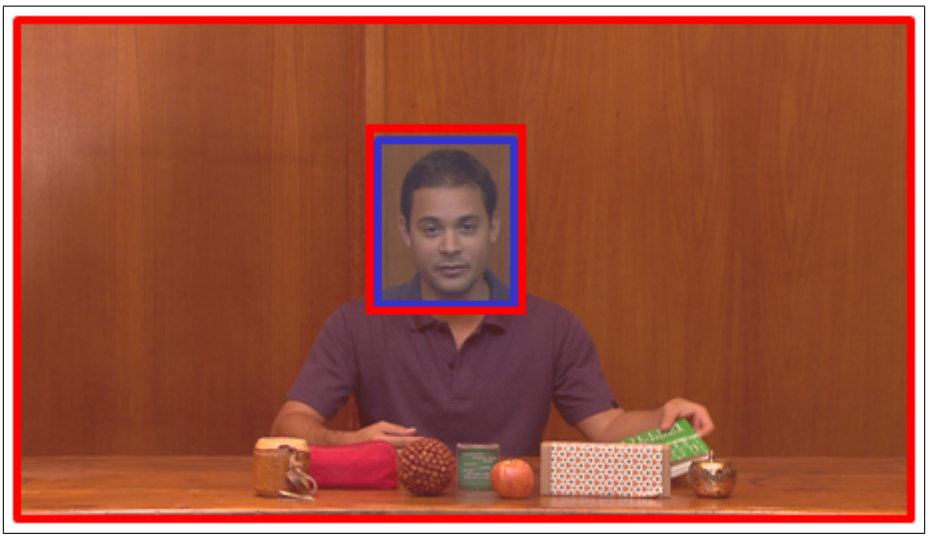


Figura 4.7: Exemplo segmentação do quadro. Em azul a região interna da ROI. Em vermelho a região externa da ROI.

quadro degradado, região com presença de distorção, e o restante é copiado do quadro de referência, garantindo a existência de regiões idênticas entre os quadro

comparados. Desta forma ao se realizar a comparação entre $f_o(s,t)$ e $f_p(s,t)$ para obter os parâmetros $p(s,t)$, todos terão valor zero se excetuando os que se encontram na ROA.

4.4 Janela Temporal de Mínimos

Uma das características do SVH chamada persistência retiniana designa o fenômeno ou a ilusão provocada quando um objeto visto pelo olho humano persiste na retina por uma fração de segundo após a sua percepção. Portanto é razoável considerar que erros que ocorrem em um único ou poucos quadros são percebidos por um período maior devido à persistência da retina. Para compensar esse fato outra abordagem é proposta como melhoria da métrica VQM.

Métricas tradicionais de qualidade avaliam os erros por computação de médias das medições ao longo de um número de quadros. Aqui propomos a abordagem de Janela Temporal de Mínimos (*Temporal Window of Minima* - TWM) que considera o efeito da persistência visual do erro. A abordagem TWM atribui o valor mínimo da medida de avaliação para todos os quadros em uma janela de tempo, simulando a persistência da retina. Desta forma a pior degradação percebida em um período de tempo se sobrepõe aos outros níveis de degradações percebidos.

No algoritmo do VQM existe um método similar de processamento dos índices de qualidade dos blocos do quadro que considera os piores casos de distorção. Diferente do método proposto onde se considera o efeito temporal de piores casos, o VQM considera apenas o pior caso espacial devido à constatação de que as maiores distorções localizadas dentro de um quadro tendem a chamar mais a atenção do espectador, fazendo com que a avaliação desta região seja fator predominante na decisão da qualidade subjetiva. Isto reforça a ideia que esta distorção percebida em uma região do quadro influência a avaliação do quadros sequintes devido à resistência retiniana.

4.4.1 Métrica VQM baseada em Janela Temporal de Mínimos

Na modificação do algoritmo do VQM foi acrescentada uma etapa de TWM para considerar os efeitos da persistência visual do erro. Esta etapa consistente na atribuição da menor medida de qualidade (maior distorção) encontrada para todos os quadros dentro de uma janela de tempo.

No caso do VQM, após a etapa de extração de parâmetros de qualidade (seção 4.1.2), uma janela temporal não sobreposta é formada com valores de $p(s,t)$ de mesmo índice espacial. O parâmetro de menor valor da janela é então atribuído a todos instantes de tempo, simulando a persistência deste erro ao longo da janela.

$$p_{min} = min(p(s,t)), \forall t \in W$$

$$p'(s,t) = p_{min}, \forall t \in W$$

sendo W a janela temporal, $p(s,t)$ valores de parâmetros de qualidade com $t \in W$ e $s = constante$. O p_{min} é o menor valor de $p(s,t)$ na janela W.

No modelo geral do VQM os blocos $b(s,t)$ possuem a duração de 6 quadros [22]. Na etapa de extração de características toda informação de cada bloco é representada por um único valor de $f(s,t)$. Como queremos encontrar o pior caso de cada quadro, adotamos blocos $b(s,t)$ com duração de um quadro. Para manter a coerência temporal com o VQM original, utilizamos uma janela temporal com 6 quadros de duração. Ao final do processo de janelamento, o número de amostras $p'(s,t)$ é o mesmo gerado pelo VQM clássico. As etapas seguintes do algoritmo não foram alterados, como pode ser visto na Figura 4.8. No final obtemos o valor (*M-VQM*) de medida de qualidade.

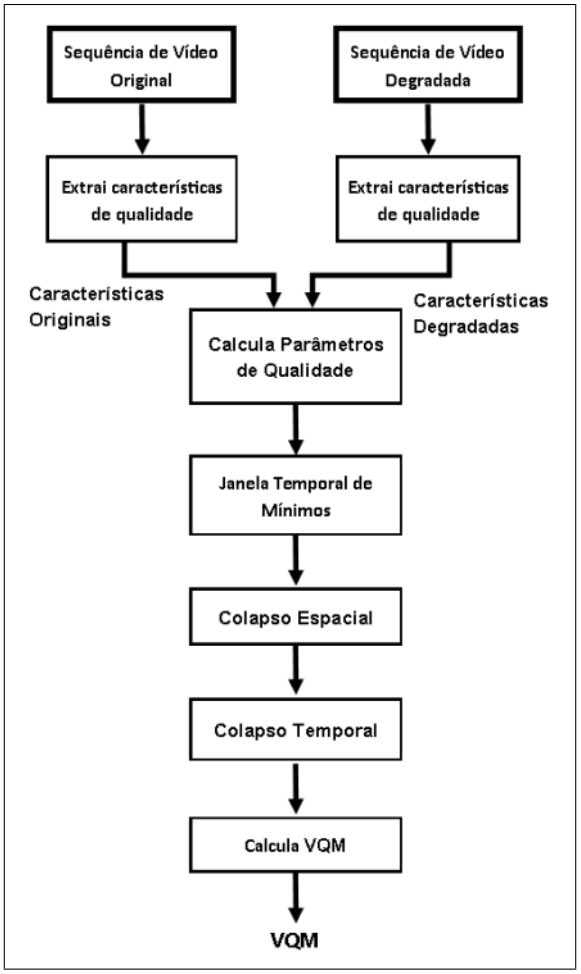


Figura 4.8: Diagrama do algoritmo VQM com Janela Temporal de Mínimos (*M-VQM*).

4.5 Métrica VCQM

Ambas as abordagens descritas acima foram combinadas na métrica VCQM (*Video Conferencing Quality Model*). Como os dois métodos são independentes um no outro, na VCQM primeiro é realizada a etapa de reconhecimento de regiões de interesse, como na *ROI-VQM*. Para cada sequência *In-ROI* e *Out-ROI* é aplicado o algoritmo *M-VQM*. Ao final obtemos os índices de qualidade $M\text{-}VQM_{In-ROI}$ e $M\text{-}VQM_{Out-ROI}$. A métrica VCQM é obtida da combinação destes dois valores e o VQM sem ROI.

Para comprovar a eficácia dos métodos propostos, avaliamos a métrica SSIM com abordagem de ROIs e Janela Temporal de Mínimos. Assim como VQM, combinamos os 3 valores de SSIM, $SSIM_{In-ROI}$ e $SSIM_{Out-ROI}$ para obter a maior correlação com os testes subjetivos. A métrica considerando os dois métodos é chamada de *Video Conferencing Structural Similarity* (VC-SSIM).

No Capítulo 5 veremos os resultados dos testes subjetivos que foram realizados para avaliar o desempenho das métricas descritas aqui e a melhor forma combiná-las para ter a maior correlação com as notas subjetivas.

Capítulo 5

Banco de Sequências de Videoconferência com Alta Definição

Uma vez que o objetivo desta pesquisa centra-se em aplicações de videoconferência, um banco de sequências de vídeos de alta definição foi gerado. Este banco de sequências foi projetado para conter situações que seriam vistas em sessões de videoconferência típicas. Tais situações incluem:

- Indivíduos do sexo masculino e feminino;
- Oclusão parcial de rosto;
- Fundo plano, simples e complexo;
- Rotação da face;
- Indivíduos usando óculos ou com barba;
- Dois indivíduos na cena;

Primeiramente vinte e duas sequências de referência foram registradas para formar o bando de Sequências de Videoconferência com Alta Definição (SVAD). Este banco foi composto de gravações em alta definição (1920x1080) e contém cenas típicas vistas em aplicações de videoconferência. Todas sequências foram gravadas por um profissional nas instalações do Laboratório de Processamento de Sinais [29] (LPS) na Universidade Federal do Rio de Janeiro (UFRJ). Estas sequências estão disponíveis em [30]. Suas principais características podem ser vistas junto com um quadro representativo no Apêndice A.

Tabela 5.1: Sequências de referência utilizadas no conjunto de testes subjetivos.

Sequência	Tipo de Fundo	Características
01	Liso	Indivíduo do sexo masculino
02	Liso	Oclusão da face
08	Simples	Indivíduo do sexo feminino; oclusão da face
14	Simples	Indivíduo com pele escura
15	Simples	Dois indivíduos
18	Complexo	Indivíduo do sexo masculino

Do banco de sequências foram selecionadas as 6 mais representativas para a elaboração do conjunto de testes subjetivos. As sequências utilizadas e suas características são listadas na Tabela 5.1.

A partir deste conjunto, sequências de testes foram geradas através de distorções de brilho e contraste (degradação perceptual) e erros de canal (perda de blocos), seguida de um algoritmo de ocultamento de erros.

