Universidade Federal de Campina Grande
Centro de Engenharia Elétrica e Informática
Programa de Pós-Graduação em Engenharia Elétrica

Tese de Doutorado

Métrica de Avaliação Objetiva de Vídeo Usando a Informação Espacial, a Temporal e a Disparidade

Aluno: Carlos Danilo Miranda Regis

Orientador: Marcelo Sampaio de Alencar

Campina Grande – PB

Março – 2013

Métrica de Avaliação Objetiva de Vídeo Usando a Informação Espacial, a Temporal e a Disparidade

Carlos Danilo Miranda Regis

Tese de Doutorado submetida à Coordenação dos Cursos de Pós-Graduação em Engenharia Elétrica da Universidade Federal de Campina Grande como parte dos requisitos necessários para obtenção do grau de Doutor no domínio da Engenharia Elétrica.

Área de Concentração: Processamento da Informação

Marcelo Sampaio de Alencar, Ph.D., UFCG
Orientador

Campina Grande, Paraíba, Brasil

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA CENTRAL DA UFCG

R337m Regis, Carlos Danilo Miranda.

Métrica de avaliação objetiva de vídeo usando a informação espacial, a temporal e a disparidade / Carlos Danilo Miranda Regis. – Campina Grande, 2013.

138 f. : color.

Tese (Doutorado em Engenharia Elétrica) - Universidade Federal de Campina Grande, Centro de Engenharia Elétrica e Informática, 2013.

"Orientação: Prof. Ph.D. Marcelo Sampaio de Alencar".

Referências.

1. Vídeos Digitais. 2. Métricas Objetivas. 3. Informação Espacial. 4. Qualidade Visual. 5. Informação Temporal. I. Alencar, Marcelo Sampaio de. II. Título.

CDU 621.397(043)

"MÉTRICA DE AVALIAÇÃO OBJETIVA DE VÍDEO USANDO A INFORMAÇÃO ESPACIAL, A TEMPORAL E A DISPARIDADE"

CARLOS DANILO MIRANDA REGIS

TESE APROVADA EM 22/03/2013

MARCELO SAMPAIO DE ALENCAR, Ph.D., UFCG
Orientador(a)

ALEXANDRE DE ALMEIDA PRADO POHL, Dr.Ing., UTFPR
Examinador(a)

FRANCISCO MADEIRO BERNARDINO JÚNIOR, D.Sc, UPE
Examinador(a)

MYLENE CHRISTINE QUEIROZ FARIAS, Ph.D, UnB
Examinador(a)

JOSEANA MACEDO FECHINE RÉGIS DE ARAÚJO, D.Sc., UFCG
Examinador(a)

WASLON TERLLIZZIE ARAÚJO LOPES, D.Sc., UFCG
Examinador(a)

CAMPINA GRANDE - PB

À Deus,

à minha esposa Gabriella,

aos meus pais, Roberto e Carla,

irmãos Caio e Tiago e

*ao meu avô João Regis (*in memorian*).*

Agradecimentos

À Deus que sempre está comigo, que me orientou e deu as habilidades necessárias para a conclusão desta tese. À minha esposa Ana Gabriella, que é a mulher virtuosa de Provérbios 31.10-12, em reconhecimento aos muitos momentos roubados do nosso convívio durante a realização deste trabalho. Muito obrigado pela compreensão e apoio.

Aos meus pais, Roberto e Carla, quero honrá-los pela confiança, carinho e ensinamentos. Assim como meus irmãos, Caio e Tiago, e meus familiares que sempre torceram por mim. Em especial, ao meu avô João Francisco Regis (*in memorian*).

Ao meu orientador Marcelo Sampaio pela dedicação na orientação deste trabalho. Obrigado pelas experiências proporcionadas e pelas lições aprendidas: elas serão de extrema importância por toda a minha vida.

Aos amigos do Iecom: Rafael Fernandes, Paulo Ribeiro, Raissa Rocha e Jean Oliveira. Agradeço pelo incentivo e apoio prestados durante todos esses anos. Aos meus alunos e ex-alunos: José Vinícius, Mikaelle, Ítalo e Nathália que de forma direita e indireta contribuíram para esse trabalho.

Ao CNPq por financiar este trabalho, ao Iecom por fornecer a infra-estrutura necessária. À todos que fazem a PPgEE pelo apoio constante. Agradeço também à UFCG.

Resumo

Vídeos digitais estão propensos a vários tipos de distorções que podem ocorrer durante os processos de aquisição, processamento, compressão, armazenamento e transmissão, resultando em perda da qualidade visual. A avaliação da qualidade de vídeo é importante para definir parâmetros em sistemas e propor a redução da degradação.

A maneira mais apropriada para verificar a qualidade visual de uma sequência de vídeo é a avaliação subjetiva, porém a avaliação objetiva é mais rápida e tem menor custo em relação às soluções subjetivas.

O desenvolvimento de métricas objetivas tem crescido significativamente, e isso se deve à necessidade de quantificar, de forma adequada e rápida, a qualidade visual percebida pelo sistema visual humano.

A produção de vídeos em três dimensões vem crescendo, assim como a necessidade de avaliá-los. Um problema encontrado para a avaliação é que os métodos usados para os vídeos em 2D não são adequados para medir a qualidade de vídeos em 3D, porque a profundidade e as distorções típicas da estereoscopia não são consideradas.

Esta tese propõe novas métricas de avaliação de vídeo em duas e três dimensões. Essa métrica usa a informação espacial e temporal e a disparidade para avaliar os vídeos de forma mais próxima das avaliações subjetivas.

Para obter a métrica para vídeos em três dimensões, foram propostas medidas para vídeos em duas dimensões que utilizam a informação espacial (PW-SSIM – *Perceptual Weighted Video Quality Approach*) e temporal (TPW-SSIM – *Temporal Perceptual Weighted Video Quality Approach*). A partir dessas métricas foram propostas as métricas DPW-SSIM (*Disparity Perceptual Weighted Video Quality Approach*) e DTPW-SSIM (*Disparity Temporal Perceptual Weighted Video Quality Approach*) para avaliar os vídeos 3D.

Na avaliação de vídeos em duas dimensões a métrica B-SSIM obteve o melhor resultado para imagens borradas. Para vídeos codificados com MPEG-2, e transmitidos por sistemas sem fio e IP, a métrica com o melhor resultado foi a TPW-SSIM. Em relação aos vídeos em três dimensões, as métricas que incluem disparidade superaram as métricas que não a contém, e as com melhores resultados foram a DPW-SSIM e a PW-SSIM.

Palavras-chave: Vídeo ditais; Métricas objetivas; Informação espacial; Qualidade Visual; Informação temporal; Disparidade.

Abstract

Digital videos are subject to several types of distortions, which may occur during the processes of acquisition, processing, compression, storage and transmission of videos, resulting in loss of visual quality. Video quality assessment is important to set system parameters and propose the reduction of degradation, to improve the systems quality.

The most appropriate way to assess the visual quality of a video sequence is subjective assessment. However, objective assessment is quicker and cheaper in comparison with the subjective solutions.

The development of objective metrics has increased significantly, because of the need to quantify appropriately the visual quality perceived by the Human Visual System (HVS). The production of 3D videos is increasing, as well as the need to evaluate them. A problem identified in the evaluation process is related to the methods used for 2D videos, which are not suitable for measuring the quality of 3D videos, since typical depth and distortions of the stereoscopy are not considered.

This study proposes new metrics for assessment of 2D and 3D videos. The proposed metric uses spatial and temporal information and disparity to evaluate videos more closely to the subjective assessment.

In order to obtain the metric for 3D videos, measurements have been proposed for 2D videos, which uses spatial information (PW-SSIM – Perceptual Weighted Video Quality Approach) and temporal information (TPW-SSIM – Temporal Perceptual Weighted Video Quality Approach). Based on these metrics, the DPW–SSIM (Disparity Perceptual Weighted Video Quality Approach) and the DTPW-SSIM metrics (Disparity Temporal Perceptual Weighted Video Quality Approach metrics) have been proposed to assess 3D videos.

For the evaluation of two dimensional videos the B-SSIM metric obtained the best result for blurry images. For MPEG-2 encoded videos, transmitted by wireless and IP systems, the metric with the best result was the TPW-SSIM. Regarding the three dimensional videos, the metrics that included disparity exceeded the metrics that did not include it, and the best results were obtained for the DPW-SSIM and PW-SSIM metrics.

Keywords: Digital videos; Objective metrics; Spatial information; Visual quality; Temporal information; Disparity.

Lista de Símbolos

a	Altura
b	Número de *bits*
V	Número de blocos
B	Distância do objeto
d	Disparidade
F	Distância focal
$f(x, y, n)$	Representação de um vídeo 2D
$f_l(x, y, n)$	Representação da vista esquerda de um vídeo 3D
$f_r(x, y, n)$	Representação da vista direita de um vídeo 3D
$h(x, y, n)$	Representação de um vídeo degradado em 2D
H	Tamanho real da imagem
n	Número de quadros
N	Número de quadros
P	Número de *Pixels*
R	Tamanho da imagem da retina
r	Raio de abertura dos olhos
S	Sobel
S_0	Fóvea
T	Foco ocular
t_c	Distância entre câmeras
v	Distância visual
x	Coordenada do eixo horizontal
X	Total de colunas em um quadro
w	Máscara
W	Largura da imagem
y	Coordenada do eixo vertical

Y	Total de linhas em um quadro
Z	Profundidade
z_i	Luminância de um *pixel*
μ	Média
σ	Desvio padrão
∇f	Operador gradiente no vídeo f

Lista de Siglas e Abreviaturas

ACR	*Absolute Category Rating*	Ordenamento por Categoria Absoluta
B-SSIM	*Blur – Structural Similarity*	Similaridade Estrutural – Borrado
CC	*Correlation Coefficient*	Coeficiente de Correlação
CIF	*Common Intermediate Format*	Formato Intermediário Comum
CV	*Coefficient of Variation*	Coeficiente de Variação
CW-SSIM	*Complex-Wavelet Structural Similarity*	Similaridade Estrutural Baseada em Wavelet
DCR	*Degradation Category Rating*	Ordenamento por Categoria de Degradação
DCT	*Discrete Cosine Transform*	Transformada Discreta do Cosseno
DIBR	*Depth-Image-Based Rendering*	Processamento da Imagem Baseada na Profundidade
DTPW-SSIM	*Disparity Temporal Perceptual Weighting – Structural Similarity*	Similaridade Estrutural com Ponderação Perceptual, Temporal e da Disparidade
DSCQS	*Double Stimulus Continuous Quality Scale*	Escala de Qualidade de Duplo Estímulo Contínuo
DSIS	*Double Stimulus Impairment Scale*	Escala de Restrição com Duplo Estímulo
ETSI	*European Telecommunications Standard Institute*	Instituto Europeu de Padrões de Telecomunicações
E-SSIM	*Edge – Structural Similarity*	Similaridade Estrutural de Borda
G-SSIM	*Gradient-based Structural Similarity*	Similaridade Estrutural Baseada no Gradiente
HDTV	*High-Definition Television*	Televisão de Alta Definição

HVS	*Human Visual System*	Sistema Visual Humano
Iecom	*Institute for Advanced Studies in Communications*	Instituto de Estudos Avançados em Comunicações
IFPB	*Federal Institute of Education, Science and Technology of Paraíba*	Instituto Federal de Educação, Ciência e Tecnologia da Paraíba
IQA	*Image Quality Assessment*	Avaliação da Qualidade da Imagem
IQSSA	*Image Quality and Stereo Sense Assessment*	Avaliação da Qualidade da Imagem e do Sentido Estéreo
ITU-R	*International Telecommunication Union – Radiocommunication Sector*	União Internacional de Telecomunicações – Setor de Rádio Comunicação
ITU-T	*International Telecommunication Union – Telecommunication Standardization Sector*	União Internacional de Telecomunicações – Setor de Padronização das Telecomunicações
KROCC	*Kendall Rank-Order Correlation Coefficient*	Coeficiente de Correlação de Kendall
MOS	*Mean Opinion Score*	Opinião Média da Avaliação
MOVIE	*Motion-based Video Integrity Evaluation*	Avaliação da Integridade do Vídeo Baseado no Movimento
MSE	*Mean Squared Error*	Erro Médio Quadrático
MS-SSIM	*Multi-Scale Structural Similarity*	Similaridade Estrutural de Múltiplas Escalas
PC	*Pair Comparison*	Comparação em Pares
PLCC	*Pearson Linear Correlation Coefficient*	Coeficiente de Correlação Linear de Pearson
PQM	*Perceived Quality Metric*	Métrica da Qualidade Percebida
PSNR	*Peak Signal-to-Noise Ratio*	Razão Sinal Ruído de Pico
PVD	*Preferred Viewing Distance*	Distância de Exibição Preferencial
PW-SSIM	*Perceptual Weighting – Structural Similarity*	Similaridade Estrutural com Ponderação Perceptual
P-SSIM	*Percentile Pooling Structural Similarity*	Similaridade Estrutural Baseada no Agrupamento Percentual
QCIF	*Quarter Common Intermediate Format*	Um Quarto do Formato Intermediário Comum

QoE	*Quality of Experience*	Qualidade da Experiência
RMSE	*Root Mean Squared Error*	Raiz Quadrada do Erro Médio
SI	*Spatial Perceptual Information*	Informação da Percepção Espacial
SROCC	*Spearman Rank-Order Correlation Coefficient*	Coeficiente de Correlação de Sperman
SSCQE	*Single Stimulus Continuous Quality Evaluation*	Avaliação Contínua da Qualidade por Estímulo Simples
SQVGA	*Sub Quarter Video Graphics Array*	Arranjo Gráfico de um Quarto de Vídeo
SS	*Single Stimulus*	Estímulo Simples
SSA	*Stereo Sense Assessment*	Avaliação no Sentido Estéreo
SIF	*Source Input Format*	Formato de Entrada na Fonte
SSCQE	*Single Stimulus Continuous Quality Evaluation*	Avaliação de Qualidade de Estímulo Simples Contínuo
SSIM	*Structural Similarity*	Similaridade Estrutural
T-DMB	*Terrestrial Digital Multimedia Broadcasting*	Transmissão Digital Multimídia Terrestre
TPW-SSIM	*Temporal Perceptual Weighting – Structural Similarity*	Similaridade Estrutural com Ponderação Perceptual e Temporal
Vídeo 2D		Vídeo em Duas Dimensões
Vídeo 3D		Vídeo em Três Dimensões
ViMSSIM	*Video Quality Metric Structural Similarity Index*	Métrica de Qualidade de Vídeo baseada no Índice de Similaridade Estrutural
VQM	*Video Quality Metric*	Métrica de Qualidade de Vídeo
WSNR	*Weighted Signal to Noise Ratio*	Razão Sinal Ruído Ponderada
3-SSIM	*Three Component Structural Similarity Index*	Índice de Similaridade Estrutural com Três Componentes

Lista de Figuras

Lista de Tabelas

Conteúdo

Capítulo 1

Introdução

Vídeos digitais estão propensos a vários tipos de distorções que podem ocorrer durante os processos de aquisição, processamento, compressão, armazenamento, transmissão ou reprodução, e geralmente resultam em perda da qualidade visual. Em princípio, os usuários de sistemas multimídias anseiam por qualidade máxima. No entanto, há uma relação entre qualidade, disponibilidade, acessibilidade e custo do sistema de vídeo, que resulta em um conceito de melhor qualidade dentro de certas condições.

A qualidade de um de vídeo é um fator subjetivo, e depende não só das distorções presentes, mas também da definição de qualidade por parte de cada indivíduo. Os usuários de serviços de vídeo móvel, por exemplo, aceitam uma queda de qualidade em função da redução de custos (Arthur, 2002).

O modelo ideal para se verificar a qualidade visual de uma sequência de vídeo é a avaliação subjetiva (Estrada, 2009). Entretanto, a realização de testes subjetivos agrega, muitas vezes, alto custo, tempo e uma complexa metodologia. Os experimentos de avaliação da qualidade subjetiva de vídeo são descritos nas recomendações BT.500-10 (ITU-R, 2010) do ITU-R, para serviços de TV, e P.910 da ITU-T para aplicações multimídia (ITU-T, 1999).

Os modelos de avaliação objetiva são mais rápidos e possuem menor custo em relação às soluções subjetivas, além disso, tais modelos indicam a existência de degradações imperceptíveis à visão humana. O conhecimento de limites da percepção de degradações pelo sistema visual humano (HVS) é importante para definir parâmetros em sistemas de compressão, transmissão e armazenamento de vídeos, isto é, para propor um conceito que limita a transmissão a partir da degradação, a fim de melhorar a qualidade proporcionada pelos sistemas.