5.1 Simulação de Erro de Brilho e Contraste

Com o objetivo de investigar a influência de parâmetros perceptivos do modelo do sistema visual humano na avaliação de qualidade, sequências de teste foram geradas distorcendo-se o brilho e o contraste das sequências de referência. Sequências com distorção de brilho tiveram o seu nível médio de luminância de cada quadro deslocado, conforme ilustrado na Figura 5.1. Para as sequências com degradação no contraste foi aplicado a cada quadro um método de realce no contraste da biblioteca OpenCV, Figura 5.2. Foram simuladas variações de ±25%, ±50% e ±75% no brilho e contraste para cada sequência.

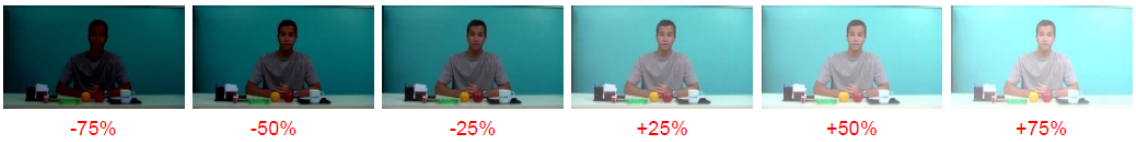


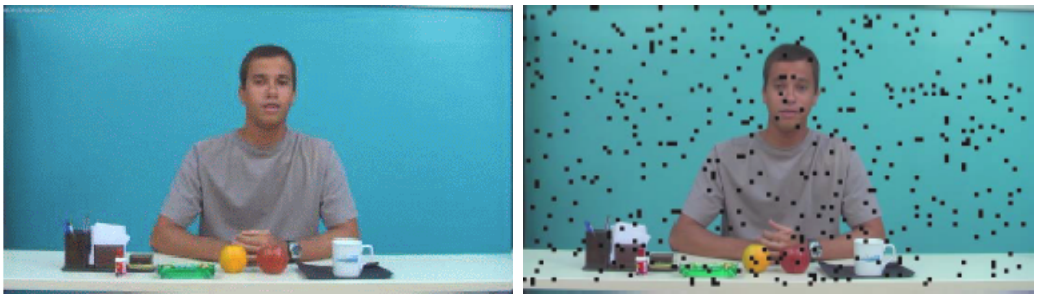
Figura 5.1: Variações no nível de brilho do quadro.

Figura 5.2: Variações de contraste do quadro.

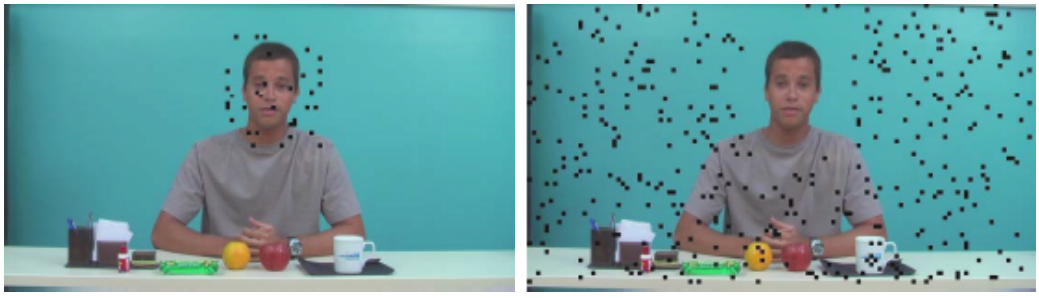
5.2 Simulação de Erros de Canal

Para compreender como os erros de canal influenciam na percepção de qualidade, sequências degradadas com percentuais variados de perda de macroblocos (aplicada a cada quadro) foram geradas. Uma técnica de ocultamento de erro padrão [31] foi aplicada para corrigir a perda dos macroblocos. Foram simulados três cenários:

- Quadro inteiro: os erros são distribuídos por todo o quadro ao acaso, Figura 5.3(b);
- Dentro da região de interesse: os erros são distribuídos apenas no interior da região de interesse (neste caso, a percentagem de macroblocos perdidos está relacionado ao tamanho da região e não o quadro), Figura 5.3(c);
- Fora da região de interesse: erros estão distribuídos aleatoriamente apenas fora da região de interesse (percentual também relacionada ao tamanho da região), Figura 5.3(d);



(a) (b)



(c) (d)

Figura 5.3: Exemplos de simulação de Erro de Transmissão. (a) Sequência original. (b) Sequência com erro distribuído pelo quadro inteiro. (c) Sequência com erro apenas no interior da região de interesse. (d) Sequência com erro apenas no exterior da região de interesse. Os erros não foram ocultados para fins de ilustração.

Três taxas de nível de erro foram utilizadas na geração de sequências de testes, 1%, 5% e 10%. Combinações de casos de erro de brilho/contraste e transmissão também foram gerados, resultando em um total de 342 sequências de teste. Deste

total, foram selecionadas as 150 sequências mais apropriadas para os testes subjetivos, considerando o nível de distorção (baixo/alto) e o tempo necessário para a realização das sessões de avaliação. As sequências de teste geradas e as selecionadas, assim como o tipo de distorção, encontram-se no Apêndice C. Os testes assumiram que a duração das sequências é de 8 a 10 segundos. Sequências com duração acima de 10 segundos foram editadas e apenas os primeiros dez segundos foram considerados nas avaliações.

5.3 Testes Subjetivos

Os experimentos de avaliação subjetiva realizadas neste trabalho seguiram as diretrizes especificadas na Recomendação ITU-R BT.500-10 [3] citadas no Capítulo 2. Estes procedimentos incluem as condições de exibição, os critérios para selecionar os observadores, os materiais de teste e métodos para analisar os dados coletados.

O método de avaliação utilizado neste trabalho foi o mesmo da fase I de experimentos do VQEG [1], o DSCQS. Neste método as sequências de teste são apresentadas em uma ordem (pseudo) aleatória. De preferência, o mesmo vídeo de referência não é exibido duas vezes em sequência. No DSCQS, os observadores assistem pares de sequências correspondentes à versão original e degradada (em ordem aleatória). As sequências são apresentadas duas vezes, como visto na Figura 5.5.

De acordo com as orientações em [3], até três observadores podem participar em cada sessão de avaliação de qualidade. O experimento é definido como visto na Figura 5.4. Pelo menos 15 observadores atribuiram uma nota a cada sequência de modo a garantir resultados válidos.

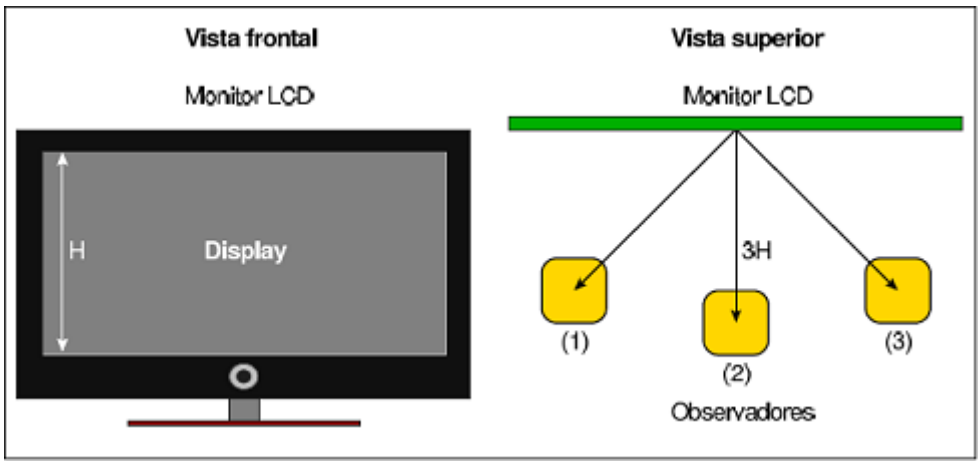


Figura 5.4: Configuração do teste para três observadores.

Neste método, os indivíduos não são informados sobre quais são as sequências originais e quais são as degradadas. Depois que cada par é exibido duas vezes, a

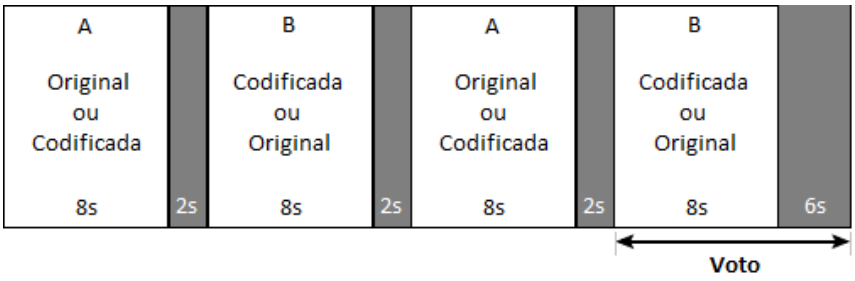


Figura 5.5: Ordem de visualização das sequências de teste do método DSCQS.

ordem entre a sequência original e a degradada é alterada aleatoriamente para o próximo par. Um nível médio de cinza (nível de luminosidade de 128) deve ser mostrado entre as sequências [3].

Durante o período de voto, o observador registra o nível de qualidade das sequências percebido por ele. A escala utilizada é mostrada na Figura 2.6. Os níveis de qualidade utilizados para classificar as sequências são explicados na Tabela 5.2.

Tabela 5.2: Níveis de qualidade e a sua degradação associados no DSCQS.

Nível	Qualidade	Degradação
5	Excelente	Imperceptível
4	Bom	Perceptível mas não incómodo
3	Razoável	Perceptível e um pouco incómodo
2	Pobre	Incômodo
1	Ruim	Muito incômodo

Antes de efetuar a análise completa dos dados, uma pós-triagem das notas subjetivas é realizada. O primeiro passo desta análise é verificar a integridade dos dados para cada observador. Um observador é descartado se houver mais de um voto perdido em uma sessão única de teste. A segunda etapa da seleção é a eliminação dos observadores com notas instáveis e os com valores extremos, isto é, *outliers*.

O processamento das notas é realizado com procedimentos de análise padrão, conforme especificado nos relatórios finais das Fases I e II de testes do VQEG [1, 21]. Mais especificamente, para cada combinação de variáveis do teste, a média e o desvio padrão das notas de qualidade devem ser computados.

Em torno de 90 observadores fizeram parte dos testes, e cada sequência degradada foi avaliada por pelo menos 15 usuários. As notas subjetivas finais entre as sequências degradadas e de referência foram computadas como o *Differential Mean Opinion Score* (DMOS).