Nas últimas décadas, o desenvolvimento de métricas objetivas tem crescido significativamente. Isso se deve à necessidade de quantificar, de forma adequada e rápida, a qualidade visual percebida pelo sistema visual humano.

Por suas características de possibilidade de repetição e baixo custo, os métodos objetivos facilitam a avaliação dos serviços e de equipamentos, facilitando a especificação e a avaliação de novos sistemas. Além disso, a implementação de medidas da qualidade de vídeo em tempo real, por meio de métodos objetivos, abre a perspectiva de monitoração contínua da qualidade do serviço fornecido (Arthur, 2002), para sistemas de vigilância (Keval, 2009), o vídeo sob demanda, IPTV, WebTV, os sistemas de transcodificação espacial e de vídeo conferência (Estrada, 2009).

No entanto, as principais métricas objetivas, tais como MSE (*Mean Squared Error*) e PSNR (*Peak Signal to Noise Ratio*), nem sempre tem uma boa correlação com os resultados fornecidos pela avaliação subjetiva. Isso se deve ao fato de a imagem formada no córtex visual primário dos mamíferos não estar representada no domínio do *pixel* e sim no conteúdo e nas relações entre os *pixels* de uma imagem (ou imagens) (Akamine & Farias, 2012).

Atualmente, as métricas objetivas que tem obtido os melhores coeficientes de correlação com testes subjetivos são aquelas com base na abordagem da semelhança estrutural (SSIM – *Structural Similarity*), proposta por Wang *et al.* (2004). Várias métricas surgiram baseadas nessa métrica como, por exemplo, *Edge-Based Structural Similarity* (E-SSIM), *Multi-Scale Structural Similarity* (MS-SSIM), *Fast* MS-SSIM, 3-SSIM, *Percentile Pooling* SSIM (P-SSIM) (Wang *et al.*, 2003), *Complex-Wavelet SSIM index* (CW-SSIM) (Wang & Simoncelli, 2005), *Gradient-based Structural Similarity* (G-SSIM) (Chen *et al.*, 2006b) e ViMSSIM (Vu & Deshpande, 2012). Cada variação do SSIM considera um tipo de efeito ou comportamento e, com isso, há o aumento do custo computacional em relação às métricas mais simples.

Na tentativa de melhorar a abordagem, muitos pesquisadores investigam como introduzir as características do sistema visual humano, a fim de aumentar a sua correlação com os resultados subjetivos. Uma das principais áreas de pesquisa que estão sendo investigadas para obter essa melhoria é a atenção visual do HVS, que envolve a seleção e o foco de estímulos relevantes (Corbetta, 1998).

Experiências mostram que a atenção visual humana não está igualmente distribuída ao longo do espaço de imagem, mas concentra-se em algumas regiões (Itti & Koch, 2001). Estima-se que a inclusão de outros métodos que podem identificar a atenção

visual de uma cena, ou seja, atribuir um peso maior às regiões que têm uma maior importância visual.

O modelo de mapa de saliência é frequentemente utilizado como um indicador da atenção visual. Estudos relatam que os algoritmos de avaliação da qualidade da imagem (IQA – *Image Quality Assessment*), quando combinados com modelos de mapa de saliência, apresentaram melhora significativa em seu desempenho (Wang & Li, 2011), (Farias & Akamine, 2012).

Akamine & Farias (2012) investigaram a modelagem computacional da atenção visual por mapas de saliência que foram incorporados em métricas objetivas (PSNR e SSIM). Esta técnica apresenta bons resultados, principalmente para mapas de saliência gerados a partir de rastreamento ocular, chamada mapa de saliência subjetiva. You *et al.* (2010), também investigaram a atenção visual modelada pelo mapa de saliência, mapa da atenção de saliência e o mapa GAFFE (Rajashekar *et al.*, 2008), como um fator importante para avaliar a qualidade da imagem objetiva. Oprea *et al.* (2009) incluíram elementos que atraem a atenção, como contraste de cor, tamanho, orientação e excentricidade a avaliação da qualidade da imagem.

Outra maneira de considerar as características espaciais como um preditor da atenção visual são os algoritmos de classificação. De um modo geral, os algoritmos de classificação comparam, individualmente, as características espaciais dos *pixels*, que são geralmente fornecidas pelos vetores de gradiente, na mesma posição espacial do vídeo de referência e do vídeo distorcido (Li *et al.*, 2004). Em Li & Bovik (2010) é relatado que algoritmos VQA (*Video Quality Assessment*) são potencializados quando combinados com algoritmos de classificação. O benefício dos algoritmos de classificação em relação ao modelo de mapa de saliência é a baixa complexidade computacional, o que é uma característica essencial, quando combinada com os algoritmos VQA.

Porém, os métodos de avaliação da qualidade de imagem em duas dimensões (2D) não são adequados para medir a qualidade da imagem em três dimensões (3D), uma vez que a profundidade (o fator mais importante) e as distorções típicas da estereoscopia (por exemplo, *crosstalk*) não são incorporadas (Meesters *et al.*, 2004).

Para confirmar que as métricas de avaliação 2D não são adequadas, Hewage *et al.* (2009) realizaram uma análise da correlação entre as métricas de qualidade existentes para vídeo 2D, e os resultados subjetivos da percepção do atributo da cor mais a profundidade estereoscópica. Nesta análise a métrica 2D (*Video Quality Metric* – VQM (Pinson & Wolf, 2004)) foi a melhor avaliada para a componente de cor e for-

temente correlacionada com a percepção geral da profundidade. Mas nenhuma das métricas avaliadas obteve uma correlação forte para os vídeos estereoscópicos.

Por esse motivo, vários estudos subjetivos estão sendo realizados para avaliar diferentes efeitos em vídeos 3D. Um dos primeiros estudos explorou a variação da experiência humana em dois atributos específicos do vídeo 3D: a percepção da presença de profundidade e a naturalidade da profundidade. Foi mostrado que com o aumento da profundidade se obtém um maior sentido de presença, desde que a profundidade seja percebida com naturalidade (Silva *et al.*, 2010).

Em Leon *et al.* (2008) a profundidade foi quantizada em diferentes taxas de *bits*, e avaliada subjetivamente para a percepção 3D em geral, chegando à conclusão que a profundidade pode ser quantizada sem afetar a qualidade da percepção 3D.

Em Tikanmaki *et al.* (2008) foi relatado que as distorções na profundidade do vídeo 3D são menos significativas do que as distorções de cor, e que a percepção da profundidade não se altera com diferentes níveis de quantização da profundidade.

Mesmo com várias pesquisas sendo desenvolvidas em vídeos 3D, poucos métodos objetivos de avaliação de imagens estéreo foram apresentados. A falta de métricas objetivas tem feito com que os pesquisadores nas áreas de codificação e transmissão de vídeos 3D tenham que testar suas melhorias usando a métrica PSNR (Hewage & Martini, 2011), (han Lu *et al.*, 2012) (Zhao & Zhang, 2012) (Maiti *et al.*, 2012) (Alajel & Xiang, 2012).

Entre as métricas já propostas para avaliação objetiva em vídeos 3D, a maioria utiliza os mapas de profundidade em junção com as métricas 2D (Shao *et al.*, 2009). Porém, para a geração dos mapas de profundidade foram usados métodos com alto custo computacional e com uma baixa precisão (Benoit *et al.*, 2008).

Entre as métricas que avaliam vídeos 3D sem o uso de mapas de profundidade pode-se citar a de Yang *et al.* (2009) e a PQM (Joveluro *et al.*, 2010). A métrica de Yang *et al.* (2009) realiza a avaliação objetiva da qualidade da imagem estéreo a partir de dois aspectos: a avaliação da qualidade de imagem (IQA – *Image Quality Assessment*) e avaliação do par estéreo (SSA – *Stereo Sense Assessment*). Portanto, o indicador de avaliação objetiva da imagem estéreo tem duas partes, uma é a qualidade da imagem, e a outra é o sentido estéreo.

1.1 Objetivo da Tese

O objetivo principal desta tese é propor novas métricas de avaliação objetiva para vídeos 3D e 2D, que leve em consideração a estrutura do vídeo, a informação espacial e temporal e a disparidade. Essa nova métrica é comparada com métricas existentes na base de dados NAMA3DS1-COSPAD1 (Urvoy *et al.*, 2012) e LIVE (Seshadrinathan *et al.*, 2010b).

Para o desenvolvimento das novas técnicas foi analisado o desempenho das métricas existentes, considerando a qualidade avaliada e o tempo de processamento (Cardoso *et al.*, 2012). A partir dessa análise, são sugeridas modificações em métricas objetivas existentes para avaliar os vídeos com uma maior correlação com as avaliações subjetivas. Essas modificações são baseadas no HVS, uma vez que os usuários têm mais interesse em determinadas áreas.

A partir da análise das métricas existentes acrescenta-se a informação espacial e temporal a uma delas e depois avalia-se a nova métrica para vídeos 2D. Depois acrescenta-se a essa métrica o efeito da disparidade para avaliar os vídeos 3D. A escolha da disparidade é devido a essa ser uma forma rápida para se obter uma ideia da profundidade dos objetos na imagem.

1.2 Organização do Texto

Esta tese está organizada em sete capítulos que apresentam aspectos teóricos sobre tópicos relacionados aos assuntos da pesquisa, bem como os resultados obtidos neste trabalho.

No Capítulo 2 são apresentados conceitos da visão e da formação da imagem no olho. As características do sistema visual humano e da atenção visual são apresentadas, uma vez que as métricas de avaliação de vídeo devem ser geradas a partir dessas características. Também é apresentado como pode ser realizada a segmentação das imagens, que é uma técnica muito utilizada em imagens e vídeos. Para finalizar o capítulo, são apresentadas as principais degradações encontradas em vídeos.

O Capítulo 3 aborda as definições sobre vídeo em três dimensões. Os fatores de profundidade são apresentados nesse capítulo, como a profundidade monocular e binocular. Dentre elas, a técnica que se destaca é a estereoscopia. Um outro ponto abordado nesse capítulo são as formas de captura dos vídeos em três dimensões.

Também são apresentados os fatores que diminuem o conforto visual dos vídeos em três dimensões e as medidas para avaliar o desconforto visual, que ainda estão em fase de padronização.

No Capítulo 4 são apresentadas e analisadas as métricas de avaliação objetiva e subjetiva. São apresentadas algumas métricas de avaliação objetiva e subjetiva que foram usadas durante a tese, assim como algumas métricas estatísticas que podem avaliar os resultados das métricas objetivas e subjetivas.

No Capítulo 5 são descritos as métricas propostas neste trabalho, considerando que existem partes do vídeo que são mais importantes que outras. As avaliações das métricas propostas, as avaliações subjetivas e a comparação entre elas são apresentadas no Capítulo 6.

Por fim, no Capítulo 7 é descrito um resumo das principais contribuições a partir dos resultados obtidos na tese e discute alguns trabalhos futuros que se pretende realizar após o término da tese. Uma lista de publicações geradas até o presente é apresentada.

Capítulo 2

Vídeos Digitais

O vídeo digital consiste na reprodução de um conjunto de imagens (quadros) em intervalos de tempo pré-determinados (Wu & Rao, 2006). Matematicamente, o nível de luminância de um vídeo digital pode ser considerado uma função $f(x, y, n)$, em que x, y e n são discretos, e x e y representam as coordenadas espaciais de um *pixel* e n representa o quadro. Um modelo estrutural para essa concepção é ilustrado na Figura 2.1.

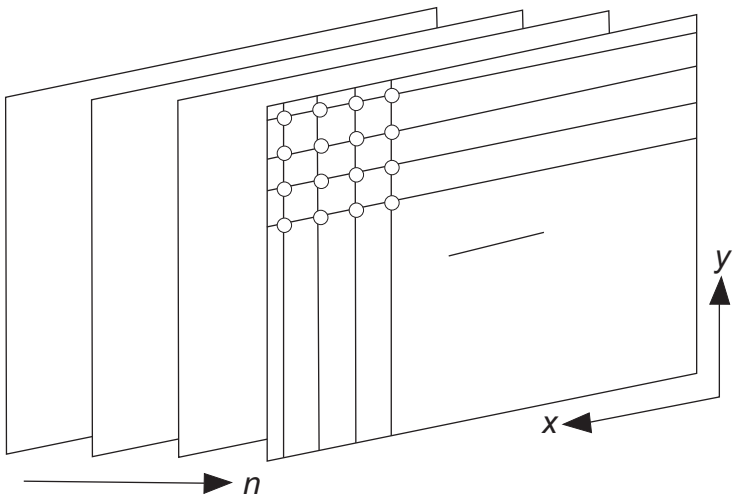


Figura 2.1: Modelo espaço-temporal de uma sequência de vídeo. Adaptada de Richardson (2010).

Os vídeos digitais são submetidos a vários tipos de processos, entre eles: codificação, armazenamento e transmissão. Esses processos muitas vezes implicam perda da

qualidade do vídeo. Assim, torna-se muito importante avaliar a qualidade dos vídeos processados, além de indicar o quanto determinado processo degradou o vídeo.

Nas próximas seções são apresentadas características da visão e da formação da imagem e algumas técnicas de processamento de vídeos digitais que estão, de alguma forma, relacionadas às métricas de avaliação objetiva da qualidade de vídeos.

2.1 Visão

A visão é um sentido essencial. Estima-se que 80 a 90% de todos os neurônios do cérebro humano são envolvidos na percepção visual (Young, 1991). O órgão responsável pela captação da informação luminosa (visual) e transformá-la em impulsos a serem decodificados pelo sistema nervoso é o olho. Ele é um órgão especializado e delicadamente coordenado, e cada uma de suas estruturas desempenha um papel específico na transformação da luz. Os olhos funcionam como um sistema de conversão seletiva do estímulo luminoso em sinais bioelétricos.

O olho humano é formado por um conjunto complexo de elementos que atuam de forma específica para que o ato de olhar, ver ou enxergar ocorra. Primeiramente existem aquelas estruturas responsáveis pela captação da luz e que desempenham função óptica, posteriormente estão os elementos que transformam o impulso luminoso em impulso elétrico, por meio de reações químicas. O sistema óptico do olho humano (Figura 2.2) é composto pela córnea, íris, pupila, cristalino, retina, esclera e nervo ótico (Ramos, 2006; Pedrini & Schwartz, 2007).

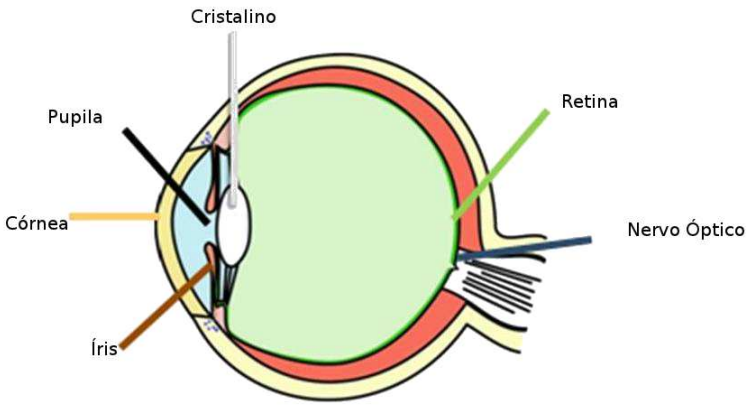


Figura 2.2: Sistema visual humano.

- Córnea – A córnea é um tecido resistente e transparente que cobre a superfície anterior do olho. É a primeira estrutura do olho que a luz atinge. A córnea tem cinco camadas de tecido transparente e resistente. A camada mais externa, o epitélio, possui uma capacidade regenerativa muito grande e se recupera rapidamente de lesões superficiais. As quatro camadas seguintes, mais internas, proporcionam rigidez e protegem o olho de infecções.

- Íris – A porção visível e colorida do olho, logo atrás da córnea. Possui músculos para aumentar ou diminuir a pupila, a fim de que o olho possa receber mais ou menos luz, conforme as condições de luminosidade do ambiente.

- Pupila – É um orgão responsável pela adaptação à luz, corresponde à abertura central da íris, pela qual a luz passa para alcançar o cristalino. O diâmetro da abertura pupilar pode variar entre 1,5 e 8 mm, que corresponde a uma mudança de 30 vezes a quantidade de luz que entra no olho.

- Cristalino – A importância do cristalino é a sua curvatura, assim, a sua capacidade óptica pode ser voluntariamente aumentada pela contração dos músculos ligados a ele. Essa capacidade de aumentar ou diminuir sua superfície curva, a fim de se ajustar às diferentes necessidades de focalização das imagens, próximas ou distantes, é chamada acomodação (Seção 3.1.1). A acomodação é essencial para trazer objetos de distâncias diferentes ao foco. Entretanto, essa habilidade é diminuída gradualmente com o aumento da idade, até ser perdida quase completamente, uma condição conhecida como presbiopia (Winkler, 2005).

- Retina – É a membrana mais interna do olho revestida por uma camada de tecidos nervosos. Sendo responsável pela sensação da imagem visual projetada pelas estruturas da parte frontal do olho, pela codificação das informações em sinais nervosos e transmissão destas para o cérebro.

  A retina contém fotorreceptores que transformam a luz em impulsos bioelétricos, que o cérebro pode interpretar como imagens. Os fotorreceptores são neurônios especializados que fazem uso de substâncias fotoquímicas sensíveis à luz, para converter a energia da luz incidente em sinais que podem ser interpretados pelo cérebro. Há dois tipos de fotoreceptores na retina, os cones e os bastonetes. Os cones são sensíveis à cor e responsáveis pela capacidade do olho em discernir detalhes nas imagens. Os bastonetes são responsáveis pela visão monocromática,

e têm sensibilidade à intensidade mais apurada que os cones. Isto faz com que, com pouca iluminação, se consiga ter percepção geral da imagem captada no campo de visão, porém em tons de cinza.

Cada olho recebe e envia ao cérebro uma imagem, no entanto, os objetos são vistos como um só, devido à capacidade de fusão das imagens. A visão binocular (com os dois olhos, Seção 3.1.1) proporciona maior campo visual e a noção de profundidade. O ponto cego é uma pequena região da retina em que está localizado o nervo óptico e não possui fotorreceptores. Por outro lado a fóvea é parte central da retina, na qual a visão é mais desenvolvida, pelo fato de possuir maior concentração de cones.

- Nervo Óptico – Transporta os impulsos elétricos do olho para o centro de processamento do cérebro, para a devida interpretação.

- Esclera – É o nome da capa externa, fibrosa, branca e rígida que envolve o olho, e contínua com a córnea. É a estrutura que dá forma ao globo ocular.

2.2 Formação da Imagem no Olho

Em câmeras fotográficas, a lente tem uma distância focal fixa, e a focalização para diferentes distâncias é obtida variando a distância entre a lente e o plano-imagem. No olho humano, a distância entre a lente e o plano-imagem é fixa, e a distância focal necessária para atingir uma focalização adequada é obtida variando o formato do cristalino (que equivale a uma lente flexível) (Gonzalez & Woods, 2006).

A disposição geométrica apresentada na Figura 2.3 ilustra como calcular as dimensões de uma imagem formada na retina. Por exemplo, suponha que uma pessoa esteja olhando para uma objeto de 15 m de altura a uma distância de 100 m e o diâmetro do seu olho seja de 17 mm. Se a for a altura do objeto na imagem formada na retina, a disposição geométrica da figura leva a $15/100 = a/17$ ou $a = 2{,}55$ mm. A imagem na retina é focalizada principalmente na área da fóvea, a percepção ocorre pela excitação relativa dos receptores de luz, que transformam a energia radiante em impulsos elétricos, posteriormente decodificados pelo cérebro.

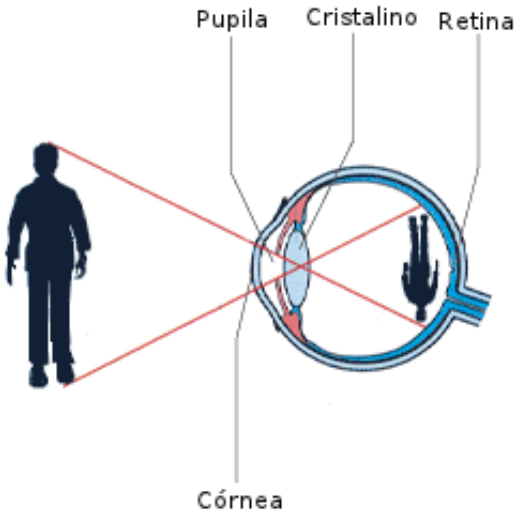


Figura 2.3: Representação gráfica do olho focalizando um objeto.

2.2.1 Determinação da distância de um objeto em relação ao olho

Uma pessoa normalmente percebe a distância por três meios principais: primeiro, com as dimensões das imagens de objetos conhecidos na retina, segundo, a partir do fenômeno da paralaxe de movimento e, terceiro, com a estereoscopia. A capacidade de determinar a distância é chamada de percepção de profundidade (Guyton & Hall, 2006).

Determinação da distância pelas dimensões de imagens retinianas de objetos conhecidos

Se uma pessoa tem 1,80 m de altura, pode-se determinar quanto ela está distante simplesmente pelo tamanho de sua imagem na retina. Não é preciso conscientemente pensar no tamanho, mas o cérebro aprendeu a calcular automaticamente, a partir da dimensão das imagens, as distâncias dos objetos quando as dimensões são conhecidas.

Determinação da distância por paralaxe de movimento

Outro meio importante pelo qual os olhos determinam a distância é a paralaxe de movimento. Se um indivíduo olhar à distância com os olhos completamente imóveis, não perceberá paralaxe de movimento, mas quando ela movimenta a cabeça para um

lado ou outro, as imagens dos objetos próximos se movimentam nas retinas, enquanto as imagens dos objetos distantes continuam quase completamente estáticas.

Determinação de distância por estereoscopia

Outro método pelo qual se percebe a paralaxe é a "visão binocular". Como um olho está a em média 6,3 cm do outro olho, as imagens nas duas retinas são diferentes entre si. Por exemplo, um objeto de 2,5 cm em frente ao nariz forma uma imagem no lado esquerdo da retina do olho esquerdo e no lado direito da retina do olho direito, enquanto um pequeno objeto posicionado 6 m à frente do nariz tem sua imagem em pontos correspondentes nos centros das duas retinas. É quase inteiramente essa paralaxe binocular (ou estereoscopia) que dá a uma pessoa com dois olhos uma capacidade muito maior para julgar distâncias relativas quando os objetos estão próximos do que uma pessoa que tenha apenas um olho.

2.3 Sistema Visual Humano

A identificação e a compreensão do funcionamento das principais características do sistema visual humano, que estão relacionadas com fatores de percepção de qualidade, são importantes para o desenvolvimento de métricas objetivas de avaliação da qualidade de vídeos.

Os mecanismos e os conceitos associados aos sinais de vídeo se baseiam no processo de percepção de imagens pelo ser humano. O sistema de visão recebe estímulos luminosos e transfere as informações ao cérebro, que as processa criando a percepção de imagens. Esse processo é dinâmico com dependências temporal e espacial, pois a cada instante e em diferentes posições dentro do campo visual os estímulos e a percepção se renovam.

A visão envolve diversas funções complexas, tais como detecção, localização, reconhecimento e interpretação de objetos no ambiente (Pedrini & Schwartz, 2007). Por esse motivo, a compreensão do sistema visual humano pode auxiliar no desenvolvimento de sistemas capazes de adquirir, analisar e interpretar informações visuais.

Por último, a capacidade de percepção de detalhes pelo olho humano é finita, limitada pela estrutura dos elementos fotossensíveis da retina. Isto se denomina acuidade visual, que é medida em graus (ou minutos de arco).

A seguir são apresentadas algumas características psicofísicas do HVS que estão diretamente relacionadas com a qualidade visual percebida.

- *Visão Central e Periférica*

  A visão central é o efeito causado quando um observador humano fixa o olhar em um ponto no seu ambiente visual. A região em torno desse ponto se apresenta com uma maior resolução espacial em relação a outras regiões.

  A visão periférica é a propriedade da visão de perceber o que está fora do foco principal de visão. À medida que um objeto está mais distante do ponto de fixação do olhar humano, sua resolução espacial é diminuída (Estrada, 2009).

  Dessa forma, devido às características de visão central e periférica, há uma segregação da imagem, por parte do HVS, em regiões potencialmente mais relevantes e menos relevantes. Além disso, insere-se o fator subjetivo de que degradações em regiões visualmente mais relevantes são mais sensíveis ao HVS do que degradações em zonas de fundo ou periféricas, consideradas potencialmente menos relevantes.

- *Adaptação à Luz e Contraste Simultâneo*

  O HVS consegue discriminar uma grande variação na intensidade luminosa, capacidade conhecida como adaptação à luz (Gonzalez & Woods, 2006). Agregado a essa característica, existe um fenômeno denominado contraste simultâneo, que está relacionado ao fato de que a luminosidade de uma região percebida pelo olho humano não depende simplesmente da intensidade dos *pixels* dessa região. Ou seja, o olho humano codifica o contraste relativo ao estímulo visual, e não a intensidade absoluta da luz (Estrada, 2009).

  O efeito de contraste simultâneo pode ser observado na Figura 2.4. Os retângulos centrais possuem o mesmo nível de cinza, por sua vez, os planos de fundo possuem níveis de cinza diferentes, provocando uma falsa percepção de que o retângulo que está sobre o fundo mais claro é mais escuro do que o outro que está sobre o fundo mais escuro.

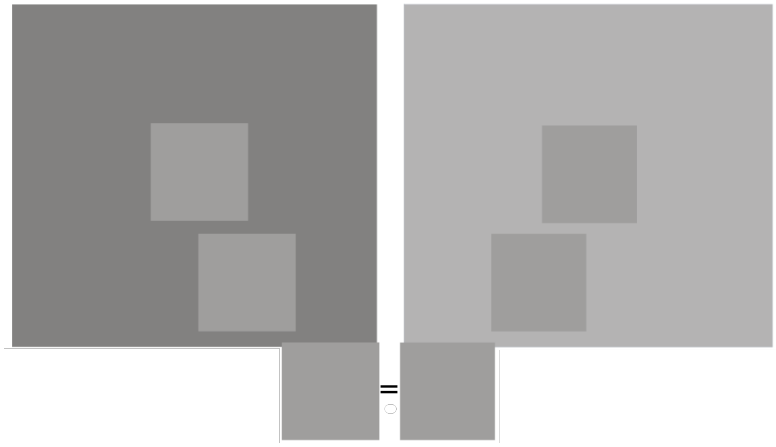


Figura 2.4: Efeito de contraste simultâneo.

2.4 Atenção Visual

Os olhos humanos recebem uma grande quantidade de estímulos visuais. No entanto, é impossível processar toda a informação que chega aos olhos de uma só vez (Tsotsos, 1990). Assim, o cérebro humano pode resolver esse problema de diferentes formas. Uma forma é movimentar rapidamente os olhos, esse movimento é chamado de sacádicos (do inglês *saccadic eye movements*). Os movimentos sacádicos servem para criar um diferencial entre as imagens, para manter a atenção, visto que o olho é mais sensível à diferença de luminosidade que à luminosidade absoluta.

Os mecanismos óculo-motores permitem que o olhar se fixe em um determinado ponto (fixação) ou mudem para outro local quando informações suficientes já foram coletadas. A seleção das fixações baseia-se nas propriedades visuais da cena e é dada prioridade às áreas com uma concentração elevada de informações, minimizando a quantidade de dados a serem processados pelo cérebro e maximizando a qualidade da informação recolhida (Akamine & Farias, 2012).

A habilidade que o sistema visual tem para selecionar e processar somente as regiões mais relevantes em uma cena visual é chamada de atenção visual. Ela pode ser entendida como um mecanismo para lidar com a incapacidade de tratar de uma só vez uma grande quantidade de informação visual, tanto em sistemas biológicos quanto em sistemas computacionais (Pereira, 2007).

Os métodos principais para obtenção da atenção visual podem ser identificados por dois métodos, o *top-down* e o *bottom-up*. O *top-down* usa conhecimentos obtidos a *priori* para detectar regiões de maior interesse na imagem. Esses conhecimentos podem ser

obtidos de várias formas. Geralmente, utiliza-se ferramentas de aprendizagem baseadas em modelos geométricos ou relacionais (como redes semânticas), ou modelos estatísticos (como redes neurais). Porém, esses conhecimentos também podem ser fornecidos por um ser humano, selecionando manualmente regiões de maior interesse da imagem.

O método de atenção visual *bottom-up* é guiado por características primitivas da imagem, como a cor, a intensidade e a orientação. Algumas características como cor, orientação ou tamanho dos objetos em uma imagem são responsáveis por guiar o mecanismo biológico de atenção visual (Wolfe & Horowitz, 2004).

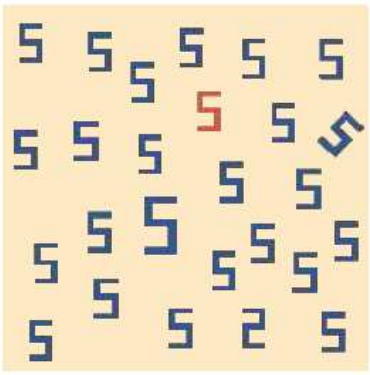
Em Itti & Koch (2001) foi proposto um modelo que analisa as imagens procurando as propriedades visuais mais relevantes: cor, intensidade e orientação. O algoritmo cria mapas correspondentes a estas três propriedades e as combina para prever a probabilidade de fixação do olhar em áreas específicas da imagem.

A junção de todos esses pontos forma o mapa de saliência, que é um mapa escalar e bidimensional, cuja atividade representa topograficamente a saliência visual, independentemente da dimensão característica que torna o local saliente.

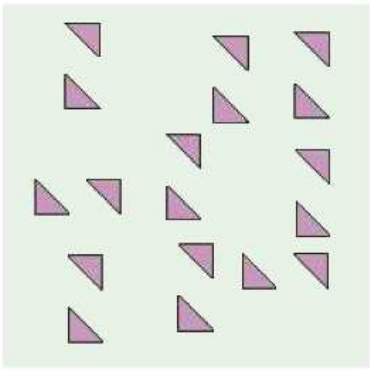
Assim, se o elemento do mapa de saliência for ativo, esta localização é saliente, não importando se ele corresponda a um objeto vermelho em um campo de objetos verde, ou a um estímulo em movimento para a direita, enquanto outros se movem para a esquerda.

Na Figura 2.5, há exemplos de tarefas de busca visual. Algumas destas tarefas são simples. Na Figura 2.5a, o contraste entre o azul e o vermelho ressalta a existência de um numeral 5 (cinco) de cor diferente dos demais. No entanto, perceber um número cinco azul e maior, é um pouco mais complicado. A Figura 2.5a também é um exemplo da importância de conhecimento *a priori* para executar determinadas buscas visuais, pois dificilmente é possível identificar o número dois existente nesta imagem sem que alguém tenha dito que há um número dois. Isto demonstra o fato de que a atenção visual *top-down* é mais lenta e necessita de conhecimento prévio sobre o que se quer encontrar.

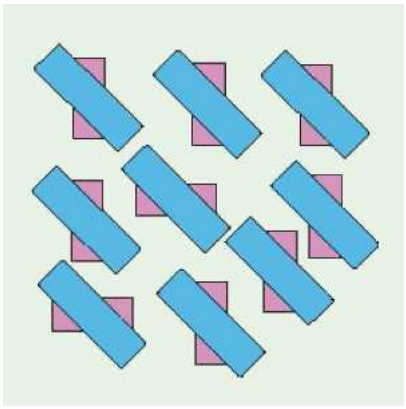
As Figuras 2.5b e 2.5c demonstram a importância da orientação e do contraste de cores para ressaltar objetos diferentes em imagens. Na Figura 2.5c é difícil encontrar os pares de triângulos horizontais, mas esta tarefa é simplificada devido ao contraste de cores entre os retângulos azuis e os retângulos rosas. Na Figura 2.5d, a busca por cruzes é ineficiente devido ao fato da informação de intersecção não guia a atenção.



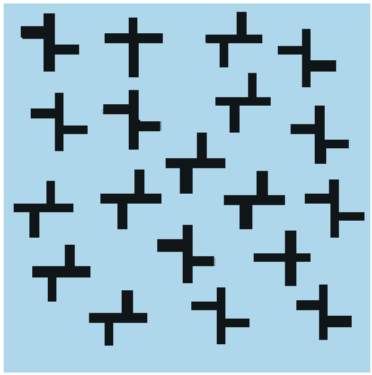
(a) Conhecimento *a priori* para executar buscas visuais.



(b) Contraste de cores.



(c) Contraste de orientações.



(d) Informação de intersecção não guia a atenção.

Figura 2.5: Exemplos de tarefas de busca visual.