Ao final dos testes subjetivos percebeu-se nas simulações de erros de canal que distorções em regiões de objetos com bordas acentuadas também resultaram em menores notas na avaliação subjetiva. Consideramos que a degradação nessas regiões também chama a atenção do observador pois apresentam objetos que contêm bordas com alto contraste e o SVH possui grande sensibilidade à degradação de borda,

como é mencionado na descrição da métrica EPSNR, seção 3.4.2. Portanto passamos a incluir estas regiões no método desenvolvido. ROI típicas são ilustradas na Figura 5.6 e incluem rostos e objetos que contêm informações de alta frequência. Como esta região é fixa nas sequências de vídeo utilizadas neste trabalho, elas foram mapeadas manualmente para cada sequência de referência utilizada nos testes subjetivos. Em sistemas práticos, a detecção da região face pode ser feita usando algoritmos conhecidos como o Viola-Jones, e a região contendo objetos pelo conteúdo de frequência.

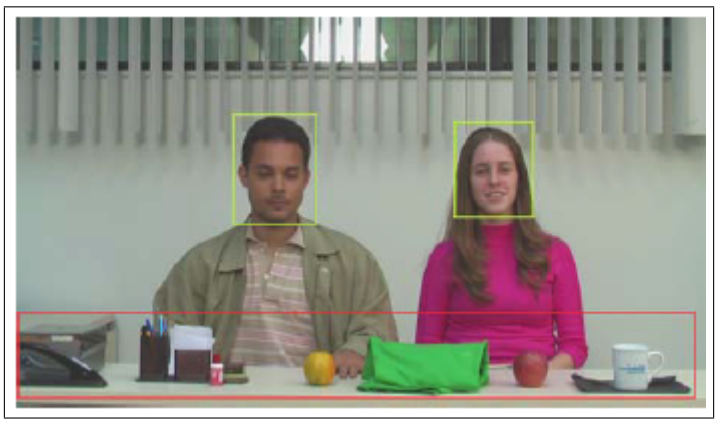


Figura 5.6: Região de interesse composta de face e objetos com alta frequência.

5.4 Correlação entre Notas Subjetivas e Objetivas

Por final, as notas objetivas de qualidade devem ser correlacionadas com as notas DMOS dos observadores. Uma função de regressão não-linear é usada para transformar o conjunto de notas objetivas em um conjunto de notas preditas $DMOS_p$, que depois são comparadas com os valores reais de DMOS dos testes subjetivos para avaliar o desempenho da métrica.

Considerando os três tipos de notas:

$$G_{no} = fm(nota\ objetiva\ sem\ ROI);$$

$$G_{in} = fm(nota\ objetiva\ interior\ da\ ROI);$$

$$G_{out} = fm(nota\ objetiva\ no\ exterior\ da\ ROI);$$

sendo fm uma função de mapeamento exponencial. A predição da nota DMOS é obtida pela combinação:

$$DMOS_p = f(G_{no}, G_{in}, G_{out}) \tag{5.1}$$

As notas preditas $DMOS_p$ são calculadas como uma média ponderada das 3 notas de entrada G_{no}, G_{in} e G_{out} após a aplicação de uma função não-linear, onde os pesos e os coeficientes de função foram obtidos usando uma estratégia de otimização com a formação de conjuntos de treinamento conhecida como *K-fold* [32], [33].

5.4.1 Estratégia de Otimização *K-fold*

A estratégia de otimização utilizada consistia em dividir o banco de dados aleatoriamente em três conjuntos, **A**, **B** e C com a mesma quantidade de sequências em cada. Em seguida, esses conjuntos foram combinados dois a dois para criar o conjunto de treinamento, enquanto o terceiro foi usado para testes. Portanto, três combinações possíveis foram utilizadas: 1) conjuntos **A** + **B** para treinamento e conjunto **C** para teste, 2) conjuntos **A** + **C** para treinamento e **B** para teste 3) **B** + **C** para treinamento e **A** para teste. Para cada combinação de conjuntos de treinamento encontramos os coeficientes da função 5.1 de melhor ajuste dos dados, de tal forma que a correlação é minimizada.

Uma variedade de funções foram consideradas durante a regressão não-linear, como exponenciais e polinômios [1]. Mais detalhes do processo de otimização encontram-se no Apêndice D.

Para a otimização conjunta das notas VQM os melhores resultados foram obtidos usando as seguintes funções:

$$G_{no,p} = \frac{x(1)}{1 + \exp[-x(2) * (x(3) * G_{no} - x(4))]} \tag{5.2}$$

$$G_{in,p} = \frac{x(5)}{1 + \exp[-x(6) * (x(7) * G_{in} - x(8))]} \tag{5.3}$$

$$G_{out,p} = \frac{x(9)}{1 + \exp[-x(10) * (x(11) * G_{out} - x(12))]} \tag{5.4}$$

$$DMOS_p = x(13) * G_{no,p} + x(14) * G_{in,p} + x(15) * G_{out,p} \tag{5.5}$$

onde $x(n)$ são os coeficientes da função a serem encontrados pelo processo de otimização. Para as métricas SSIM os melhores resultados foram obtidos com as seguintes funções:

$$G_{no,p} = x(1) * (G_{no})^3 + x(2) * (G_{no})^2 + x(3) * G_{no} + x(4) \tag{5.6}$$

$$G_{in,p} = x(5) * (G_{in})^3 + x(6) * (G_{in})^2 + x(7) * G_{in} + x(8) \tag{5.7}$$

$$G_{out,p} = x(9) * (G_{out})^3 + x(10) * (G_{out})^2 + x(11) * G_{out} + x(12) \tag{5.8}$$

$$DMOS_p = x(13) * G_{no,p} + x(14) * G_{in,p} + x(15) * G_{out,p} \tag{5.9}$$

A correlação final é obtida como a correlação média do conjunto de testes.

5.5 Resultados

Na Tabela 5.3 estão presentes os índices de correlação da métrica SSIM e as versões considerando regiões de interesse, $SSIM_{In-ROI}$ e $SSIM_{Out-ROI}$. O $DMOS_p$ foi obtido pela equação 5.9, sendo $G_{no} = f(SSIM)$, $G_{in} = f(SSIM_{In-ROI})$, $G_{out} = f(SSIM_{Out-ROI})$ e f uma função de mapeamento exponencial. Para o conjunto de testes contendo apenas sequências com distorção perceptual (brilho e contraste) , a métrica SSIM obteve uma correlação de 0,95 com os testes subjetivos, comprovando o seu bom desempenho mesmo para sequências de vídeo. Para o conjunto de testes contendo distorção de canal seu desempenho cai para 0,75. Por ser uma métrica para avaliar imagens estáticas, o SSIM não considera as características temporais da sequência de vídeo. A característica temporal que consideramos os erros de canal possuir devido à persistência retiniana pode ser um dos motivos desta queda.

Com a melhor combinação das três métricas SSIM, $SSIM_{In-ROI}$ e $SSIM_{Out-ROI}$ na nota $DMOS_p$ conseguimos uma correlação de 0,91 para distorções perceptuais e 0,72 para distorções de canal, ainda abaixo dos valores obtidos utilizando apenas a métrica SSIM.

Tabela 5.3: Correlação média utilizado SSIM e o método ROI. $DMOS_p = f(SSIM, SSIM_{In-ROI}, SSIM_{Out-ROI})$.

Conjunto de testes	SSIM	$SSIM_{In-ROI}$	$SSIM_{Out-ROI}$	$DMOS_p$
Distorção Perceptual	0,95	0,85	0,94	0,91
Distorção de Canal	0,75	0,74	0,73	0,72

A Tabela 5.4 mostra os resultados da métrica SSIM utilizado os métodos ROI e Janela de Mínimos (TWM). O $DMOS_p$ foi obtido com $G_{no} = f($ M-$SSIM$, $G_{in}) = f($ M-$SSIM_{In-ROI})$ e $G_{out} = f($ M-$SSIM_{Out-ROI})$. Na avaliação de distorções perceptuais o método de TWM não alterou os resultados obtidos anteriormente pois este tipo de distorção não possui efeito temporal. Na presença de distorções de canal vemos uma melhora marginal em relação à versão sem estratégia temporal de mínimos, principalmente na combinação otimizada destas notas ($DMOS_p$). Este resultado indica que a abordagem TWM pode trazer ganhos.

Tabela 5.4: Correlação média utilizado SSIM e os métodos ROI e Janela de Mínimos. $DMOS_p = f(M-SSIM, M-SSIM_{In-ROI}, M-SSIM_{Out-ROI})$.

Conjunto de testes	M-$SSIM$	M-$SSIM_{In-ROI}$	M-$SSIM_{Out-ROI}$	DMOSp
Distorção Perceptual	0,95	0,87	0,94	0,91
Distorção de Canal	0,76	0,74	0,74	0,74

Na Tabela 5.5 estão presentes os índices de correlação das métricas VQM (sem região de interesse), VQM_{In-ROI} e $VQM_{Out-ROI}$. O $DMOS_p$ foi obtido com $G_{no} = f(VQM)$, $G_{in} = f(VQM_{In-ROI})$ e $G_{out} = f(VQM_{Out-ROI})$ pelas equações 5.5, 5.2, 5.3 e 5.4.

Um conjunto de testes contendo ambos tipos de distorção (percetual e de canal) também foi avaliado pelas métricas VQM, VQM_{In-ROI} e $VQM_{Out-ROI}$. Neste conjunto foi utilizado a taxa fixa de 5% para o nível de erro de canal (Apêndice C).

Observamos que para distorções do tipo perceptual o $DMOS_p$ combinando as três métricas apresentou melhorias em relação à métrica original, obtendo uma correlação de 0,84 contra 0,82 do VQM sem ROI.

Na avaliação das simulações de distorção do canal a abordagem por ROIs não mostrou ganhos em relação à métrica original, similar ao comportamento observado nos testes com SSIM. No entanto as métricas tiveram uma queda considerável de desempenho quando erros de canal estavam presentes. Este resultado não era esperado para o VQM pois, ao contrário do SSIM, a métrica possui procedimentos que consideram as características temporais do SHV.

Tabela 5.5: Correlação média utilizado VQM e o método ROI. $DMOS_p = f(VQM, VQM_{In-ROI}, VQM_{Out-ROI})$.

Conjunto de testes	VQM	VQM_{In-ROI}	$VQM_{Out-ROI}$	DMOSp
Distorção Perceptual	0,82	0,52	0,84	0,84
Distorção de Canal	0,62	0,54	0,58	0,58
Distorção Perceptual + de Canal	0,71	0,59	0,72	0,72

Por último a Tabela 5.6 apresenta os resultados combinados das abordagens por ROI e Janela de Mínimos utilizando VQM. Eles foram equivalentes aos para o VQM sem estratégia de janela de mínimos.

Tabela 5.6: Correlação média utilizado VQM e os métodos ROI e Janela de Mínimos. $DMOS_p = f(M-VQM, M-VQM_{In-ROI}, M-VQM_{Out-ROI})$.