2.5 Segmentação de Imagens

A segmentação é um dos primeiros passos para a análise de imagens. A segmentação subdivide uma imagem em suas partes ou objetos constituintes, que devem corresponder às áreas de interesse da aplicação (Gonzalez & Woods, 2006). O processo de segmentação deve localizar corretamente a posição, a topologia e a forma dos objetos para que as informações resultantes da análise da imagem sejam partes dessa imagem.

Os algoritmos de segmentação são geralmente baseados na busca por descontinuidades ou pelas similaridades dos níveis de cinza. Na primeira categoria, os métodos visam particionar a imagem com base em mudanças abruptas nos níveis de cinza, que são caracterizadas pela detecção de pontos isolados, linhas ou bordas na imagem. Para

$$w = \begin{bmatrix} w_1 & w_2 & w_3 \\ w_4 & w_5 & w_6 \\ w_7 & w_8 & w_9 \end{bmatrix}$$

Figura 2.6: Máscara de 3×3 *pixels*.

a segunda categoria, os métodos baseiam-se em agrupar pontos da imagem que apresentam valores similares para um determinado conjunto de características (Gonzalez & Woods, 2006; Pedrini & Schwartz, 2007).

O conceito de segmentação de uma imagem, baseado em descontinuidade ou em similaridade dos valores de níveis de cinza de seus *pixels*, pode ser aplicado tanto a imagens estáticas como a imagens dinâmicas.

2.5.1 Detecção de Descontinuidades

Existem três tipos básicos de descontinuidades detectadas em imagens digitais: pontos, linhas e bordas. A maneira mais rápida para a identificação de descontinuidades é com a varredura da imagem por uma máscara. No caso de uma máscara w com tamanho 3×3 *pixels*, Figura 2.6, esse procedimento envolve o cálculo da soma dos produtos dos coeficientes pelos níveis de cinza da região delimitada pela máscara. Dessa forma, a resposta E da máscara posicionada sobre um ponto da imagem é dada por

$$E = w_1 z_1 + w_2 z_2 + \ldots + w_9 z_9 = \sum_{i=1}^{9} w_i z_i, \tag{2.1}$$

na qual z_i é o nível de luminância associado ao coeficiente w_i da máscara. A resposta da máscara é definida em relação à sua posição central. Quando a máscara é posicionada em um *pixel* da borda, a resposta é calculada, utilizando a vizinhança parcial apropriada.

2.5.2 Detecção de Pontos

A detecção de pontos isolados pode ser obtida pela aplicação direta da máscara h, definida como

$$h = \begin{bmatrix} -1 & -1 & -1 \\ -1 & 8 & -1 \\ -1 & -1 & -1 \end{bmatrix}$$

Um ponto é detectado na posição central da máscara se

$$|R| > T, \tag{2.2}$$

em que T é um limiar não negativo e o R é dado pela Fórmula 2.1. Calculando as diferenças ponderadas entre os valores do ponto central e de seus vizinhos, um ponto é detectado se houver uma discrepância entre o seu valor de nível de cinza e de seus vizinhos (Pedrini & Schwartz, 2007).

2.5.3 Detecção de Linhas

A detecção de linhas pode ser realizada com o uso de máscaras. As máscaras da Equação 2.3 indicam como podem ser detectadas as linhas na horizontal (h_1), linhas orientadas a 45 graus (h_2), linhas verticais (h_3) e linhas orientadas a 135 graus (h_4).

$$h_1 = \begin{bmatrix} -1 & -1 & -1 \\ 2 & 2 & 2 \\ -1 & -1 & -1 \end{bmatrix}, \qquad h_2 = \begin{bmatrix} -1 & -1 & 2 \\ -1 & 2 & -1 \\ 2 & -1 & -1 \end{bmatrix},$$
$$h_3 = \begin{bmatrix} -1 & 2 & -1 \\ -1 & 2 & -1 \\ -1 & 2 & -1 \end{bmatrix}, \qquad h_4 = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix}. \tag{2.3}$$

Para uma imagem com intensidade de fundo constante, a resposta máxima da máscara h_1 ocorre quando a linha passa pelo meio da máscara. Isso pode ser verificado rascunhando uma matriz simples de elementos "1" com uma linha composta por elementos com um nível de cinza diferente posicionada horizontalmente na matriz (Gonzalez & Woods, 2006).

2.5.4 Detecção de Bordas

Uma borda é o limite ou a fronteira entre duas regiões com valores de luminância relativamente distintos. Assume-se que as regiões em questão são suficientemente homogêneas, de maneira que a transição entre duas regiões pode ser determinada com base apenas na descontinuidade dos níveis de cinza.

A maioria das técnicas para detecção de bordas utiliza o cálculo de um operador local diferencial. Esse operador é utilizado porque em imagens reais as descontinuidades abruptas não são comuns, devido às componentes de baixa frequência ou da suavização produzida pela maior parte dos dispositivos. Bordas em imagens digitais são, em geral, levemente borradas devido à amostragem (Pedrini & Schwartz, 2007).

A utilização do operador local diferencial pode ser realizada usando a primeira e a segunda derivada. A primeira derivada em imagens é positiva nas transições da região escura para a região clara, negativa nas transições da região clara para escura e nula nas áreas de nível de cinza constante. Por outro lado, a segunda derivada é positiva na parte da transição associada ao lado mais escuro da borda, negativa na parte da transição associada ao lado mais claro da borda e nula nas áreas de nível de cinza constante.

Portanto, a magnitude da primeira derivada pode ser utilizada na detecção de uma borda em uma imagem, enquanto a segunda derivada gera um cruzamento em zero, ou seja, uma indicação de que há uma mudança de sinal na transição dos níveis de cinza, permitindo a localização das bordas em uma imagem.

A análise está limitada a um perfil horizontal unidimensional, mas a abordagem pode ser aplicada a uma borda de qualquer orientação da imagem. A primeira derivada em qualquer ponto da imagem é obtida usando a magnitude do gradiente naquele ponto e a segunda derivada é obtida similarmente utilizando o laplaciano.

Operadores de Gradiente

Como uma imagem depende de duas coordenadas espaciais, as bordas da imagem podem ser expressas por derivadas parciais. Um operador comumente utilizado em diferenciação de imagens é o gradiente, que é um vetor cuja direção indica os locais nos quais os níveis de cinza sofrem maior variação.

O vetor gradiente, na forma matricial, de uma imagem pode ser expresso como

$$\begin{bmatrix} G_x \\ G_y \end{bmatrix} = \begin{bmatrix} \dfrac{\partial f}{\partial x} \\ \\ \dfrac{\partial f}{\partial y} \end{bmatrix}. \tag{2.4}$$

A partir da análise vetorial, observa-se que o vetor gradiente aponta na direção da mudança mais rápida de f na posição (x, y). Em detecção de bordas, a magnitude desse vetor é uma quantidade importante, geralmente chamada simplesmente de gradiente e denotada por ∇f, em que G_x e G_y são os gradientes em x e y, respectivamente.

$$\nabla \mathrm{f} = \sqrt{G_x^2 + G_y^2} = \sqrt{\left(\frac{\partial f}{\partial x}\right)^2 + \left(\frac{\partial f}{\partial y}\right)^2} \tag{2.5}$$

Essa quantidade equivale à maior taxa de aumento de $f(x, y)$ por unidade de distância na direção de ∇f. Comumente, aproxima-se o gradiente com valores absolutos

$$\nabla \mathrm{f} \approx |G_x| + |G_y|. \tag{2.6}$$

A direção do vetor gradiente é também uma medida importante. Seja $\theta(x, y)$ o ângulo de direção do vetor ∇f na posição (x, y). Então, a partir da análise vetorial, tem-se

$$\theta(x, y) = \arctan\left(\frac{G_y}{G_x}\right) \tag{2.7}$$

em que o ângulo é medido em relação ao eixo x.

Uma mudança em intensidade pode ser detectada pela diferença entre os valores de *pixels* adjacentes. Bordas verticais podem ser detectadas pela diferença horizontal entre pontos, enquanto bordas horizontais podem ser detectadas pela diferença vertical entre os pontos adjacentes da imagem.

Seja a região da imagem mostrada na Figura 2.7, em que os valores denotam os níveis de cinza dos *pixels*. A magnitude do gradiente (Fórmula 2.5), pode ser aproximada

no ponto $f(x, y)$ de várias maneiras. Uma forma simples consiste em usar a diferença $f(x, y) - f(x + 1, y)$ na direção de x e $f(x, y) - f(x, y + 1)$ na direção y, combinadas como

$$\nabla f \approx \sqrt{[f(x, y) - f(x + 1, y)]^2 + [f(x, y) - f(x, y + 1)]^2}. \tag{2.8}$$

$f(x-1, y-1)$	$f(x, y-1)$	$f(x+1, y-1)$
$f(x-1, y)$	$f(x, y)$	$f(x+1, y)$
$f(x-1, y+1)$	$f(x, y+1)$	$f(x+1, y+1)$

Figura 2.7: Região da imagem formada por 3×3 *pixels*.

Uma outra abordagem para a aproximação da Fórmula 2.5 é usar as diferenças cruzadas

$$\nabla f \approx \sqrt{[f(x, y) - f(x + 1, y + 1)]^2 + [f(x, y + 1) - f(x + 1, y)]^2}, \tag{2.9}$$

ou usar os valores absolutos

$$\nabla f \approx |f(x, y) - f(x + 1, y + 1)| + |f(x, y + 1) - f(x + 1, y)|. \tag{2.10}$$

Essa equação pode ser implementada com máscaras de tamanho 2×2 *pixels*. Por exemplo, a Fórmula 2.9 pode ser implementada tomando o valor absoluto das respostas das duas máscaras apresentadas na Fórmula 2.11 e somando os resultados. Essas máscaras são chamadas de operadores cruzados de gradiente de Roberts (Gonzalez & Woods, 2006).

$$G_x = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}, \quad G_y = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}. \tag{2.11}$$

Uma aproximação para a Fórmula 2.5 no ponto $f(x, y)$, mas usando uma vizinhança de 3×3 *pixels*, é dada por

$$\begin{aligned}\nabla f \approx |&[f(x+1,y-1)+f(x+1,y)+f(x+1,y+1)]-\\&[f(x-1,y-1)+f(x-1,y)+f(x-1,y+1)]|+\\&|[f(x-1,y+1)+f(x,y+1)+f(x+1,y+1)]-\\&[f(x-1,y-1)+f(x,y-1)+f(x+1,y-1)]|.\end{aligned} \tag{2.12}$$

A diferença entre a terceira e a primeira coluna da região 3×3 *pixels* aproxima a derivada na direção x, enquanto a diferença entre a terceira e a primeira linhas aproxima a derivada na direção y. As máscaras mostradas na Fórmula 2.13, chamadas de operadores de gradiente de Prewitt, podem ser usadas para implementar a Fórmula 2.12 (Pedrini & Schwartz, 2007).

$$G_x = \begin{bmatrix}-1 & 0 & 1\\ -1 & 0 & 1\\ -1 & 0 & 1\end{bmatrix}, \qquad G_y = \begin{bmatrix}-1 & -1 & -1\\ 0 & 0 & 0\\ 1 & 1 & 1\end{bmatrix}. \tag{2.13}$$

O operador de Sobel aproxima a magnitude do gradiente com a diferença de valores ponderados dos níveis de cinza da imagem, como

$$\begin{aligned}G_x \approx &[f(x+1,y-1)+2f(x+1,y)+f(x+1,y+1)]-\\&[f(x-1,y-1)+2f(x-1,y)+f(x-1,y+1)]\\G_y\ &[f(x-1,y+1)+2f(x,y+1)+f(x+1,y+1)]-\\&[f(x-1,y-1)+2f(x,y-1)+f(x+1,y-1)].\end{aligned} \tag{2.14}$$

As máscaras mostradas na Fórmula 2.15 implementam o operador de Sobel

$$G_x = \begin{bmatrix}-1 & 0 & 1\\ 2 & 0 & 2\\ -1 & 0 & 1\end{bmatrix}, \qquad G_y = \begin{bmatrix}-1 & -2 & -1\\ 0 & 0 & 0\\ 1 & 2 & 1\end{bmatrix} \tag{2.15}$$

em que, os níveis de cinza dos *pixels* de uma região são sobrepostos pelas máscaras

centradas no *pixel* (x, y) da imagem. A Figura 2.8 ilustra a detecção de bordas de uma imagem utilizando o operador de Sobel. A detecção é obtida pela combinação dos resultados obtidos pelo operador de Sobel, G_x e G_y (Chen *et al.*, 2006a).

(a) Quadro original (b) Quadro após a convolução

Figura 2.8: Exemplo da aplicação do operador de Sobel. Na Figura 2.8a é apresento o quadro 1 do vídeo Foreman. Na Figura 2.8b é apresentado o mesmo quadro do vídeo após a convolução com o operador de Sobel.

Outros operadores podem ser também utilizados, como o Kirsch, Robinson e o Frei-Chen. O operador de Kirsch e Robinson tem oito máscaras de convolução e o operador de Frei-Chen consiste de nove máscaras de convolução, o que gera uma maior complexidade ao sistema (Pedrini & Schwartz, 2007).

Laplaciano

O operador laplaciano de uma função bidimensional contínua $f(x, y)$ é definido por uma derivada de segunda ordem, como

$$\nabla^2 f = \frac{\partial^2 f}{\partial x} + \frac{\partial^2 f}{\partial y}. \tag{2.16}$$

Como no caso do gradiente, a Equação 2.16 pode ser implementada na forma digital de diferentes maneiras. Para o caso de uma região 3×3, a forma mais frequentemente encontrada na prática é

$$\nabla^2 f = 4f(x, y) - f(x, y-1) + f(x-1, y) + f(x+1, y) + f(x, y+1). \tag{2.17}$$

A exigência para a definição do operador laplaciano na forma discreta é que o

coeficiente associado ao *pixel* central seja positivo e que os outros *pixels* externos sejam negativos. Uma vez que o laplaciano é uma derivada, a soma dos coeficientes tem que ser nula. Portanto, a resposta é nula sempre que o ponto em questão e seus vizinhos tiverem o mesmo valor, ou seja, pertencerem a uma região homogênea da imagem.

A máscara apresentada na Equação 2.18 pode ser usada na implementação da Fórmula 2.17, tal que, as duas matrizes que compõem a máscara correspondem às derivadas segundas ao longo de todas as linhas e colunas, respectivamente, assim como no laplaciano contínuo expresso na Equação 2.16.

$$\begin{bmatrix} 0 & 0 & 0 \\ -1 & 2 & -1 \\ 0 & 0 & 0 \end{bmatrix} + \begin{bmatrix} 0 & -1 & 0 \\ 0 & 2 & 0 \\ 0 & -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{2.18}$$

2.6 Artefatos

As técnicas de compressão buscam eliminar as redundâncias presentes na sequência de imagens, entretanto, estão diretamente relacionadas com o surgimento de artefatos nos vídeos. Isto ocorre porque quanto mais uma sequência de vídeo for submetida a processos de compressão, maior será a chance de perda ou alteração das informações, o que implica perda de qualidade visual percebida. Alguns exemplos de artefatos encontrados em vídeos são o ruído branco gaussiano, o ruído sal & pimenta, o borramento (*blurring*), o efeito da blocagem *(blocking)*, *color bleeding*, efeito escada, *ringing*, ruído tipo mosquito, cintilação, perda de pacotes e *aliasing* (Estrada, 2009). O efeito dos artefatos usados nesta tese no vídeo Foreman é apresentado na Figura 2.9.

(a) Blocagem. (b) Borramento. (c) Ruído Guassiano. (d) Ruído Sal & Pimenta.

Figura 2.9: Quadro do vídeo Foreman com diferentes artefatos.

Além do processo de compressão, o surgimento de artefatos é influenciado pela origem, pelo conteúdo, pelas condições do canal de transmissão, por processos como

codificação, armazenamento, filtragem, conversão ou transformação (Arthur, 2002).

Para verificar a eficiência de determinada métrica objetiva é necessário possuir amostras de vídeos que apresentem degradações encontradas em condições reais. A partir de então, é preciso que haja a produção controlada desses defeitos, isto é, que se conheça a quantidade e o artefato que está sendo inserido na amostra.

A seguir, estão descritos alguns tipos de artefatos que ocorrem em vídeos.

- Ruído Gaussiano Branco

  O ruído gaussiano branco é visualmente percebido como "chuviscos"na imagem (Albini, 2009). O ruído gaussiano branco é distribuído uniformemente quanto ao local de ocorrência, isto é, a probabilidade de um *pixel* ser afetado pelo ruído gaussiano branco independe da sua posição espacial, e apresenta uma distribuição gaussiana quanto ao valor da amplitude do ruído que é adicionado aos *pixels*.

- Sal & Pimenta

  O ruído sal & pimenta é visualmente percebido por pontos pretos e brancos ao longo da amostra de vídeo, que representam os valores de luminância mínimo e máximo, respectivamente. O ruído sal & pimenta ocorre por conta de falhas em sensores ou em memórias.

- Blocagem

  O efeito de bloco é causado pela perda de informação dos *pixels*, geralmente resultante de processos de compressão. Em vídeos, esse efeito é percebido como uma descontinuidade entre blocos adjacentes, que gera cantos acentuados (Arthur, 2002; Wu & Rao, 2006). O efeito de bloco é percebido pelo HVS quando ocorre em zonas transitórias de níveis de luminância, por exemplo, bordas de objetos, contornos e linhas.

- Borramento

  O efeito de borramento consiste na perda dos detalhes espaciais da imagem, e é visualmente caracterizado por uma imagem turva ou desfocada (Albini, 2009). Como o HVS é altamente sensível a estruturas em objetos de uma imagem, o borramento causa um maior desconforto ao HVS, em relação aos outros tipos de artefatos (Wang *et al.*, 2004).

- *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 por um macrobloco inteiro.

- Efeito escada

  A representação de linhas com outras orientações exige o uso dos coeficientes de frequência altas para a reconstrução exata. Assim, os codificadores baseados na transformada discreta de cosseno (*Discrete Cosine Transform* – DCT) quando eliminam as frequências mais altas geram o efeito de escada nas linhas inclinadas.

- *Ringing*

  É mais visível ao longo das bordas de alto contraste do que em áreas de superfície lisa. Esse efeito gera ondulações, que são maiores à medida que se aumenta o *ringing* e é uma consequência direta da quantização levando a irregularidades de alta frequência na reconstrução. O efeito de *ringing* ocorre nas componentes de iluminação e cor (Estrada, 2009).

2.7 Considerações Finais

O principal componente do sistema visual humano é o olho. No olho destacam-se o cristalino e a retina, que são os principais responsáveis pela detecção da imagem pelo cérebro. A visão central e periférica é uma das características do HVS, que depende da retina. Essa característica é utilizada para o desenvolvimento de novas métricas de avaliação de vídeo.

Um dos processamentos realizados em vídeos digitais é a segmentação. Com a segmentação é possível detectar partes do vídeo, como: linhas, descontinuidades, pontos e bordas. Para detecção de bordas são utilizados os operadores de gradiente e o laplaciano.

Os vídeos digitais sofrem degradações durante os processos de transmissão, codificação, filtragem e quantização. Os artefatos abordados neste capítulo foram o ruído sal & pimenta, borramento (*blurring*), o efeito da blocagem *(blocking)*, *color bleeding*,

efeito escadas, *ringing* e perda de pacotes. Com esses artefatos é possível simular o efeito de diferentes processos e assim avaliar a qualidade dos vídeos degradados.

Capítulo 3

Vídeo 3D

Vídeos em três dimensões (3D) oferecem uma sensação de profundidade da cena observada. Para que essa sensação possa ser percebida, é necessário telas específicas, existindo vários tipos de telas 3D disponíveis, que necessitam do auxílio de óculos ou não, e também diferentes tipos de algoritmos de renderização[1] 3D.

A terceira dimensão (profundidade) não existe na tela, ocorrendo por uma ilusão da mente. Isto é possível graças a um fenômeno natural chamado estereoscopia, que é a projeção de duas imagens, da mesma cena, em pontos de observação ligeiramente diferentes. O olho recebe a imagem de forma que em cada um seja projetada uma imagem do mesmo objeto e o cérebro automaticamente analisa e gera duas imagens sob perspectivas diferentes, o que produz a profundidade estereoscópica. Nesse processo são obtidas informações quanto à profundidade, distância, posição e tamanho dos objetos, gerando uma ilusão de visão em 3D.

Com o advento da tecnologia 3D, o usuário tem a liberdade de escolher o ângulo de visão para assistir determinada cena, e até mesmo produzir efeitos de rotação e pausa na imagem, o que não era possível em vídeos 2D, pois eram percebidas apenas as alterações das imagens com relação ao tempo e não um dos eixos do espaço, a profundidade. Além do mais, o efeito da estereoscopia melhora a percepção de nitidez, sentido de presença e a naturalidade das imagens.

[1] O processo de tratamento digital de imagens e vídeos consome muitos recursos dos processadores. Deste modo, são utilizados algoritmos com a função de fazer esses vídeos ou imagens trabalharem em baixa resolução, chamados de algoritmos de renderização.

3.1 Fatores Humanos na Percepção da Profundidade

A visão humana é considerada a referência na elaboração de técnicas para processamento de imagem, e a análise do seu comportamento define o Sistema de Visão Humana (HVS – *Human Vision System*). A imagem tridimensional se forma a partir do recebimento de duas imagens de uma mesma cena, uma em cada olho, que se sobrepõem com alguma diferença, originando perspectivas diferentes de uma mesma cena.

Uma variedade de opções é explorada pelo ser humano, para perceber a profundidade de uma imagem em três dimensões, tipicamente classificadas em profundidade binocular e monocular. O cérebro humano utiliza informações monoculares, como acomodação, oclusão linear, perspectiva aérea, tamanho relativo, densidade relativa e paralaxe[2] de movimento, para construir a percepção de profundidade com apenas um olho. Essas características podem ser observadas em *displays* 2D, como por exemplo, as televisões tradicionais.

A profundidade binocular requer que as informações de profundidade da imagem sejam exploradas com os dois olhos, de forma que as diferenças entre as imagens sejam percebidas e o efeito da profundidade seja formado.

3.1.1 Profundidade Binocular

Na profundidade binocular as imagens são formadas a partir da disparidade na projeção das imagens na retina. Ela é subdividida em estereoscopia e convergência.

Estereoscopia

O termo estereoscopia ocorre da justaposição dos termos gregos *stereo*, relativo a dois (duplo), e *scopos*, relativo a visão (observador). Em linhas gerais, a estereoscopia ou visão binocular trata-se de duas imagens de uma cena, que são projetadas nos olhos a partir de pontos de observação ligeiramente diferentes. O cérebro funde as duas imagens e, nesse processo, obtém informações quanto à profundidade, distância, posição e tamanho dos objetos, gerando uma sensação de visão tridimensional.

A estereoscopia ocorre devido ao fato de os olhos estarem horizontalmente separados por aproximadamente 6,3 cm, proporcionando a cada olho um ponto de vista único.

[2]Paralaxe é a distância horizontal entre a imagem esquerda e a direita em que aparecem os objetos em relação ao observador.

Utilizam-se duas perspectivas diferentes de uma mesma imagem vista pelos dois olhos. A imagem é formada na região central da retina (fóvea) de cada olho sob ângulos diferentes. A diferença entre os ângulos é chamada de disparidade binocular e fornece informações sobre as distâncias relativas dos objetos até o observador, a estrutura de profundidade e o ambiente em geral (Silva *et al.*, 2011).

No entanto, a obtenção das informações relativas, como distância, posição e tamanho dos objetos, não se deve somente à fusão do chamado par estereoscópico. Esses efeitos são inerentes à cena e independentes de mecanismos para captação de imagens, que são a iluminação, a oclusão e as sombras. A percepção do espaço decorre de características próprias dos mecanismos (ângulo de abrangência do campo visual), não sendo condicionada significativamente pela fusão do par estereoscópico.

A oclusão (no sentido de obstrução visual), a iluminação e as sombras (consequências diretas da iluminação) facilitam a identificação da posição relativa e do tamanho dos objetos. Pela iluminação e pelas sombras compreende-se a forma e o tamanho dos volumes. Pela oclusão, sabe-se qual objeto está mais próximo do ponto de vista (Malard *et al.*, 2008).

Existem vários tipos de estereoscopia, os principais são:

- Estereoscopia voluntária – utiliza um estereoscópio composto por lentes que direcionam uma das imagens do par estereoscópico para o olho direito e a outra para o olho esquerdo, permitindo visualizar a imagem em 3D.

- Estereoscopia polarizada – a luz é uma onda eletromagnética. No processo de estereoscopia por polarização da luz são utilizados filtros polarizadores que fazem com que as imagens do par estereoscópico projetadas sejam polarizadas em planos ortogonais (por exemplo, um plano vertical e um horizontal). Desta forma, cada olho recebe uma imagem diferente e a fusão dessas imagens, resulta na visão estereoscópica.

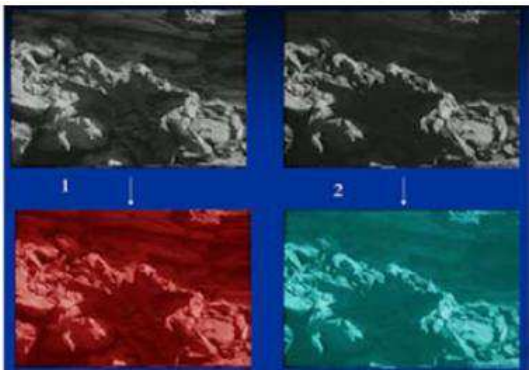
- Estereoscopia intermitente ou cintilamento – este tipo de estereoscopia baseia-se em estudos do olho humano. As imagens formadas na retina do olho humano persistem por cerca de 0,1 segundo, após a ocultação do objeto. Assim, o processo explora este fato para estabelecer a separação dos campos visuais dos dois olhos. Primeiro, projeta-se alternadamente as imagens da esquerda e da direita, durante cerca de 1/60 de segundo. Em seguida, com sincronia, veda-se o campo visual do

olho direito, enquanto a imagem da foto esquerda é projetada. Depois fecha-se o campo visual do olho esquerdo, enquanto a imagem da foto direita é projetada. Como a frequência de projeções sucessivas é alta, os olhos vêem, continuamente as imagens correspondentes e, assim, se obtém a visão tridimensional.

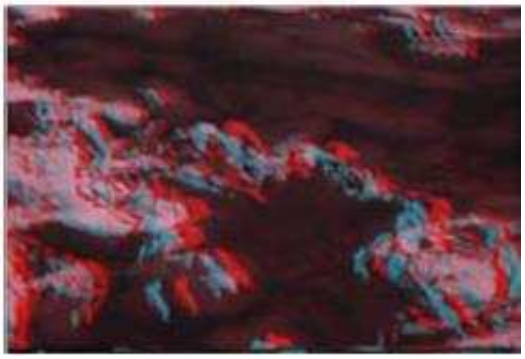
- Estereoscopia por holografia – a estereoscopia por holografia não utiliza pares de imagens estereoscópicas, pois a holografia é uma técnica que registra em filme a informação relativa a um objeto ou cena. Ela capta as informações de uma imagem em 3D incluindo profundidade e as grava também em 3D.

- Estereoscopia Anaglífica – O método de visualização estereoscópica anaglífica é o mais simples. Essa técnica caracteriza-se por colorizar com uma cor primária diferente cada uma das imagens referentes a cada olho, de modo que o espectador possa separar cada uma das imagens que se encontram misturadas na tela, utilizando óculos com uma lente vermelha e outra ciano (Mancini, 1994).

  Para a codificação desse tipo de vídeo estereoscópico, é necessário separar os canais RGB dos vídeos do par estereoscópico. Do vídeo que corresponde à visão do olho esquerdo é extraída a informação do canal vermelho, e do vídeo que corresponde à visão do olho direito extraem-se os canais azul e verde, como mostrado na Figura 3.1a. Forma-se, então, com a componente vermelha da visão do olho direito e as componentes azul e verde da visão do olho esquerdo um novo vídeo RGB. A imagem anaglífica resultante pode ser observada na Figura 3.1b.



(a) Processo de extração do canal vermelho e dos canais verde e azul do vídeo 3D.



(b) Exemplo de um quadro anaglífico.

Figura 3.1: Quadros anaglíficos (Andrade & Goularte, 2009).

Acomodação e Convergência

Os olhos automaticamente focam (acomodação) o objeto, fazendo com que ele sobressaia entre os outros objetos ao seu redor. Assim, imagens duplas na frente ou atrás do plano de fixação tendem a estar fora de foco e vão desaparecer com o aumento do borrado.

A acomodação é o processo responsável pela mudança do poder refrativo do olho, garantindo que a imagem seja focada no plano da retina. Quando um objeto de interesse é fixado, o olho se acomoda de tal forma que uma imagem nítida é percebida na retina. Uma boa acomodação exige um tempo de fixação mínimo de um segundo ou mais. No entanto, o olho humano pode tolerar uma certa desfocagem da retina sem reajustar a acomodação.

Os processos de acomodação e convergência estão intimamente relacionados, visto que a acomodação produz movimentos de convergência e a convergência produz acomodação. A acomodação é dirigida a imagens de objetos a uma distância de tela, enquanto a convergência está direcionada para as distâncias percebidas dos objetos. Em condições normais, essas distâncias coincidem, mas a situação é diferente quando o indivíduo está assistindo vídeos 3D, pois devido ao efeito da profundidade, os objetos apresentam distâncias diferentes.

3.1.2 Profundidade Monocular

O borramento de uma imagem simula a percepção da profundidade, pois os olhos automaticamente focam (acomodação) no objeto fixado, fazendo com que este objeto sobressaia dentre os outros objetos ao seu redor, provocando um borramento nos objetos mais distantes (Silva *et al.*, 2011).

Na Figura 3.2 pode ser observado o esquema geométrico para análise da formação da imagem na retina, por meio do borramento da imagem.

Fixando o ponto P, a capacidade óptica (convergir ou divergir a luz) dos olhos é ajustada para que a imagem do ponto P esteja focada na região central da retina (fóvea). No entanto, a capacidade óptica dos olhos varia em certa frequência e, assim, o olho pode tolerar certa quantidade de desfoque da retina sem reajuste de acomodação.

Esta diferença de capacidade óptica é conhecida como a profundidade de foco ocular e é apresentado em dioptrias[3] (D). Em outras palavras, o ponto P pode ser deslocado

[3]Dioptria é uma unidade de medida que afere a capacidade de divergência de um sistema óptico

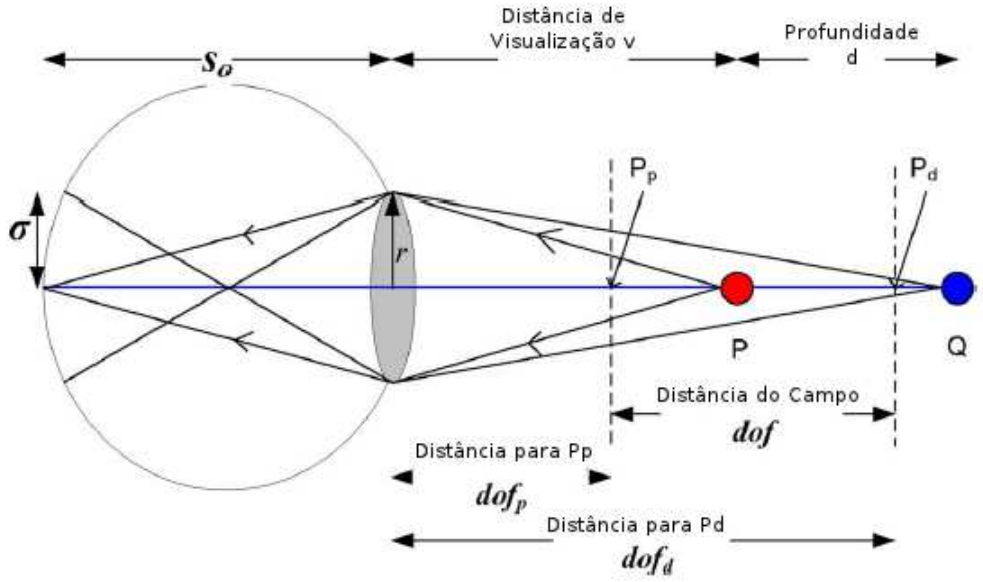


Figura 3.2: Geometria do desfoque na retina.

ao longo do eixo óptico de uma distância determinada, sem que se perceba o desfoque da imagem. Essa distância é conhecida como a campo de profundidade dof.