Conjunto de testes	M-VQM	M-VQM_{In-ROI}	M-$VQM_{Out-ROI}$	DMOSp
Distorção Perceptual	0,74	0,43	0,78	0,78
Distorção de Canal	0,68	0,55	0,63	0,60
Distorção Perceptual + de Canal	0,70	0,54	0,70	0,67

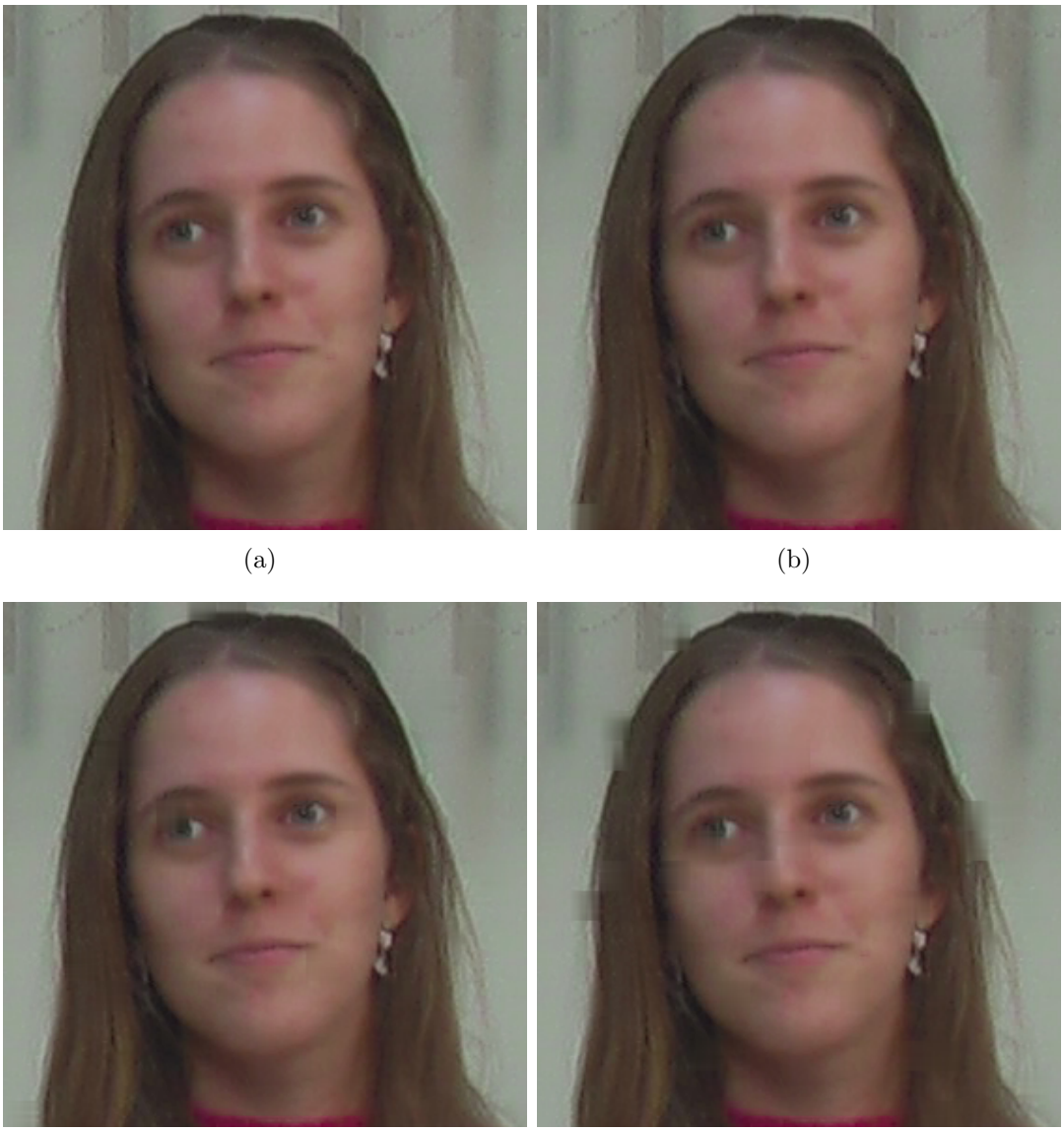
As alterações necessárias para adaptar o VQM ao método TWM não contribuíram para melhorar o seu desempenho frente a distorções de canal. Consideramos que este resultado esteja ligado à própria natureza do VQM que, de certa forma, já considera blocos de dados no tempo.

Desta forma consideramos que a combinação otimizada das notas baseadas em VQM não foram satisfatórias. Mais testes devem ser realizados para se descobrir a melhor forma de combinar os métodos propostos com o algoritmo VQM afim de obter os ganhos apresentados com o SSIM.

A estratégia de otimização em *k-folds* pode estar relacionada com o fato dos resultados das notas combinadas $DMOSp$ não apresentarem o ganho de desempenho esperado. Os valores apresentados são a média entre todos os *k-folds*, o que significa que o $DMOSp$ foi melhor do que as métricas isoladas para alguns *k-folds* e pior para outros. Entretanto, a segmentação do conjunto de sequências em "folds"equivalentes do ponto de vista estatístico é um problema bastante complexo. Um particionamento do conjunto de sequências em conjuntos estatisticamente mais equivalentes resultaria num melhor treinamento e, consequentemente, na melhoria do desempenho da métrica otimizada.

Outra possível causa de tão baixo desempenho pode estar relacionada às simulações geradas para os testes subjetivos. Em um cenário real de transmissão de videoconferência esperamos que as distorções causadas pela perda de blocos sejam mais perceptivas quando ocorrem nas regiões da face. As distorções no fundo da imagem devem ser imperceptíveis ou terem menor influência na avaliação da qualidade. Nas simulações de distorção de canal avaliamos que esta premissa não foi respeitada. Durante a realização dos testes subjetivos as sequências com níveis de 10%, e algumas de 5%, de erro de canal no exterior da região de interesse (fundo) apresentaram distorções que ficaram evidentes demais para o observador, não sendo consideradas muito realistas.

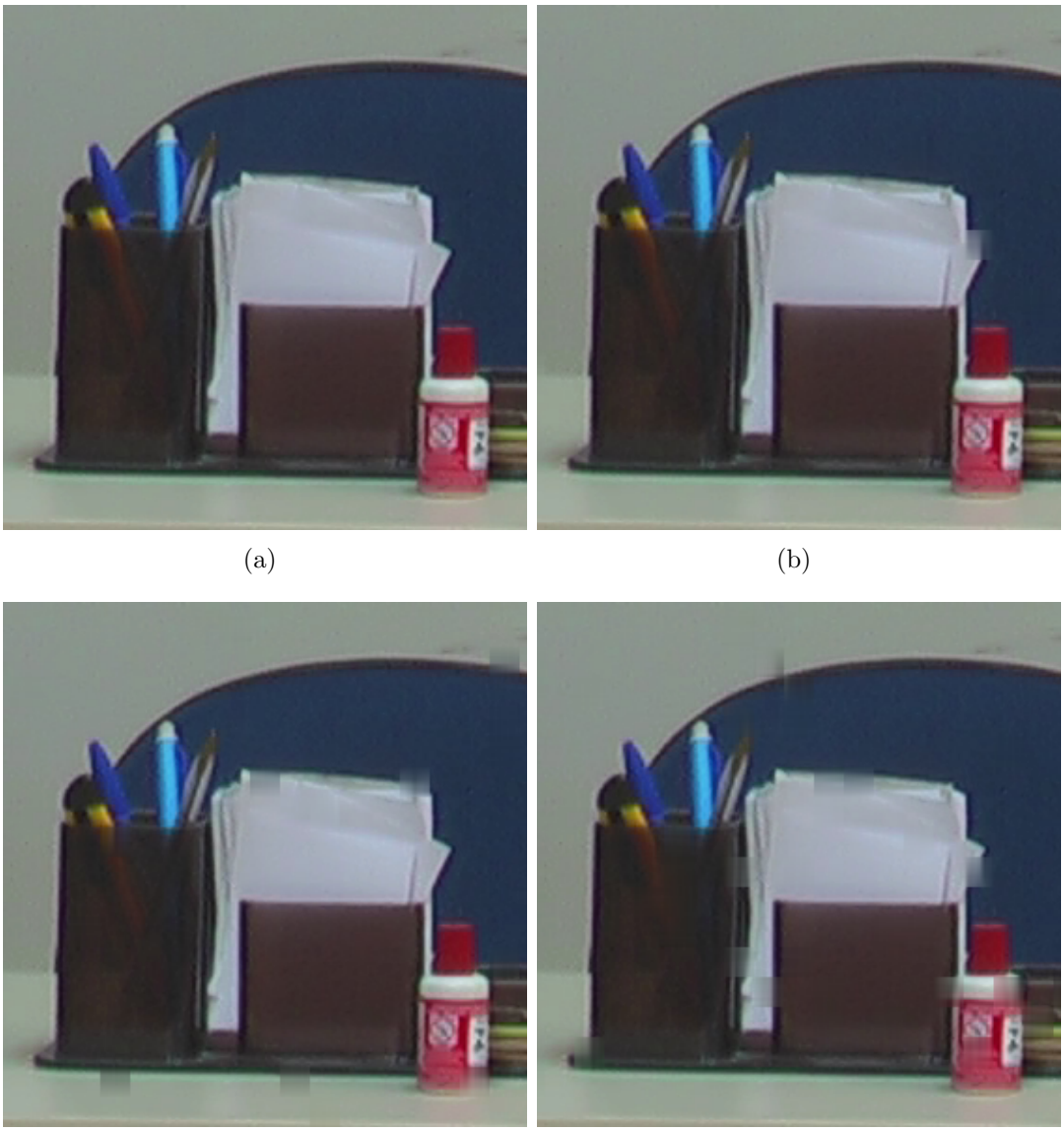
Na Figuras 5.7 e 5.8 são ilustradas os diferentes níveis de distorção em uma região ampliada da *In-Roi* e *Out-Roi* respectivamente. Nos casos 5.7(c) e 5.7(d) a distorção na região interna da ROI são evidentes. Nas simulações considerando a região externa da ROI, para a taxa de 1% de erro (5.8(b)) há pouca percepção de distorção da sequência. O mesmo não ocorre nas taxas de 5% (5.8(c)) e 10% (5.8(d)) onde as distorções passam a ser mais perceptíveis.



(a) (b)

(c) (d)

Figura 5.7: Níveis de erro *In-Roi*. (a) Região sem distorção. (b)*In-Roi* com 1% de erro. (c)*In-Roi* com 5% de erro. (d)*In-Roi* com 10% de erro.



(a) (b)

(c) (d)

Figura 5.8: Níveis de erro *Out-Roi*. (a) Região sem distorção. (b) *Out-Roi* com 1% de erro. (c) *Out-Roi* com 5% de erro. (d) *Out-Roi* com 10% de erro.

Capítulo 6

Banco de Sequências de Videoconferência com Alta Qualidade

Sistemas comerciais de alto nível de transmissão de vídeo devem atingir elevados níveis de satisfação dos usuários para oferecer uma experiência verdadeiramente realista. Tais sistemas devem enfrentar os compromissos entre os recursos do sistema e a qualidade oferecida aos usuários.

Ainda assim, pouco se sabe sobre os melhores pontos de operação para codificadores de alta definição em aplicações de transmissão de vídeo digital. Neste trabalho, utilizamos métricas do estado-da-arte de avaliação objetiva de qualidade com a finalidade de avaliar como diferentes configurações de codificação afetam a qualidade final em aplicações de videoconferência de alta definição.

Para tal experimento, um segundo banco de Sequências de Videoconferência com Alta Qualidade (SVAQ) foi criado a partir de gravações de cenas em alta definição (1080i) com muito pouca compressão (perceptualmente sem perdas) realizadas em um estúdio da emissora TV Futura utilizando equipamentos profissionais de alta qualidade. O banco contém 31 sequências de referência com as mesmas características do banco SVAD. Estas sequências estão disponíveis em [34]. Um quadro representativo de cada sequência pode ser visto no Apêndice B.

6.1 Configurações de codificação das SVAQ

Para as simulações, cada uma das 31 sequências de referência foi codificada com as configurações mostradas na Tabela 6.1, em um total de 31 x 45 = 1395 sequências degradadas. Para fins de comparação, as sequências codificadas foram convertidas de volta para sua condição nativa de exibição (resolução de 1080i e taxa de quadros

30Hz) usando interpolação espacial e temporal antes de serem avaliadas pela métrica. As notas finais obtidas para as 45 configurações são as médias do conjunto de 31 sequências.

Tabela 6.1: Configurações de codificação das SVAQ.

Taxa de Quadros	30 Hz			15 Hz			10 Hz		
Resolução	270p	540p	1080i	270p	540p	1080i	270p	540p	1080i
Taxa de bits mínima (Mbps)	0,05	0,05	0,25	0,05	0,05	0,3	0,05	0,05	0,1
Taxa de bits máxima (Mbps)	3	25	25	3	12	25	2.5	12	25

6.2 Avaliação Objetiva de Qualidade das SVAQ

Duas métricas de referência completa foram selecionadas a partir da literatura, devido à sua excelente correlação com as avaliações subjetivas de qualidade. O VQM, já descrito e utilizado em diversas etapas deste trabalho e o amplamente utilizado SSIM, introduzido na subseção 3.1.1. Embora a métrica SSIM tenha sido desenvolvida para avaliar a percepção de qualidade de imagens estáticas, pode facilmente ser estendida para as sequências de vídeo, bastanto calcular o SSIM para cada quadro do vídeo e a média entre eles. Apesar de não considerar as características temporais da sequência, como o VQM, sua vantagem sobre VQM é que tem uma complexidade computacional muito menor. Neste trabalho, embora a métrica VQM tenha sido considerada mais adequada do que SSIM para vídeo, o SSIM é muito útil para validar os resultados obtidos com o VQM. A métrica VQM utilizada foi a implementada neste trabalho sem a introdução da região de interesse nem a janela de mínimos. Para o calculo do SSIM foi utilizado o código de referência disponibilizado em [15].

Com base nas avaliações obtidas pelas métricas de qualidade analisamos os resultados e sugerirmos as melhores combinação de taxa de quadros e resolução a serem usadas para uma determinada taxa de bits, a fim de obter a melhor qualidade visual.

6.3 Resultados com o banco SVAQ

As Figuras 6.1, 6.2 e 6.3 contêm os gráficos Taxa-Distorção das sequências codificadas a 10, 15 e 30 quadros por segundos, respectivamente. Observamos que o comportamento entres elas se mantém para todas as sequências. Para uma taxa fixa de quadros por segundos, a medida que aumentamos a taxa de bits, as sequências de maior resolução passam a ter maior qualidade.

As Figuras 6.4, 6.5 e 6.6 contêm os gráficos Taxa-Distorção das sequências codificadas com quadros de resolução 480x270, 960x540 e 1920x1080, respectivamente. Os gráficos possuem um comportamento muito parecido com os de taxa de quadros

fixa. Para um resolução de quadro fixa, com o aumento da taxa de bits obtemos maior qualidade em sequências com maior taxa de quadros.

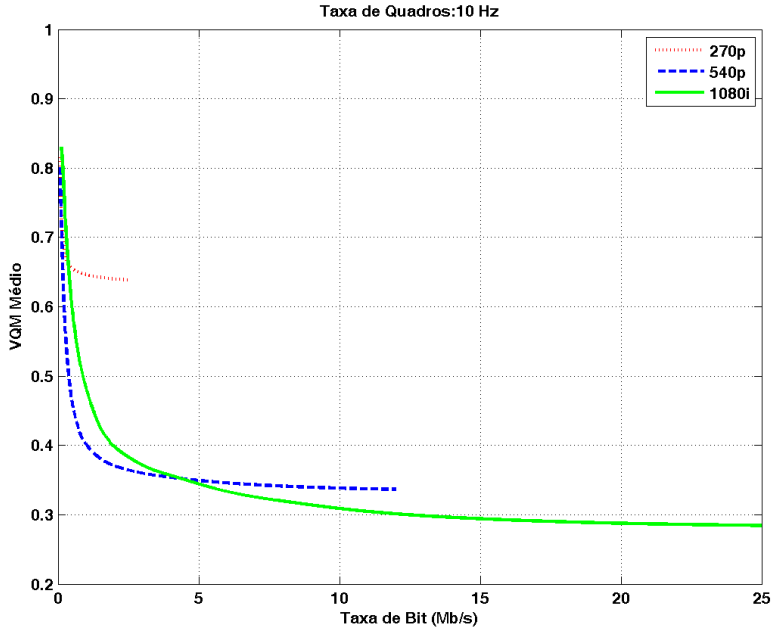


Figura 6.1: Curvas Taxa-Distorção (VQM) das sequências com 10 Hz.

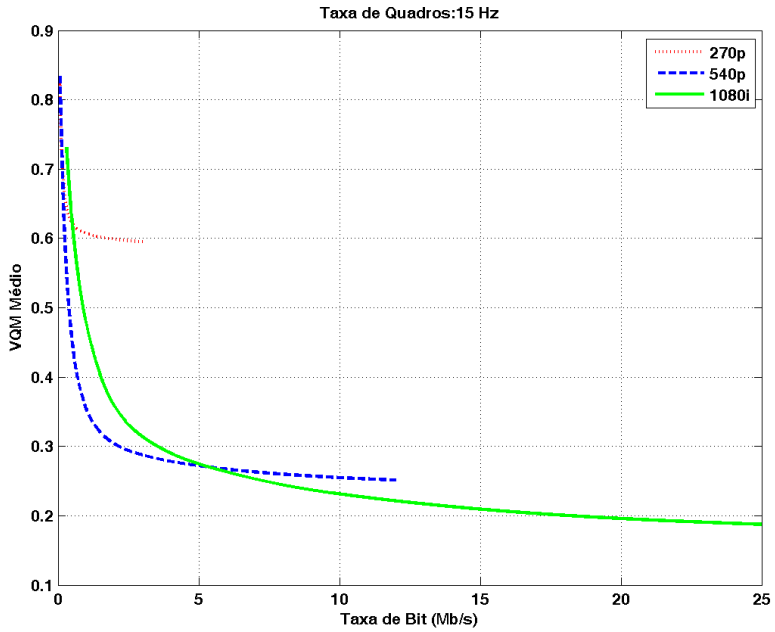


Figura 6.2: Curvas Taxa-Distorção (VQM) das sequências com 15 Hz.

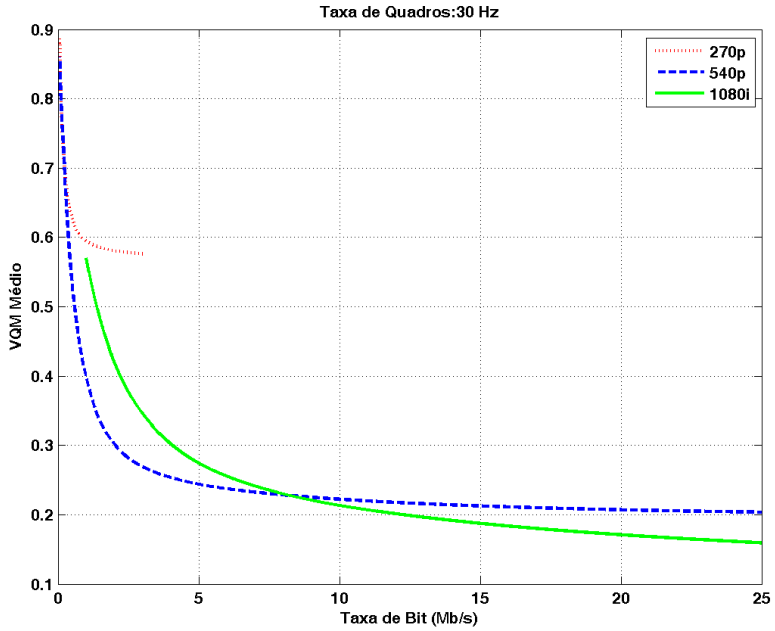


Figura 6.3: Curvas Taxa-Distorção (VQM) das sequências com 30 Hz.

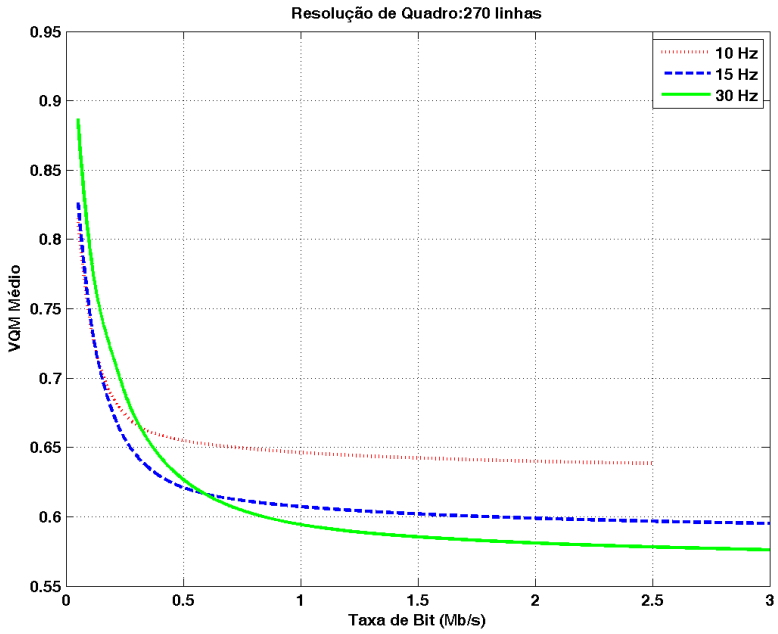


Figura 6.4: Curvas Taxa-Distorção (VQM) das sequências 270p.