Os pontos mais próximo e mais distante do limite são conhecidos como ponto proximal (P_p) e ponto distal (P_d). Portanto, qualquer objeto dentro do limite é percebido nitidamente e os objetos fora do limite são percebidos borrados, o que estimula a sensação de profundidade (Silva *et al.*, 2011). A Equação 3.1 relaciona a distância do olho ao ponto Q, que é $v + d$, e o valor do borrão da retina σ

$$v + d = \frac{F \cdot r \cdot S_o}{r \cdot S_o - F}, \tag{3.1}$$

em que v representa a distância visual, d a profundidade, F a distância focal das lentes, r o raio de abertura dos olhos e S_o é a distância do olho a partir da retina. A magnitude do campo de profundidade (dof) difere de pessoa para pessoa, dependendo da profundidade do foco ocular. A Equação 3.2 representa o cálculo da profundidade do foco ocular T relacionando o ponto proximal e distal, de acordo com os padrões das equações ópticas,

(m^{-1}). A óptica é a unidade de medida da potência de uma lente corretiva (popularmente conhecido como grau).

$$T = \frac{1}{dof_p} + \frac{1}{dof_d}, \tag{3.2}$$

e os valores de dof_p e dof_d são dados por:

$$dof_p = \frac{2v}{2 + v \cdot T}, \tag{3.3}$$

$$dof_d = \frac{2v}{2 - v \cdot T}. \tag{3.4}$$

O valor de T ele varia entre 0,6D e 0,8 D. Portanto, quando um objeto está no foco e é visto nítido e um outro objeto, que está além do dof, é visto borrado, tem-se a sensação de profundidade. Desta forma, são definidos limites para os objetos projetados em uma tela estereoscópica para minimizar o desconforto visual.

Uma outra forma da profundidade monocular é a profundidade por estímulos geométricos. As geometrias relacionadas às profundidades sugeridas são a perspectiva linear, o tamanho conhecido, o tamanho relativo, a altura da imagem, a interposição e o gradiente da textura (Zhang *et al.*, 2011).

Alguns desses estímulos são mais fortes do que outros. A interposição, por exemplo, pode ensinar a ordem de profundidade dos objetos, mas não a distância em profundidade entre eles. Alguns estímulos podem ser difíceis de serem usados em uma aplicação para a estimativa da profundidade. Por exemplo, as informações relacionadas à dimensão dos objetos são difíceis de serem usadas, pois requerem a identificação de objetos e o conhecimento das dimensões originais desses objetos. Os mais comuns são os estímulos geométricos da perspectiva linear e da altura da imagem.

A perspectiva linear se refere à propriedade de linhas paralelas convergirem para uma distância infinita, ou equivalentemente, um objeto de tamanho fixo vai produzir um menor ângulo de visão quanto mais distante estiver do olho. Esta característica é utilizada na estimativa da profundidade por meio da detecção de linhas paralelas nas imagens e identificação do ponto em que essas linhas convergem (ponto de fuga). Em seguida, uma atribuição adequada de profundidade pode ser derivada com base na posição das linhas e do ponto de fuga (Zhang *et al.*, 2011).

A altura na imagem indica que os objetos que estão mais próximos da parte inferior da imagem estão geralmente mais próximos dos objetos que estão na parte superior da imagem.

Além da perspectiva linear e altura da imagem, também é possível recuperar a profundidade a partir da textura. O objetivo é estimar a forma de um objeto com base em sugestões de marcas na superfície ou sua textura (Torralba & Oliva, 2002). Esses métodos, no entanto, são normalmente restritos a tipos específicos de imagens e não podem ser aplicados à conversão do conteúdo de vídeo 2D para 3D em geral.

Tamanho Relativo

O tamanho relativo representa a profundidade da imagem, levando em consideração a geometria formada na retina. A imagem projetada atua como um estímulo importante para a percepção da profundidade, em que o ângulo visual de um objeto projetado diminui à medida que a distância até o objeto aumenta e vice-versa.

Utilizando a propriedade de semelhança de triângulos, a Equação 3.5 mostra a relação entre o tamanho da imagem na retina R em função do tamanho real da imagem H, a distância focal F e a distância do objeto D. Assim,

$$R = \frac{HF}{D}. \tag{3.5}$$

O aumento na distância do objeto reduz o tamanho da imagem formada na retina. O cérebro interpreta essa diferença como uma mudança de profundidade ΔR. A Equação 3.6 representa matematicamente esse comportamento.

$$R - \Delta R = \frac{HF}{(D + \Delta d)}. \tag{3.6}$$

É possível estimular uma mudança no tamanho da imagem projetada na retina por meio da mudança do tamanho do objeto ΔH. Com isso, a mudança no tamanho dos objetos que são fisicamente similares proporciona a sensação de profundidade devido a uma mudança do tamanho da imagem na retina.

$$R - \Delta R = (H - \Delta H) \cdot \frac{F}{D}. \tag{3.7}$$

Profundidade de cor e intensidade

Variações na quantidade de luz que chega ao olho também podem fornecer informações da profundidade dos objetos. Este tipo de variação aparece nas imagens captadas como variações de intensidade ou alterações na cor. Os cálculos da profundidade que são baseadas neste mecanismo de dispersão atmosférica incluem a distribuição de luz e de sombra, a percepção da imagem de fundo e o contraste local.

Dispersão atmosférica refere-se à dispersão dos raios de luz pela atmosfera, produzindo um tom azulado e menor contraste entre os objetos que estão longe e melhor contraste entre os objetos que estão em uma estreita faixa (Cozman & Krotkov, 1997). Com base em regras de cor, que são aprendidas heuristicamente usando um grande número de imagens da paisagem, a detecção da região semântica é realizada para dividir imagens da paisagem em seis regiões, como céu, montanha mais distante, montanha distante, montanha próxima, terra e outros (Zhang *et al.*, 2011).

3.2 Captura e Formação da Imagem Tridimensional

Uma imagem que tem ou parece ter altura, largura e profundidade é tridimensional (ou 3D). Uma imagem que possui altura e largura mas não possui profundidade é bidimensional (ou 2D). As imagens 2D são úteis para comunicar algo simples, rapidamente. Por outro lado as tridimensionais relatam histórias mais elaboradas, vídeos que transmitem uma sensação de realidade aumentada, mas precisam carregar muito mais informações (Smolic *et al.*, 2005).

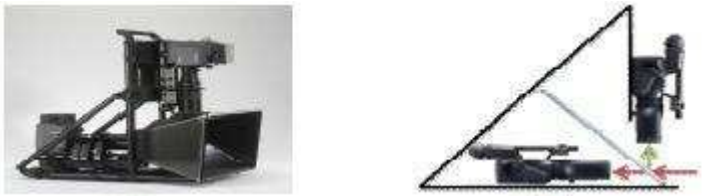
A visualização de vídeos em 3D por parte do espectador é um processo complexo, e assim sendo, a captação das imagens que formaram os vídeos tridimensionais deve simular as características do sistema visual humano para se ter um resultado final de qualidade aceitável. Grande parte dos sistemas de filmagem para conteúdos em 3D requer que a captação seja feita em dois canais independentes, um para cada olho e, para tal, recorre-se a duas câmeras. Uma câmera capta a imagem do olho esquerdo e outra a do olho direito. Por essas duas perspectivas é gerada a percepção espacial dos eventos (Fragoso *et al.*, 2012).

Para esse sistema se faz necessária a observação de dois aspectos essenciais: os parâmetros de gravação e a distância interocular. Os parâmetros de gravação devem ser os mesmos nas duas câmeras; caso contrário, as imagens teriam características

diferentes, resultando em uma baixa qualidade de visualização, ou em vídeos com efeito fantasma.

Atualmente, existem dois métodos de captura de vídeos 3D, utilizando pares de câmeras. O primeiro método é denominado *Side-By-Side Rigs* (Figura 3.3a), que consiste na colocação das câmaras lado a lado, permitindo um alinhamento mais rápido e preciso. Essa técnica é mais indicada para filmagens a grandes distâncias e imagens abertas (cenários, paisagens), uma vez que não permite a aproximação requerida entre lentes, de modo a criar o efeito 3D em curtas distâncias.

(a) Câmera *side-by-side rigs*.



(b) Câmera *beam-splitter rigs*.

Figura 3.3: Tipos de Câmeras.

O segundo método de captura é chamado de *beam-splitter rigs* (Figura 3.3b) e consiste em uma câmera horizontal que filma a imagem que atravessa o espelho e uma câmera vertical que filma a imagem refletida pelo espelho. Ambas as câmeras podem capturar uma parte ou a totalidade da área de imagem, podem possuir ajuste interocular e mecanismo de convergência (por meio de ângulos de inclinação) para obter melhores resultados (Fragoso *et al.*, 2012).

3.3 Conforto Visual em Vídeos 3D

Na literatura, os termos fadiga visual e desconforto visual têm sido usados como sinônimos para descrever o desconforto que pode acompanhar o uso de tecnologias de imagem. No entanto, uma distinção entre esses dois termos é apresentada para auxiliar

as metodologias de medição (Tam *et al.*, 2011).

O termo fadiga visual refere-se a uma diminuição no desempenho do sistema visual produzido por uma alteração fisiológica. Portanto, a fadiga visual pode ser avaliada com medida fisiológica, tal como mudança na resposta da acomodação, no diâmetro da pupila e na característica do movimento ocular. Desconforto visual, por outro lado, refere-se à sensação subjetiva de desconforto que acompanha a mudança fisiológica. Assim, o conforto visual pode ser medido pelo questionamento ao espectador para relatar o seu nível de percepção de conforto visual (Tam *et al.*, 2011).

3.3.1 Medidas de Conforto Visual

Não existe metodologia padrão para a medição do conforto visual para imagens estereoscópicas. Por exemplo, a *International Telecommunications Union* (ITU) tem apenas uma recomendação sobre os métodos subjetivos para geração de imagens estereoscópicas (ITU-R, 2000). No entanto, a recomendação considera apenas a qualidade de imagem e a profundidade.

Por conta dessa deficiência, muitos pesquisadores têm utilizado modificações indicadas na ITU-R Rec. BT. 500 (ITU-R, 2010), que se destina à avaliação da qualidade da imagem. Alguns aspectos desse método, como modos de apresentação (apresentação simples, dupla, ou contínua) e sequência de duração (de alguns segundos até vários minutos de duração) geralmente têm sido mantidos. No entanto, na ausência de orientações comuns, outros aspectos, tais como condições de visualização, critérios para seleção de materiais e escalas de classificação, têm sido sempre diferentes.

Em particular, a maioria dos pesquisadores tem preferido utilizar escalas de conforto personalizadas no lugar das que são utilizadas para a avaliação da qualidade de imagem (Tam *et al.*, 2007).

O desconforto visual pode variar ao longo do tempo, provavelmente aumentando após a exposição prolongada ao estímulo visual. Assim, pesquisadores têm sugerido a utilização de uma avaliação contínua do conforto visual semelhante ao método *Single Stimulus Continuous Quality Evaluation* (SSCQE) descrito na ITU-R Rec. BT. 500 (ITU-R, 2010). Na avaliação contínua, os espectadores são apresentados a uma sequência de vídeo de longa duração, por exemplo, 5, 15 ou 60 minutos, e eles são convidados a avaliar a característica de interesse, neste caso o nível de conforto, de forma contínua.

Alguns pesquisadores usaram questionários para capturar a natureza complexa do conforto visual. Esses questionários listam um conjunto de sintomas ou fontes potenciais de desconforto visual e pedem aos telespectadores para identificar com mais precisão a fonte de seu desconforto (Kuze & Ukai, 2008).

3.3.2 Fatores que Afetam o Conforto Visual

Com o aumento da demanda por serviços de TV 3D, cresce a preocupação com a segurança e a saúde de visualização de imagens estereoscópicas. Alguns pesquisadores mencionam que as imagens estereoscópicas poderiam trazer prejuízos para os espectadores, especialmente para as crianças, cujo sistema visual ainda está em desenvolvimento e, assim, são mais suscetíveis às influências externas (Tam *et al.*, 2011).

Apesar de muita pesquisa, ainda não há como produzir conteúdos estereoscópicos 3D que tenham garantia de estarem livres do desconforto visual. No entanto, as pesquisas têm identificado vários fatores que poderiam afetar negativamente o conforto visual.

Em relação à concisão e à clareza, os fatores são agrupados em cinco categorias: conflito de acomodação da convergência, distribuição de paralaxe, incompatibilidade binocular, inconsistência de profundidade e inconsistência cognitiva (Tam *et al.*, 2011).

Conflito de Acomodação e Convergência

Em geral, a paralaxe excessiva causa desconforto visual (Tam *et al.*, 2011). Isso ocorre, pois as imagens com paralaxes maiores são mais difíceis de se fundir. Outra possível causa de desconforto visual de paralaxe excessiva é a acomodação e o conflito de convergência criados pelo tipo de monitor estereoscópico em uso atualmente.

A acomodação e a convergência são normalmente atreladas ao se ver os objetos em uma cena natural. No entanto, a interação normal entre esses dois processos pode ser interrompida durante a visualização de imagens estereoscópicas (Miksicek, 2006).

Em geral, assume-se que, para minimizar o conflito da acomodação da convergência, as disparidades em uma imagem estereoscópica devem ser pequenas o suficiente para que a profundidade percebida dos objetos caia em de uma zona de conforto.

Assim, o conflito da acomodação da convergência é reduzido se a profundidade percebida dos objetos for limitada à profundidade de campo do olho, para a qual as respostas da acomodação são minimizadas. Resultados consistentes sugerem que a

profundidade de campo pode ser usada para definir uma zona de visão confortável (Tam *et al.*, 2011).

Distribuição de Paralaxe

Não só a disparidade da magnitude, mas também o tipo e a distribuição, no espaço e no tempo, das disparidades parecem afetar o conforto visual. Os resultados em Tam *et al.* (2011) indicam que as cenas estereoscópicas são mais confortáveis para assistir quando a distribuição de paralaxe é tal que a parte inferior da imagem aparece mais perto e a parte superior aparece mais distante.

Além disso, imagens que têm a maior parte dos objetos no fundo da imagem proporcionaram maior conforto. Os resultados também indicaram uma diminuição do conforto visual para as cenas que têm uma alta paralaxe, e uma grande variação ao longo do tempo de paralaxe (Tam *et al.*, 2011).

Os resultados em Tam *et al.* (2011), indicam que a taxa de mudança na magnitude da disparidade com o tempo foi mais prejudicial para o conforto visual do que a magnitude absoluta das disparidades cruzadas e descruzadas. Em suma, a distribuição das disparidades em imagens estereoscópicas e sua mudança no tempo parecem ter um impacto significativo sobre o conforto visual.

Distorção *Keystone* e Curvatura do Plano de Profundidade

A distorção *keystone* é tipicamente introduzida na captura de imagens estereoscópicas que são configuradas por câmeras convergentes, de forma que a câmera direita e esquerda estejam posicionadas em um ângulo específico. Neste caso, os sensores da imagem de cada uma das câmeras são direcionados para planos ligeiramente diferentes. Isso resulta em uma imagem de forma trapezoidal em direções opostas para o olho esquerdo e direito.

Em imagens estereoscópicas essas formas opostas da imagem trapezoidal induzem paralaxe vertical e horizontal incorretas. O erro na paralaxe vertical é chamado de *distorção de keystone*, que é mais visível nos cantos da imagem e aumenta com o aumento da distância da base da câmera, diminuindo a distância de convergência e diminuindo a distância focal da lente. O erro introduzido na paralaxe horizontal é chamado de curvatura do plano de profundidade, segundo o qual os objetos no canto da imagem aparecem mais longe do observador em comparação com os objetos no meio

da imagem (Tam *et al.*, 2011).

É visível que os dois fenômenos causam um certo desconforto visual ao telespectador. No entanto, se as câmeras forem posicionadas corretamente a distâncias de convergência entre 60 a 240 cm, os efeitos negativos podem ser minimizados de tal forma que a exposição prolongada a imagens ou vídeos 3D não causem desconforto.

Inconsistências na Profundidade

A inconsistência na profundidade refere-se à informação de profundidade contraditória, resultante de erros de disparidade. Esses erros na informação da profundidade da imagem estereoscópica podem afetar o conforto visual.

Normalmente, a informação da profundidade de uma imagem estereoscópica é incorporada na disparidade horizontal entre as imagens do olho esquerdo e do olho direito. O mapa de profundidade é um método alternativo de transmissão de informações estereoscópicas. Um mapa de profundidade é geralmente associado a uma imagem ou quadro de vídeo.

O mapa de profundidade é uma matriz contendo a profundidade dos *pixels* na imagem ou quadro de vídeo associado. O uso de mapas de profundidade é baseado em uma técnica chamada de processamento da imagem baseada na profundidade (DIBR), que pode ser usada para gerar novos pontos de vista da câmera virtual de uma cena, dada uma imagem bidimensional (2D) da cena e seu correspondente mapa de profundidade. Sendo uma possível fonte de erro para as informações de profundidade (Zhang & Tam, 2005).

Uma segunda fonte de inconsistência na profundidade decorre da conversão de 2D em 3D. Essa técnica calcula um mapa de profundidade a partir de um sinal padrão 2D e usa esse mapa para gerar uma versão do sinal em 3D. Os mapas de profundidade obtidos a partir de um sinal 2D, inevitavelmente contêm informações erradas (Tam *et al.*, 2011). Esses erros também podem resultar em inconsistências na profundidade e em uma diminuição do conforto (Tam & Zhang, 2006).

3.4 Extração de Profundidade em Vídeos 3D

A disparidade refere-se às diferenças espaciais entre duas imagens, esquerda e direita, e é definida como a distância horizontal entre dois *pixels* correspondentes. Eles usualmente se encontram em posições diferentes nos quadros ou pode acontecer que o

pixel não tenha seu homólogo. Determinar o valor da disparidade é essencial para se determinar a distância dos objetos às câmeras.

Dada a informação da disparidade associada a qualquer par de imagens, o cérebro pode gerar a percepção de profundidade, fundindo estas imagens. A ideia de gerar a percepção de profundidade a partir das informações da disparidade inspirou a captura de imagens estereoscópicas, de uma mesma cena, usando duas câmeras com a mesma configuração.

Os dados do cálculo da disparidade são armazenados na forma de mapas de disparidades, que são vetores de informação com os valores da distância dos *pixels* de vistas diferentes, para cada coordenada da imagem. Podem ser representados também por uma imagem resultante de todas as diferenças entre os *pixels*, que são os mapas de profundidade. Alguns problemas, como oclusão, podem dificultar a obtenção da disparidade.

A forma mais simples de calcular a disparidade entre todos os *pixels* da imagem direita e esquerda é expressa pela Equação 3.8. Nessa equação, são calculadas as diferenças entre dois *pixels* homólogos, chamado de Método de Correspondência.

$$d(x, y) = |f_l(x, y) - f_r(x, y)|. \tag{3.8}$$

O Método de Correspondência consiste em determinar um *pixel* homólogo da imagem esquerda, na imagem da direita. A determinação desses *pixels* homólogos depende da qualidade da imagem e da precisão com que a cena foi reconstruída. Os mapas de profundidade podem ser classificados como mapas esparsos ou mapas densos. Os mapas esparsos são obtidos a partir do cálculo da disparidade para alguns pontos, e os mapas densos são obtidos pelo cálculo da disparidade por todas as áreas da imagem.

Os mapas de profundidade relacionam as distâncias das superfícies dos objetos de uma cena a partir de um ponto de vista. Essa representação é feita por intermédio de uma imagem em escala de cinza, na qual a intensidade dos *pixels* é mais clara quanto mais próximo da câmera estiver o objeto correspondente. Eles são constituídos majoritariamente por zonas suaves correspondentes a regiões com profundidade semelhante, e zonas com variações abruptas associadas às bordas dos objetos localizados a diferentes profundidades (Graziosi *et al.*, 2012), (Lucas *et al.*, 2012).

Uma vez determinado o valor da disparidade entre os *pixels* da imagem, pode-se

converter esse dado em distância física (Chan *et al.*, 2007). Para se determinar a profundidade Z com base nos valores de disparidade, é necessário conhecer os parâmetros físicos das câmeras, t_c e f, como demonstrado na Fórmula 3.9

$$Z = \frac{t_c \cdot F}{d(x, y)}, \tag{3.9}$$

em que t_c representa a distância entre as câmeras e F representa a distância focal de cada câmera. Os valores de t_c e f devem ser descritos nas especificações de cada vídeo, pois são inerentes a ele, levando em consideração a estrutura da câmera (Zhang & Tam, 2005).

Convém ressaltar que o valor da disparidade pode ser positivo, negativo ou nulo. No caso de uma disparidade nula, entende-se que foi encontrada uma distância tal que, não se consegue distinguir uma diferença de posição, pois aquele ponto encontra-se na mesma posição para ambas as câmeras. Na disparidade positiva, a imagem será formada "para dentro" do plano do *display*, enquanto que a disparidade negativa será formada "para fora" do plano do *display*.

A renderização baseada na disparidade da imagem (*Depth Image Based Rendering* – DIBR) é o método que vem sendo bastante utilizado por conseguir uma maior precisão para o mapa de profundidade, porém mais complexo do que os outros métodos.

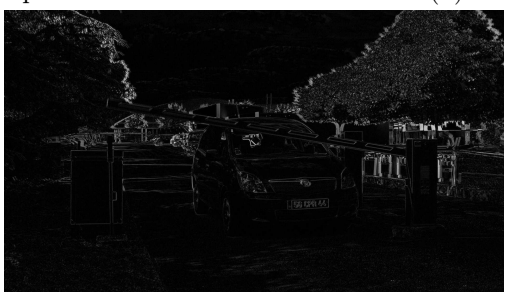
3.4.1 Renderização Baseada na Disparidade da Imagem

A renderização baseada na disparidade da imagem é uma técnica utilizada para formação da imagem 3D, em que ela é formada a partir da multiplexação das imagens esquerda (Figura 3.4a) e direita (Figura 3.4a) em uma imagem de referência e no mapa de profundidade, Figura 3.4c. O mapa de profundidade correspondente, às informações armazenadas de profundidade com 8-*bits* em valores de cinza, com 0 no lugar mais distante e 255 no local mais próximo, e a mesma resolução espaço-temporal (Hur *et al.*, 2005).

O posicionamento das câmeras ou a seleção dos parâmetros de renderização tem um impacto direto sobre a profundidade percebida de uma cena. Vários parâmetros têm influência sobre o efeito estereoscópico incluindo a distância real da cena, a distância interocular, a distância entre as lentes da câmera, o ângulo entre as câmeras, que podem ser tanto paralelas ou convergentes. A fim de proporcionar ao observador um ótimo

(a) Vista Esquerda.

(b) Vista Direita.



(c) Disparidade.

Figura 3.4: Vistas esquerda e direita e a disparidade do vídeo src01_hrc00 do banco de dados NAMA3DS1-COSPAD1.

efeito 3D é importante compreender a influência destes parâmetros sobre a qualidade do conteúdo percebido 3D. A distância interaxial e a distância de convergência são cruciais para determinar a qualidade da imagem estereoscópica, mas é difícil de determinar a distância e o ângulo adequados dentro de uma faixa de valores.

Na Figura 3.5 é apresentada a geometria das câmeras, em paralelo, para a geração da imagem 3D. Os parâmetros F e t_c são o distância focal e a distância entre as duas câmeras, C_r e C_l, respectivamente. Os pontos (X_r, y), (X_l, y) e (X_c, y) correspondem às posições dos *pixels* na imagem de referência e o ponto virtual, que corresponde ao ponto P com profundidade Z.

As Equações 3.10 e 3.11 mostram como encontrar esses valores (Zhang & Tam, 2005).

$$X_l = X_c + \frac{t_c \cdot \alpha_u}{2Z(X_c, y)} - z, \tag{3.10}$$

$$X_r = X_c - \frac{t_c \cdot \alpha_u}{2Z(X_c, y)} + z. \tag{3.11}$$

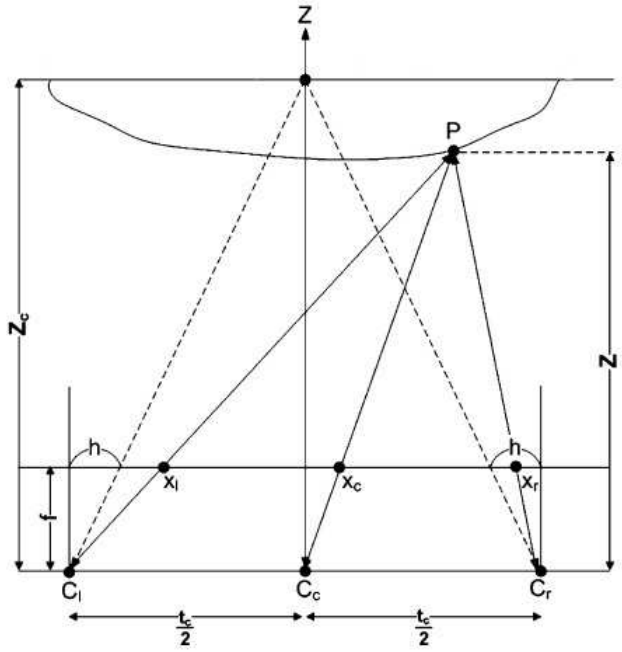


Figura 3.5: Geometria da câmera para geração das duas imagens.

O valor do fator de escala α_u é obtido dividindo a distância focal pelo tamanho do *pixel* e pode ser determinado pela Fórmula 3.13. O valor de z depende do valor da distância de convergência Z_c (a distância além das câmaras em que os eixos ópticos das câmeras se cruzam) e é calculado como segue

$$z = \frac{t_c \cdot \alpha_u}{2Z_c}. \tag{3.12}$$

$$\alpha_u = 0.0005 \cdot W \tag{3.13}$$

Os dois parâmetros, Z_c e t_c, podem ser usados para controlar e adaptar a impressão da profundidade pelo receptor. Por exemplo, se a distância entre as câmeras aumenta, a disparidade entre as duas imagens aumenta e, consequentemente, a percepção da profundidade aumenta. Essa característica do DIBR é bastante útil pois a percepção da profundidade pode ser controlada pelo usuário de acordo com sua preferência.

O telespectador pode sentir algum desconforto visual se as diferenças entre X_l e X_r forem maiores do que 3% do valor da largura da imagem (Cheolkon & Jiao, 2011).

Desta forma, a máxima diferença será de 3% do valor da largura da imagem, w.

$$X_l - X_r < 0,03 \cdot W, \tag{3.14}$$

em que w é o valor da largura da imagem.

Alguns problemas são inerentes ao DIBR, como a desoclusão[4]. Este problema se caracteriza por áreas que não são visíveis nos mapas de profundidade e são visíveis nas imagens esquerda e direita. A Figura 3.6 representa uma imagem gerada com o mapa de profundidade com a desoclusão. Para reduzir esse problema, um filtro passa-baixa é aplicado para reduzir as áreas de desoclusão e manter uma boa qualidade na profundidade das imagens (Cheolkon & Jiao, 2011).

Figura 3.6: Áreas de desoclusão na imagem do vídeo "Ballet" representado por pontos brancos. a) Fenômeno da desoclusão. b) Vista esquerda de "Ballet"com buracos.

Muitos pesquisadores têm estudado técnicas DIBR para serem aplicadas em serviços T-DMB (*Terrestrial - Digital Multimedia Broadcasting*), juntamente com técnicas de compressão e transmissão, de forma a melhorar os serviços 3D em dispositivos móveis.

Pré-Processamento

O pré-processamento de mapas de profundidade inclui duas questões: escolher a distância de convergência $Z(c)$ (o chamado *zero-parallax setting* (ZPS)) e o ajuste do mapa de profundidade.

A utilização de métodos de pré-processamento da imagem se faz necessária para que, ao se calcular a profundidade, não haja deformidades que impossibilitem sua percepção

[4]Problema relacionado à visualização de pontos sem textura nas imagens virtual esquerda e direita.

adequada. O parâmetro Distância de Convergência, conhecida como paralaxe zero é dado por

$$Z(c) = \frac{Z_{near} - Z_{far}}{2}, \tag{3.15}$$

na qual Z_{near} e Z_{far} são os planos mais próximo e mais distante do mapa de profundidade.

Para o ajuste do mapa de profundidade utilizam-se filtros assimétricos para a retirada do borramento da imagem, entre eles o sobel e o gaussiano (Zhang & Tam, 2005). Os dois filtros são importantes, pois o primeiro remove os ruídos e preserva a profundidade, o segundo baseado na suavização do gradiente, suaviza a profundidade nas imagens originais na direção horizontal e reduz os buracos.

Conversão *Disparity to Depth* (DtoD)

Em métodos como o DIBR a conversão de disparidade para profundidade (DtoD) é realizada no receptor. No entanto, nesse método, a conversão é realizada no transmissor porque a informação é transmitida pelo canal de transmissão em vez das informações de profundidade. Isto é muito eficiente, pois o transmissor tem grande poder computacional e maior capacidade de armazenamento em relação ao receptor (Cheolkon & Jiao, 2011).

Após o pré-processamento, a conversão DtoD é realizada por

$$Z(v) = \frac{1}{\frac{1}{Z_{near}}\left(\frac{v}{255}\right) + \frac{1}{Z_{far}}\left(1 - \frac{v}{255}\right)}, \qquad v \in [0, \cdots, 255], \tag{3.16}$$

em que $Z(v)$ é dada em milímetros, mede a v-ésima distância entre um objeto e um observador, Z_{near} e Z_{far} representam a maior e a menor distância entre o objeto e um observador, respectivamente.

Com isso, a disparidade de cada *pixel* é calculada a partir da distância de profundidade da Equação 3.16. É escrita desta forma, devido a fatores humanos que foram levados em consideração, como a distância de profundidade percebida dos objetos mais próximos, que é muito menor que a distância da profundidade de objetos mais distantes.

A disparidade entre cada *pixel*, apresentada na Equação 3.17, é calculada utilizando os valores de $Z(v)$.

$$D(v) = \frac{\alpha_u t_c}{2}\left(\frac{1}{Z(v)} - \frac{1}{Z_c}\right) \tag{3.17}$$

Os valores de $Z(v)$ e $D(v)$ são inversamente proporcionais. Além disso, o máximo valor de $D(v)$ é aproximadamente cinco. O valor máximo é devido ao fato de que os usuários se sentem desconfortáveis quando o valor da paralaxe entre as vistas esquerda e direita é maior do que 3% do tamanho da largura da imagem (Cheolkon & Jiao, 2011).

A maioria dos vídeos estereoscópicos tem valores de disparidade positivas ou negativas, que constituem os objetos reais 3D para frente e para trás do monitor 3D.

3.5 Considerações Finais

Os vídeos em três dimensões oferecem a sensação de profundidade, que acrescenta maior realidade às cenas. Essa realidade é percebida pela captação da imagem por pontos de observações diferentes, um em cada olho. O cérebro processa as informações produzindo a profundidade.

A junção das duas imagens é chamada estereoscopia, que é dividida em voluntária, polarizada, intermitente, holográfica e anaglífica. Para cada tipo de estereoscopia, é necessário um dispositivo específico para a visualização do vídeo 3D.

Para a captura dessas imagens são utilizados dois métodos, com duas câmeras cada. O primeiro método, o *side-by-side rigs*, que tem duas câmeras lado a lado, é hoje o método mais utilizado, e o método *beam-splitter rigs*, que consiste em uma câmera na horizontal e outra na vertical, uma com a perspectiva frontal e a outra com a profundidade.

O conforto visual é uma característica importante para a aceitação da tecnologia de vídeo 3D. Assim, vários estudos têm sido realizados para encontrar métodos para avaliar o possível desconforto.

O desconforto visual pode ser provocado pelo conflito de acomodação de convergência, que é provocado quando a profundidade passa do limite do campo do olho. Uma outra maneira de introduzir desconforto, é colocar a parte inferior da imagem mais longe que a parte superior.

O último ponto que foi tratado neste capítulo foi o mapa de profundidade. Ele pode

ser definido como um canal de imagem ou a imagem que contém as informações relativas da distância das superfícies de objetos da cena a partir de um ponto de vista. Além disso, os mapas de profundidade são imagens que possuem características bem distintas das imagens de textura, geralmente apresentam superfícies suaves, sem textura e com bordas bem definidas.

Capítulo 4

Métricas para Avaliação de Vídeo

A Qualidade da Experiência (QoE – *Quality of experience*) tornou-se um termo comumente usado para descrever a aplicação e a qualidade dos serviços de vídeo e multimídia para o utilizador. A QoE abrange muitos aspectos diferentes, na qual a qualidade do vídeo é apenas um deles (Winkler, 2005). Para a avaliação da QoE existem vários fatores que podem interferir, como por exemplo:

- Interesses individuais do espectador, como programas favoritos, que determinam o nível e o foco de atenção;
- Expectativas da qualidade do espectador, por exemplo, um filme exibido em um cinema contra um pequeno clipe assistido em um dispositivo móvel;
- Experiência de vídeo do espectador, que também determina as expectativas da qualidade (uma vez que ele assistiu o conteúdo em alta definição é difícil se acostumar com vídeos de baixa definição);
- Tipo de *Display* (CRT, LCD, etc) e propriedades (dimensão, resolução, brilho, contraste, cor, tempo de resposta);
- Exibição de configurações e condições, tais como a distância de visualização ou a luz do ambiente;
- Qualidade e sincronização do áudio que o acompanha;
- Interação com o serviço ou com o *display* do dispositivo (por exemplo, controle remoto, guia eletrônico de programação).