Os resultados mostraram que a taxa de quadros tende a ser mais importante do que a resolução espacial em termos de qualidade percebida. Ao contrário do que é normalmente esperado, quando há requisição de baixa taxa de transmissão (devido às limitações inerentes do canal) para a aplicação de videoconferências de alta definição, melhores resultados globais de qualidade são obtidos através da redução

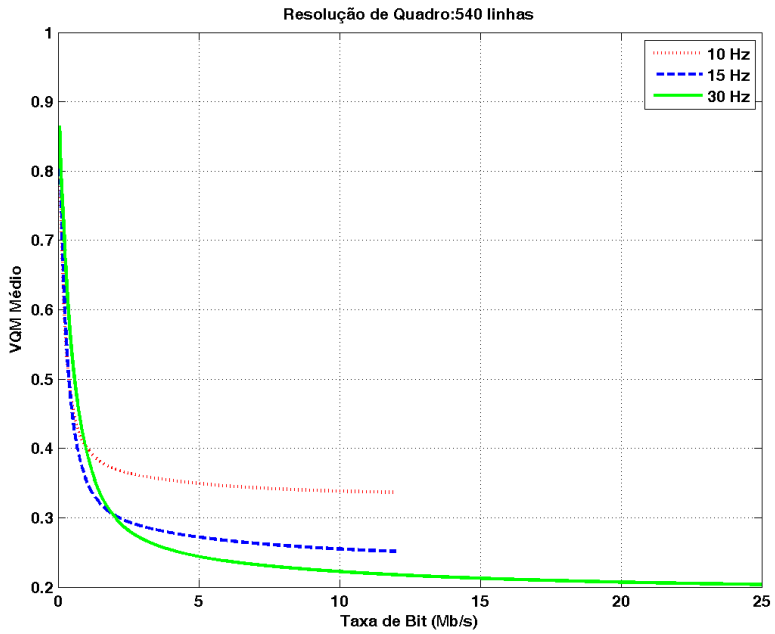


Figura 6.5: Curvas Taxa-Distorção (VQM) das sequências 540p.

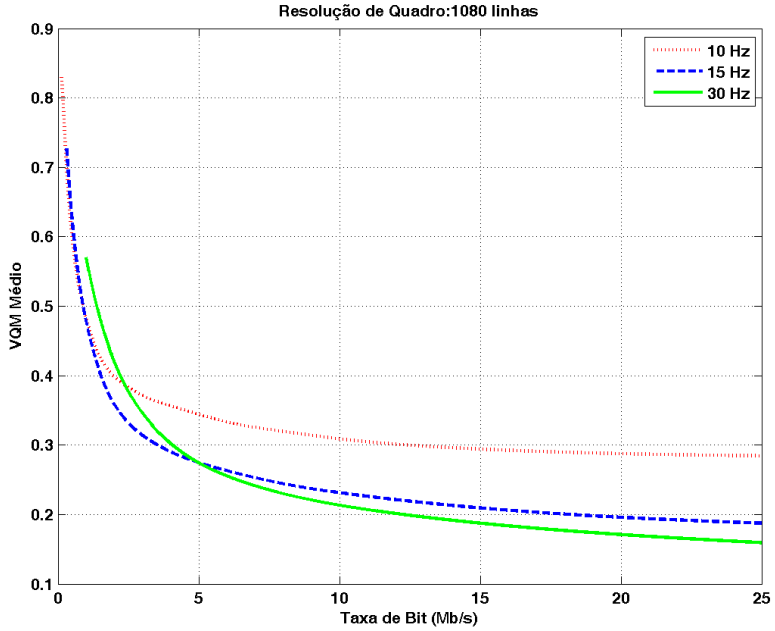


Figura 6.6: Curvas Taxa-Distorção (VQM) das sequências 1080i.

da resolução espacial das sequências de vídeo para 540p e resolução temporal de 10 quadros por segundo antes de comprimir e transmiti-los através do canal, ao invés de aumentar a taxa de bits na resolução máxima.

À medida que a taxa de bits aumenta, é melhor aumentar gradualmente a taxa de quadros para 30 quadros por segundo para só então aumentar a resolução espacial para 1080i. Note que a resolução de 270p não se mostrou útil para esta aplicação.

Além disso, analisando a Figura 6.7, observamos que para uma dada resolução e qualidade, quando a qualidade é baixa, as taxas de quadros menores tendem a dar taxas de bits menores. A medida que o grau de qualidade aumenta, em primeiro lugar a taxa de quadros 10 Hz dá lugar à 15 Hz e então para 30 Hz. Para índices de qualidade mais elevados, taxas de quadros maiores são os que dão taxas de bits menores. Tais conclusões podem fornecer uma orientação útil para o projeto de tais aplicações.

Os resultados finais mostraram que para aplicações com limitações de largura de banda, uma melhor qualidade geral pode ser obtida através da redução da resolução espacial e temporal das sequências para 540p a 10 quadros por segundo e aumentando gradualmente a taxa de quadros antes de aumentar a resolução espacial, a medida que restrições de taxa de bits sejam menos limitadas.

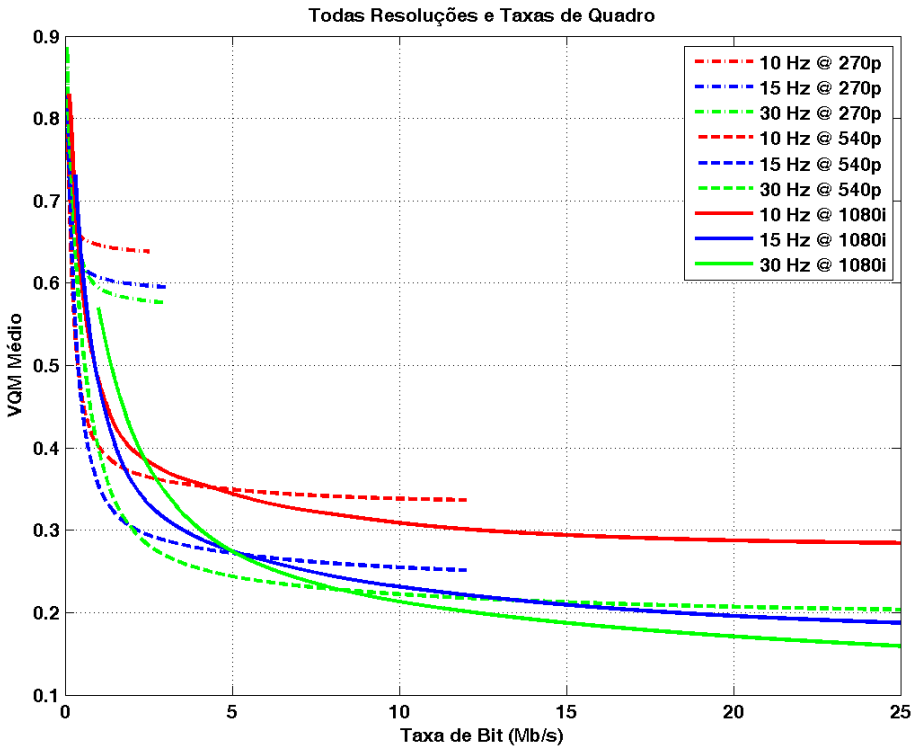


Figura 6.7: Curvas Taxa-Distorção (VQM) das SVAQ.

Nas Tabelas 6.2 e 6.3 encontram-se as melhores configurações obtidas nas simulações baseadas no VQM e SSIM, respectivamente. Analisado as tabelas observamos que baseado no VQM as mudanças entre as melhores configurações ocorrem para faixas diferentes de taxas de bits que as obtidas com o SSIM. Os resultados com o SSIM servem para confirmar as conclusões retiradas dos resultados com o VQM.

Mesmo apresentando valores específicos diferentes, os resultados com SSIM apre-

sentam a mesma tendência considerando o VQM. À medida que a largura de banda disponível aumenta, a melhor qualidade de vídeo é obtida com um aumento gradativo da taxa de quadros para só então aumentar a resolução do quadro.

Tabela 6.2: Melhor configuração baseada em notas VQM

	Taxa de Bits (Mbs/s)			
Faixa	$0 \rightarrow 0.4$	$0.4 \rightarrow 2.0$	$2.0 \rightarrow 8.0$	$8.0 \rightarrow 25.0$
Melhor configuração	540p @ 10 Hz	540p @ 15 Hz	540p @ 30 Hz	1080i @ 30 Hz

Tabela 6.3: Melhor configuração baseada em notas SSIM

	Taxa de Bits (Mbs/s)			
Faixa	$0 \rightarrow 0.5$	$0.5 \rightarrow 1.0$	$1.0 \rightarrow 4.4$	$4.4 \rightarrow 25.0$
Melhor configuração	540p @ 10 Hz	540p @ 15 Hz	1080i @ 15 Hz	1080i @ 30 Hz

Capítulo 7

Conclusões

Neste trabalho abordamos o tema da avaliação objetiva de qualidade de vídeo visando principalmente a aplicação em sistemas de videoconferência em alta definição. Resumimos as abordagens tradicionais de avaliação de qualidade de vídeo, para depois elaborar uma nova abordagem que apresentasse um ganho no desempenho das métricas de avaliação objetiva.

Dois extensos bancos de sequências com características típicas de videoconferência foram gerados para realização dos nossos testes. Um banco de sequências em alta definição [30] foi gerado um total de 23 sequências de referência e mais de 1200 sequências de teste contendo distorções perceptuais (brilho e contraste) e de erro de canal. O segundo banco contendo 30 sequências de referência em alta definição e maior qualidade [34] foi utilizado para gerar 1395 sequências de testes a partir de diferentes configurações de codificação (taxa de bits, taxa de quadros e resolução). Em conjunto com as notas DMOS obtidas dos testes subjetivos realizados, passamos a possuir um banco dados de avaliação de qualidade que permitirá pesquisadores avaliar o desempenho de algoritmos de medida de qualidade em trabalhos futuros.

As métricas FR recomendadas em [2] foram estudas e implementadas, fornecendo um importante conhecimento sobre diferentes métodos estado-da-arte de avaliação objetiva de qualidade de vídeo. A possibilidade da combinação de alguns desses métodos em conjunto com novas abordagens mostra-se atraente para o desenvolvimento futuro de métricas mais eficientes.

Propomos uma nova estratégia de avaliação de qualidade baseada na propriedade que certas regiões de uma imagem podem ser visualmente mais importantes que outras. Métricas objetivas que ponderam os índices de qualidade nestas regiões de interesse indicaram obter maior correlação com avaliações subjetivas.

Para aplicações de videoconferência foi assumido que as regiões de interesse natural seriam as regiões de face do vídeo. Como contribuição desta tese, técnicas de detecção de face foram incorporadas ao método VQM para selecionar as regiões de interesse do vídeo. Durante nossos testes percebemos que distorções em regiões

de alta frequência desviam a atenção inicial do observador. Para compensar esta propriedade foi incorporado ao método uma segunda região de interesse baseada na informação de frequência.

Também mostramos que métricas de avaliação de qualidade típicas como o VQM não são adequadas para lidar com erros de codificação, tendo uma queda considerável de desempenho quando erros de canal estão presentes. Para compensar essa queda de desempenho elaboramos o método de Janela Temporal de Mínimos que considera os efeitos da resistência retiniana do erro.

Comparamos o desempenho da métrica proposta contra a métrica VQM, usando o banco de sequências em alta definição contendo uma variedade de cenários de distorção. Os resultados indicam que os métodos propostos podem oferecer uma melhor correlação com as notas subjetivas, porém há muito o que melhorar na implementação destes métodos.

Por fim, realizamos estudos dos efeitos de parâmetros de codificação (taxa de bits, taxa de quadros e resolução) na percepção de qualidade. Com os resultados obtemos os pontos de melhor desempenho em termos de qualidade para diversas faixas de taxa de bits.

7.1 Trabalhos Futuros

A seguir listamos as considerações a serem utilizadas em trabalhos futuros.

- Combinação dos métodos apresentados em [2] para o desenvolvimento de uma métrica mais eficiente.
- A utilização em nosso método de algoritmos mais robustos e eficientes de detecção de face que o Viola-Jones.
- Melhorias no método de Janela Temporal de Mínimos considerando o efeito de "perdão "(*Forgiveness Effect*) apresentado em [35].
- A incorporação do método de pontuação percentual de [25] ao método proposto nessa dissertação.
- Uso de degradações de canal mais realistas (menores taxas de distorção).
- A realização de testes subjetivos e objetivos para investigar a influência do erros de canal em diversas configurações de codificação.

Referências Bibliográficas

[1] VQEG. *Final report from the video quality experts group on the validation of objective models of video quality assessment.* Relatório técnico, Março 2000. Disponível em: <`http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseI/`>.

[2] ITU-T. "Objective perceptual video quality measurement techniques for digital cable television in the presence of a full reference." Rec. J.144, 2004.

[3] ITU-R. "Methodology for the subjective assessment of the quality of television pictures." Rec. BT.500, 2000.

[4] ITU-T. "Subjective video quality assessment methods for multimedia applications." Rec. P.910, 1999.

[5] WINKLER, S. *Digital Video Quality: Vision Models and Metrics.* Wiley, Março 2005.

[6] BOVIK, A. C., WANG, Z., SHEIKH, H. R. "Objective video quality assessment". In: *The Handbook of Video Databases: Design and Applications*, pp. 1041–1078. CRC Press, Setembro 2003.

[7] WANDELL, B. A. *Foundations of Vision.* Sinauer Associates, Inc., 1995.

[8] BOVIK, A. C., CLARK, M., GEISLER, W. S. "Multichannel Texture Analysis Using Localized Spatial Filters", *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 12, pp. 55–73, Janeiro 1990.

[9] WEBER, E. H. *Die Lehre vom Tastsinn und Gemeingefühl, auf Versuche gegründet.* Vieweg, 1851.

[10] RIDDER, H. "Minkowski-metrics as a combination rule for digital-image-coding impairments". v. 1666, pp. 16–26. SPIE, 1992.

[11] WANG, Z., BOVIK, A. C., SHEIKH, H. R., et al. "Image Quality Assessment: From Error Visibility to Structural Similarity", *IEE Transactions on Image Processing*, v. 13, n. 4, pp. 600–612, 2004.

[12] DALY, S. J. "Visible Differences Predictor: An Algorithm for The Assessment of Image Fidelity". v. 1666, pp. 2–15. SPIE, 1992.

[13] LUBIN, J. "A Visual Discrimination Model for Imaging System Design and Evaluation". In: Peli, E. (Ed.), *Vision Models for Target Detection and Recognition*, World Scientific, pp. 245–283, 1995.

[14] WINKLER, S. "A Perceptual Distortion Metric for Digital Color Video". In: *Proceedings of the SPIE - Human Vision and Electronic Imaging*, v. 3644, pp. 175–184, San Jose, CA, Janeiro 1999.

[15] WANG, Z. "Structural Similarity (SSIM) Index Software". Disponível em: <https://ece.uwaterloo.ca/~z70wang/research/ssim/>.

[16] VQEG. "Validation of Reduce-Reference and No-Reference Objetive Models for Standard Definition Television, Phase I". Junho 2009. Disponível em: <http://www.its.bldrdoc.gov/vqeg/projects/rrnr-tv/>.

[17] WEBSTER, A. A., JONES, C. T., PINSON, M. H., et al. "An Objective Video Quality Assessment System Based on Human Perception". In: *SPIE Human Vision, Visual Processing and Digital Display IV*, pp. 15–26, 1993.

[18] GUNAWAN, I. P., GHANBARI, M. "Image Quality Assessment Based on Harmonics Gain/Loss Information". In: *ICIP05*, pp. I: 429–432, 2005.

[19] FARIAS, M. C. Q., MITRA, S. K. "No-Reference Video Quality Metric Based on Artifact Measurements". In: *ICIP05*, pp. III: 141–144, 2005.

[20] CIANCIO, A., DA COSTA, A. L. N. T., DA SILVA, E. A. B., et al. "Objective no-reference image blur metric based on local phase coherence", *IEE Electronics Letters*, v. 45, pp. 1162–1163, Novembro 2009.

[21] VQEG. *Final Report From The Video Quality Experts Group on The Validation of Objective Models of Video Quality Assessment, Phase II*. Relatório técnico, Agosto 2003. Disponível em: <http://www.its.bldrdoc.gov/vqeg/projects/frtv_phaseII>. Technical report.

[22] WOLF, S., PINSON, M. H. "A New Standardized Method for Objectively Measuring Video Quality Abstract". In: *IEEE Transactions on Broadcasting*, v. 50, pp. 312–322, Setembro 2004.

[23] ITS. "Video Quality Metric (VQM) Software". Disponível em: <www.its.bldrdoc.gov/n3/video/vqmsoftware.htm>.

[24] BRADSKI, G. "The OpenCV Library", *Dr. Dobb's Journal of Software Tools*, 2000.

[25] MOORTHY, A. K., BOVIK, A. C. "Visual Importance Pooling for Image Quality Assessment". In: *IEEE journal of Selected Topics in Signal Processing*, pp. 193–201, 2009.

[26] WANG, Z., SHANG, X. "Spatial Pooling Strategies for Perceptual Image Quality Assessment". In: *IEEE International Conference on Image Processing*, pp. 2945–2948, 2006.

[27] VIOLA, P., JONES, M. "Rapid object detection using a boosted cascade of simple features". In: *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, pp. I–511–I–518. IEEE Comput. Soc, Abril 2001.

[28] ASTOLA, J., HAAVISTO, P., NEUVO, Y. "Vector median filters", *Proceedings of the IEEE*, v. 78, n. 4, pp. 678–689, 1990.

[29] LPS. "Laboratório de Processamento de Sinais - COPPE/Poli/UFRJ". Disponível em: <`https://www.lps.ufrj.br`>.

[30] "HDVC Database - High Definition Video Conferencing Database". . Disponível em: <`http://www02.lps.ufrj.br/~tvdigital/muque/data/lps`>.

[31] MOCHNAC, J., MARCHEVSKY, S. "Error concealment scheme implemented in H.264/AVC". In: *50th International Symposium ELMAR*, v. 1, pp. 13–16, Setembro 2008.

[32] STONE, M. "Cross-validatory choice and assessment of statistical predictions". In: *Journal of the Royal Statistical Society B (Methodological)*, v. 36, pp. 111–147, 1974.

[33] KOHAVI, R. "A study of cross-validation and bootstrap for accuracy estimation and model selection". In: *Proceedings of the International Joint Conference on Artificial Inteligence*, Agosto 1996.

[34] "HQVC Database - High Quality Video Conferencing Database". . Disponível em: <`http://www02.lps.ufrj.br/~tvdigital/muque/data/futura`>.

[35] LIU, T., WANG, Y., BOYCE, J. M., et al. "A Novel Video Quality Metric for Low Bit-rate Video Considering both Coding and Packet-loss Artifacts". In: *IEEE Journal of Selected Topics in Signal Processing*, v. 2, pp. 280–293, Abril 2009.

Apêndice A

Banco de Sequências de Videoconferência com Alta Definição (SVAD)

Figura A.1: Sequência 01 - Fundo plano; indivíduo do sexo masculino.

Figura A.2: Sequência 02 - Fundo plano; oclusão da face.

Figura A.3: Sequência 03 - Fundo plano; indivíduo com óculos.

Figura A.4: Sequência 04 - Fundo plano; indivíduo com barba.

Figura A.5: Sequência 05 - Fundo plano; indivíduo do sexo feminino.

Figura A.6: Sequência 06 - Fundo simples; oclusão da face.

Figura A.7: Sequência 07 - Fundo simples; rotação da face.

Figura A.8: Sequência 08 - Fundo simples; indivíduo do sexo feminino; oclusão da face.

Figura A.9: Sequência 09 - Fundo simples; indivíduo do sexo feminino; rotação da face.

Figura A.10: Sequência 10 - Fundo simples; indivíduo com óculos.

Figura A.11: Sequência 11 - Fundo simples; indivíduo com óculos; rotação da face.

Figura A.12: Sequência 12 - Fundo simples; indivíduo com barba.

Figura A.13: Sequência 13 - Fundo simples; indivíduo com barba e óculos.

Figura A.14: Sequência 14 - Fundo simples; indivíduo de pele escura.

Figura A.15: Sequência 15 - Fundo simples; dois indivíduos.

Figura A.16: Sequência 16 - Fundo simples; dois indivíduos; rotação da face.

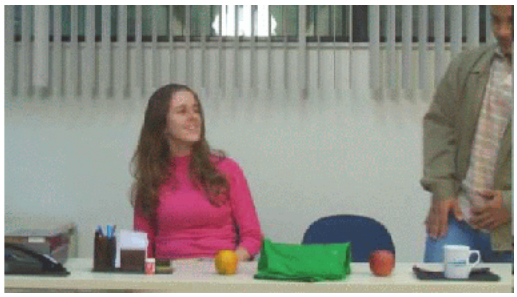


Figura A.17: Sequência 17 - Fundo simples; começa com uma pessoa e outra entra na cena.

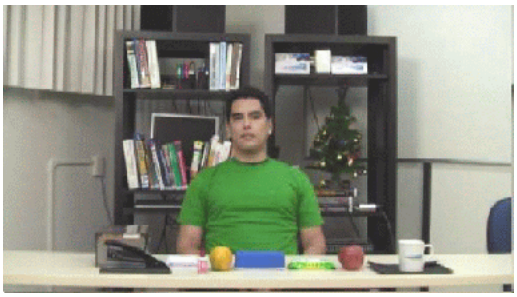


Figura A.18: Sequência 18 - Fundo complexo; indivíduo do sexo masculino.

Figura A.19: Sequência 19 - Fundo complexo; rotação da face.

Figura A.20: Sequência 20 - Fundo complexo; oclusão da face.

Figura A.21: Sequência 21 - Fundo complexo; dois indivíduos.

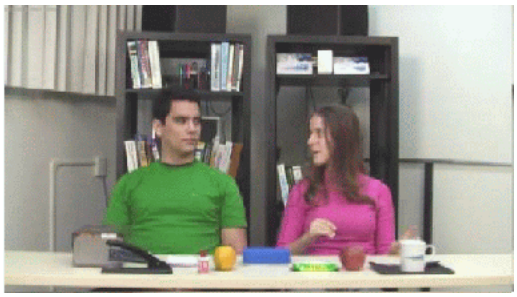


Figura A.22: Sequência 22 - Fundo complexo; dois indivíduos; rotação da face.

Apêndice B

Banco de Sequências de Videoconferência com Alta Qualidade (SVAQ)

Figura B.1: Sequência 01 - Fundo liso; indivíduo do sexo masculino.

Figura B.2: Sequência 02 - Fundo liso; oclusão da face.

Figura B.3: Sequência 03 - Fundo liso; indivíduo do sexo feminino com óculos.

Figura B.4: Sequência 04 - Fundo liso; camisa listrada (Efeito de Moiré).

Figura B.5: Sequência 05 - Fundo liso; indivíduo do barba.

Figura B.6: Sequência 06 - Fundo liso; indivíduo do sexo feminino.

Figura B.7: Sequência 07 - Fundo liso; indivíduo de pele escura.

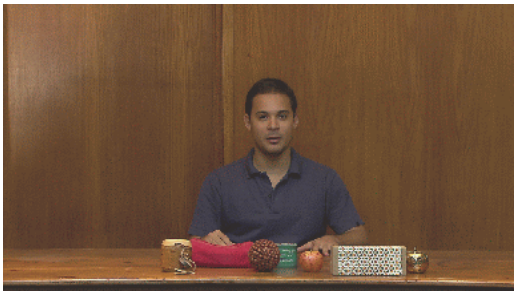


Figura B.8: Sequência 08 - Fundo simples; oclusão da face.

Figura B.9: Sequência 09 - Fundo simples; camisa listrada.

Figura B.10: Sequência 10 - Fundo simples; camisa listrada, rotação da face.

Figura B.11: Sequência 11 - Fundo simples; indivíduo do sexo feminino.

Figura B.12: Sequência 12 - Fundo simples; indivíduo do sexo feminino, oclusão da face.

Figura B.13: Sequência 13 - Fundo simples; indivíduo do sexo feminino, oclusão da face.

Figura B.14: Sequência 14 - Fundo simples; indivíduo do sexo feminino com óculos.

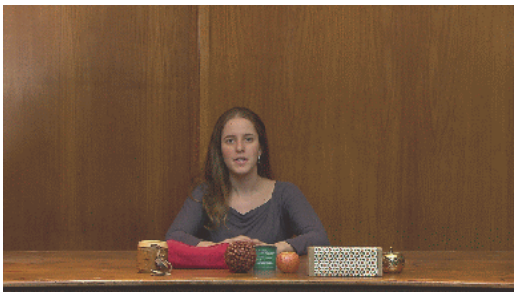


Figura B.15: Sequência 15 - Fundo simples; indivíduo do sexo feminino, rotação da face

Figura B.16: Sequência 16 - Fundo simples; indivíduo com barba.

Figura B.17: Sequência 17 - Fundo simples; indivíduo de pele escura.

Figura B.18: Sequência 18 - Fundo simples; dois indivíduos.

Figura B.19: Sequência 19 - Fundo simples; dois indivíduos, rotação da face.

Figura B.20: Sequência 20 - Fundo simples; começa com uma pessoa e outra entra na cena.

Figura B.21: Sequência 21 - Fundo simples; rotação da face, manipulando objetos.

Figura B.22: Sequência 22 - Fundo simples; manipulando objetos.

Figura B.23: Sequência 23 - Fundo simples; manipulando objetos.

Figura B.24: Sequência 24 - Fundo simples; manipulando objetos.

Figura B.25: Sequência 25 - Fundo complexo; movimentos de mão.

Figura B.26: Sequência 26 - Fundo complexo; rotação da face.

Figura B.27: Sequência 27 - Fundo complexo; oclusão da face.

Figura B.28: Sequência 28 - Fundo complexo; dois indivíduos.

Figura B.29: Sequência 29 - Fundo complexo; dois indivíduos; rotação da face.

Figura B.30: Sequência 30 - Fundo complexo; indivíduo de pele escura.

Figura B.31: Sequência 31 - Fundo complexo; indivíduo com barba.

Apêndice C

Sequências de teste do banco SVAD

Tabela C.1: Sequências de testes.

Sequência	Região de Interesse	Blocos com erro (%)	Distorção de Brilho (%)	Distorção de Contraste (%)	Testes Subjetivos
001	Sem	0	-75	0	Não
002	Sem	0	-50	0	Sim
003	Sem	0	-25	0	Sim
004	Sem	0	25	0	Sim
005	Sem	0	50	0	Sim
006	Sem	0	75	0	Não
007	Sem	0	0	-75	Não
008	Sem	0	0	-50	Sim
009	Sem	0	0	-25	Sim
010	Sem	0	0	25	Sim
011	Sem	0	0	50	Sim
012	Sem	0	0	75	Não
013	Sem	1	0	0	Sim
014	Sem	5	0	0	Sim
015	Sem	10	0	0	Sim
016	Interior	1	0	0	Sim
017	Interior	5	0	0	Sim
018	Interior	10	0	0	Sim
019	Exterior	1	0	0	Sim
020	Exterior	5	0	0	Sim
021	Exterior	10	0	0	Sim
022	Sem	5	-75	0	Não
023	Sem	5	-50	0	Não
024	Sem	5	-25	0	Não
025	Sem	5	25	0	Não
026	Sem	5	50	0	Não
027	Sem	5	75	0	Não
028	Interior	5	-75	0	Não
029	Interior	5	-50	0	Não
030	Interior	5	-25	0	Sim
031	Interior	5	25	0	Sim
032	Interior	5	50	0	Não
033	Interior	5	75	0	Não
034	Exterior	5	-75	0	Não
035	Exterior	5	-50	0	Não
036	Exterior	5	-25	0	Sim
037	Exterior	5	25	0	Sim
038	Exterior	5	50	0	Não
039	Exterior	5	75	0	Não
040	Sem	5	0	-75	Não
041	Sem	5	0	-50	Não
042	Sem	5	0	-25	Não
043	Sem	5	0	25	Não
044	Sem	5	0	50	Não
045	Sem	5	0	75	Não
046	Interior	5	0	-75	Não
047	Interior	5	0	-50	Não
048	Interior	5	0	-25	Sim
049	Interior	5	0	25	Sim
050	Interior	5	0	50	Não
051	Interior	5	0	75	Não
052	Exterior	5	0	-75	Não
053	Exterior	5	0	-50	Não
054	Exterior	5	0	-25	Sim
055	Exterior	5	0	25	Sim
056	Exterior	5	0	50	Não
057	Exterior	5	0	75	Não

Apêndice D

Procedimentos de Otimização das Notas

Este apêndice mostra o procedimento, em linguagem Matlab©, para otimização entre o conjunto de treinamento das notas objetivas e o DMOS correspondente.

O conjunto de treinamento das notas objetivas é linearizado em relação ao conjunto de notas DMOS correspondente. A função de mapeamento exponencial escolhida para a linearização das métricas individuais foi:

```
fittype ( 'a/( 1+exp( -b*( c*x-d )))' )
```

Um processo de regressão não-linear é aplicado ao conjunto de notas linearizadas. As funções consideradas na estratégia de otimização foram 5.

```
% Funções de otimização conjunta
% j            índice da interação
% x(n)       índice do vetor de parâmetros que será otimizado
% G_No     vetor de notas objetivas sem ROI linearizadas
% G_In      vetor de notas objetivas In-ROI linearizadas
% G_Out    vetor de notas objetivas Out-ROI linearizadas
% DMOSp  nota combinada usando os valores ótimos de parâmetros x

if OPT == 1,

  y = x(1)*G_In + x(2)*G_Out+ (1-x(1)-x(2))*G_No;

  DMOSp_VQM{j} = x(3)./(1+exp(-x(4)*(x(5).*y-x(6))));

elseif OPT == 2,
```

```
     GNop  = x(1)./(1+exp(-x(2)*(x(3).*G_No-x(4))));
     GInp  = x(5)./(1+exp(-x(6)*(x(7).*G_In-x(8))));
     GOutp = x(9)./(1+exp(-x(10)*(x(11).*G_Out-x(12))));

     DMOSp{j} = x(13)*GInp + x(14)*GOutp+ x(15)*GNop;

elseif OPT == 3,

     DMOSp{j} = x(1)*G_In + x(2)*G_Out + x(3)*G_No;

elseif OPT == 4,

     y = x(1)*G_In + x(2)*G_Out+ x(3)*G_No;

     DMOSp{j} = x(4)*y.^3 + x(5)*y.^2 + x(6)*y + x(7);

elseif OPT == 5,

     GNop = x(1)*G_No.^3+x(2)*G_No.^2+x(3)*G_No+x(4);
     GInp = x(5)*G_In.^3+x(6)*G_In.^2+x(7)*G_In+x(8);
     GOutp = x(9)*G_Out.^3+x(10)*G_Out.^2+x(11)*G_Out+x(12);

     DMOSp{j} = x(13)*GInp + x(14)*GOutp+ x(15)*GNop;

end
```

Depois das interações de otimização, chegamos ao valor ótimo de parâmetros x. DMOSp é a nota combinada usando os valores ótimos de parâmetros.