A variedade e a subjetividade de alguns destes fatores indicam que a medição da qualidade de um sistema de vídeo digital é um problema complexo. A maioria das métricas de qualidade leva em consideração um pequeno subconjunto dos fatores listados e se concentra em medir a fidelidade visual do vídeo em termos de distorções introduzidas por diversas etapas de processamento (principalmente de compressão e transmissão). Mesmo restringindo o problema, duas questões permanecem desafiadoras:

- Sistemas de vídeo são complexos e têm muitos componentes, incluindo aqueles para captura e o *hardware* da tela, conversores, multiplexadores, *codecs*, *streamers*, roteadores, *switches*. Todos eles processam o vídeo de alguma forma, o que pode potencialmente afetar a sua qualidade.

- A percepção visual é ainda mais complexa. Para medir a qualidade de uma forma significativa, precisa-se entender como as pessoas percebem o vídeo e sua qualidade.

A importância da medida da qualidade é notada quando se considera a diversidade de serviços disponíveis que usam vídeo. A princípio, as pessoas pedem qualidade máxima. Mas há uma relação entre qualidade, disponibilidade, acessibilidade e custo, que cria um conceito de melhor qualidade dentro de certas condições.

Os usuários de alguns tipos de serviços aceitam o sacrifício na qualidade para redução de custos ou para que seja possível assistir o vídeo em um equipamento portátil (Arthur, 2002). Portanto, precisa-se medir ou qualificar um sinal dentro de certas condições para se determinar o quanto ele atende as expectativas.

A qualidade do vídeo pode ser medida utilizando métricas de qualidade objetiva ou de forma subjetiva, por meio de experimentos com observadores humanos. Os procedimentos utilizados em um experimento da qualidade subjetiva de vídeo são descritos nas recomendações BT.500 do ITU-T e BT.500-11 do ITU-R, para serviços de TV, e P.910 da ITU-T, para aplicações multimídia (ITU-T, 1999).

As medidas subjetivas são realizadas por um grupo de pessoas, que subjetivamente classificam a qualidade do sinal processado, comparando-o com outro sinal (Bernardino Júnior, 1998), (Kim *et al.*, 2008).

As medidas objetivas como a relação sinal ruído (SNR – *Signal-to-Noise Ratio*), relação sinal-ruído de pico (PSNR – *Peak Signal to Noise Ratio*) ou erro médio quadrático (MSE – *Mean Square Error*) podem operar em tempo real, comparando a imagem sob

teste como a imagem original. Essas medidas mostram uma diferença em relação à qualidade percebida por observadores humanos, por consistirem de uma comparação matemática entre o sinal original e o sinal processado. Desta forma, um quadro deslocado de um *pixel* pode apresentar uma PSNR baixa, mas um observador não nota a diferença entre a imagem deslocada e a original.

4.1 Métricas de Avaliação Objetiva

A avaliação objetiva consiste em empregar modelos matemáticos para estimar a qualidade de uma sequência de vídeo ou de uma imagem automaticamente, isto é, sem necessidade de avaliadores e de critérios subjetivos.

Para a análise de vídeo decodificado, pode-se distinguir as métricas de dados, que medem a fidelidade do sinal sem considerar o seu conteúdo e métricas de imagens, que tratam os dados de vídeo de acordo com a informação visual que ele contém. Para vídeos comprimidos e transmitidos em redes de pacotes, existe a métrica baseada em pacotes ou *bitstream*, que observa as informações do cabeçalho do pacote e do *bitstream* diretamente, sem decodificar totalmente o vídeo (Winkler & Mohandas, 2008).

Além disso, as métricas objetivas podem ser classificadas de acordo com a disponibilidade do sinal original, com o qual os sinais distorcidos serão comparados, sendo portanto denominadas (Fonseca, 2008):

- *Métricas com referência total* – as métricas com referência total, ou *full reference* – FR, baseiam-se na comparação quadro a quadro entre o sinal do vídeo original, também denominado sinal referenciado, e o sinal do vídeo degradado, para exprimir o nível de dispersão da qualidade do vídeo degradado em relação ao vídeo original.

  Essas métricas exigem que o vídeo de referência esteja disponível, de forma perfeita e sem compressão, o que é restritivo em relação à usabilidade prática de tais métricas. Além disso, as métricas de referência total geralmente impõem um alinhamento preciso dos dois vídeos, de modo que cada *pixel* em cada quadro pode ser comparado com o seu homólogo do outro vídeo.

- *Métricas com referência reduzida* – as métricas com referência reduzida, ou *reduced reference* – RR, necessitam apenas de partes do sinal do vídeo original para

comparar com um sinal de um vídeo processado e definir a qualidade do vídeo processado.

- *Métricas sem referência* – as métricas sem referência, ou *no reference* – NR, adotam como base apenas o sinal do vídeo degradado. Por esse fator, essa classe de métricas é a ideal para avaliar a qualidade de vídeos digitais, principalmente em serviços que necessitem de monitoramento da qualidade em tempo real. Entretanto, a sua implementação é complexa, visto que se torna necessário considerar inúmeras características do sistema visual humano.

Estas três classes de métricas também têm diferentes utilizações operacionais. As métricas FR são mais adequadas para a medição da qualidade de vídeos *off-line*, como o ajuste *codec* ou testes de laboratório, nas quais as condições podem ser bem controladas, e que uma análise detalhada e precisa do vídeo é mais importante do que resultados imediatos. As métricas NR e RR são mais adequadas para monitoramento de sistemas em serviço de vídeo, para os quais a medição em tempo real é essencial. As métricas RR ainda requerem um canal de retorno e acesso à referência em algum ponto.

Além dessa classificação, as métricas objetivas podem ser classificadas segundo a sua filosofia de estimação da qualidade. Neste sentido, as principais correntes são as métricas de dados e as métricas de imagem, descritas a seguir.

4.1.1 Métricas de Dados

A comunidade de processamento de imagem e vídeo tem usado o MSE e a PSNR como métricas de fidelidade (matematicamente, a PSNR é apenas uma representação logarítmica de MSE) (Winkler & Mohandas, 2008).

O MSE é calculado a partir do valor médio dos erros quadráticos entre os *pixels* do vídeo original e do vídeo degradado, sendo definido como

$$\mathrm{MSE} = \frac{1}{NYX} \sum_{n=1}^{N} \sum_{y=1}^{Y} \sum_{x=1}^{X} (f(x,y,n) - h(x,y,n))^2, \tag{4.1}$$

em que N é o número total de quadros, Y o número total de linhas e X o número total de colunas no vídeo. Quanto menor o valor do MSE, melhor é a qualidade do vídeo avaliado em relação ao original. O valor de MSE igual a zero significa que as amostras de vídeo submetidas à avaliação são idênticas.

A PSNR é definida como

$$\text{PSNR} \;=\; 10\;\log_{10}\left[\frac{255^2}{\text{MSE}}\right], \tag{4.2}$$

para vídeos codificados com 8 *bits*. Quanto maior o valor da PSNR, mais próximo é o vídeo degradado do vídeo original.

Há várias razões para a popularidade destas duas métricas. As fórmulas para seu cálculo são simples de entender, fáceis de implementar e rápido para calcular. Apesar de sua popularidade, a PSNR tem apenas uma relação aproximada com a qualidade do vídeo percebido por observadores humanos. Isso ocorre porque ela se baseia em uma comparação *byte* a *byte* dos dados, sem considerar o que eles realmente representam. Assim, a medida fornecida por essas métricas não apresenta uma boa correlação com a qualidade realmente percebida (Zhou Wang & Bovik, 2004).

No exemplo apresentado na Figura 4.1, as imagens têm o mesmo PSNR, mas a sua qualidade percebida é muito diferente. Não se percebe erro na Figura 4.1a, ao passo que as distorções são bastante evidentes na Figura 4.1b. Há duas razões principais para esta discrepância, sendo que ambos estão intimamente ligados à forma como o sistema visual humano processa as informações:

(a) (b)

Figura 4.1: Ilustração da influência do tipo de deficiência e do conteúdo da imagem na visibilidade das distorções. Ambas as imagens têm PSNR idêntica, mas a sua qualidade percebida é muito diferente (Winkler & Mohandas, 2008).

- *Métricas de dados baseadas na distorção.* As distorções podem ser mais ou menos evidentes para o espectador, dependendo do tipo e das propriedades. O sistema visual humano não é sensível ao ruído de alta frequência inserido na imagem

esquerda (Figura 4.1a). O ruído na imagem à direita (Figura 4.1b) é bem localizada, tem menor frequência de ruído, na qual o padrão é muito mais aparente.

- *Métricas de dados baseadas no conteúdo.* A percepção do visualizador varia de acordo com a parte da imagem ou do vídeo, na qual a distorção ocorre. O ruído na imagem da esquerda está contido quase exclusivamente na região inferior da imagem, na qual tem-se vários conteúdos (bordas, textura das rochas e mar). O conteúdo da imagem mascara a distorção nesta região. O ruído na imagem direita está contida em uma região desprovida de conteúdo (céu liso), destacando as distorções.

Várias métricas de dados adicionais foram propostas e avaliadas. Embora algumas destas métricas possam prever classificações subjetivas com bastante sucesso, para uma dada distorção, ou conteúdo da cena, elas não são de confiança para as avaliações em todas as técnicas. O MSE foi proposto para ser uma métrica precisa para ruído aditivo, mas é superado em vídeos com artefatos de codificação por métricas de qualidade baseadas na visão (Avcibas *et al.*, 2002).

4.1.2 Métricas baseadas na Imagem

Devido aos problemas com as métricas de dados, muito esforço tem sido realizado em projetar melhores métricas de qualidade visual, de forma que representem os efeitos de distorções e o conteúdo sobre a qualidade percebida. As abordagens das métricas podem ser classificadas em dois grupos, uma abordagem de acordo com o modelo visual e uma abordagem baseada em engenharia (Winkler, 2005).

A abordagem de acordo com o modelo visual, como o nome implica, baseia-se na modelagem de vários componentes do sistema visual humano (HVS). As métricas baseadas no HVS tentam incorporar os aspectos da visão humana considerados relevantes para a qualidade da imagem, como percepção de cores, sensibilidade ao contraste e máscara padrão, usando modelos e dados de experimentos psicofísicos (Zhou Wang & Bovik, 2004).

Para a engenharia, por outro lado, a abordagem é baseada principalmente na extração e análise de determinadas características ou artefatos no vídeo. Eles podem ser elementos de imagem estruturais, tais como contornos, ou distorções específicas que são introduzidas por um passo do processamento de vídeo, compressão ou transmissão.

Isso não significa necessariamente que tais métricas desconsideram a visão humana, uma vez que muitas vezes consideram os efeitos psicofísicos, mas também o conteúdo e a análise da distorção na imagem.

Essa abordagem ganhou popularidade nos últimos anos com o índice de similaridade estrutural (SSIM – *Strucutural Similarity Index*), que calcula a média, a variância e a covariância de pequenas partes de um quadro e combina as medições em um mapa de distorção (Zhou Wang & Bovik, 2004). Uma outra métrica importante, a VQM, divide as sequências em blocos espaço-temporais, e um número de características da medição da quantidade e da orientação de atividade, em cada um destes blocos é calculado a partir do gradiente de luminância espacial (Pinson & Wolf, 2004). As características extraídas do teste em vídeos de referência são então comparadas utilizando um processo semelhante ao de mascaramento.

Índice de Similaridade Estrutural (SSIM)

Um dos modelos mais recentes e consolidados para as métricas objetivas de avaliação da qualidade de vídeos é o SSIM (Zhou Wang & Bovik, 2004). A métrica SSIM foi proposta com base no pressuposto de que o Sistema Visual Humano é adaptado para extrair informações estruturais de imagens.

No método SSIM, os *pixels* possuem forte dependência entre si e ela aumenta consideravelmente de acordo com a proximidade entre eles. Supõe-se que essa dependência carrega informações importantes sobre a estrutura dos objetos na imagem, e que quantificar a mudança estrutural de uma imagem pode fornecer uma boa aproximação para a qualidade percebida (Zhou Wang & Bovik, 2004).

Como consequência dessa nova proposta, foram desenvolvidos muitos algoritmos baseados no SSIM. Entre eles, pode-se citar o *Edge-Based Structural Similarity* (E-SSIM), *Multi-Scale Structural Similarity* (MS-SSIM), *Fast* MS-SSIM, 3-SSIM, *Percentile Pooling* SSIM (P-SSIM) (Wang *et al.*, 2003), Complex-Wavelet SSIM index (CW-SSIM) (Wang & Simoncelli, 2005), *Gradient-based Structural Similarity* (G-SSIM) (Chen *et al.*, 2006b). Cada variação do SSIM considera um tipo de efeito ou comportamento e, com isso, há o aumento do custo computacional em relação às métricas mais simples. Alguns desses algoritmos são apresentados nesta seção.

Funcionamento do SSIM

O SSIM é um algoritmo que utiliza a estatística da imagem para a avaliação da qualidade, ou seja, os atributos que constituem a informação estrutural dos objetos da imagem dependem da média da luminância e do contraste da imagem (Zhou Wang & Bovik, 2004). O funcionamento do SSIM inicia com a divisão da imagem em blocos $m \times m$ e então são calculados, para cada bloco, a média (μ), o desvio padrão (σ^2) e a covariância (σ_{fh}). A média e o desvio padrão são considerados estimativas da luminância e do contraste da imagem, respectivamente. A covariância é a medida de quanto um sinal é diferente do outro (Zhou Wang & Bovik, 2004; Vranjes *et al.*, 2007).

Sejam então dois sinais representados por $f = \{f_i | i = 1, 2, \cdots, P\}$ e $h = \{h_i | i = 1, 2, \cdots, P\}$, as características estatísticas são dadas por:

$$\mu_f = \frac{1}{P}\sum_{i=1}^{P} f_i \quad , \quad \mu_h = \frac{1}{P}\sum_{i=1}^{P} h_i, \tag{4.3}$$

$$\sigma_f^2 = \frac{1}{P-1}\sum_{i=1}^{P}(f_i - \mu_f)^2 \quad , \quad \sigma_h^2 = \frac{1}{P-1}\sum_{i=1}^{P}(h_i - \mu_h)^2, \tag{4.4}$$

$$\sigma_{fh} = \frac{1}{P-1}\sum_{i=1}^{P}(f_i - \mu_f)(h_i - \mu_h). \tag{4.5}$$

Em seguida, por meio da intensidade média de cada imagem, obtém-se a comparação de luminância $l(f, h)$, calculada por

$$l(f, h) = \frac{2\mu_f\mu_h + C_1}{\mu_f^2 + \mu_h^2 + C_1}, \tag{4.6}$$

em que a constante C_1 é incluída para evitar uma divisão por zero, quando o valor $\mu_f^2 + \mu_h^2$ estiver próximo de zero.

Em Zhou Wang & Bovik (2004), foi definido $C_1 = (K_1 L)^2$, em que $L = 2^b - 1$ é o valor máximo de um *pixel* de acordo com o número de *bits*, b e $K_1 \ll 1$.

Analogamente à Equação 4.6, a comparação entre os contrastes é dada por

$$c(f,h) = \frac{2\sigma_f\sigma_h + C_2}{\sigma_f^2 + \sigma_h^2 + C_2}, \tag{4.7}$$

em que $C_2 = (K_2L)^2$ e $K_2 \ll 1$. Os valores de K_1 e K_2 utilizados na literatura são, $K_1 = 0,01$ e $K_2 = 0,03$ (Wang *et al.*, 2004).

A comparação entre as estruturas é definida por

$$s(f,h) = \frac{\sigma_{fh} + C_3}{\sigma_f\sigma_h + C_3}. \tag{4.8}$$

A combinação das três comparações resulta no índice de similaridade estrutural

$$\text{SSIM}(f,h) = [l(f,h)]^\alpha \cdot [c(f,h)]^\beta \cdot [s(f,h)]^\gamma, \tag{4.9}$$

em que α, β e γ são parâmetros utilizados para ajustar a importância de cada componente. Para simplificar, considerou-se $\alpha = \beta = \gamma = 1$ e $C_3 = \dfrac{C_2}{2}$ (Zhou Wang & Bovik, 2004).

Assim, a Equação 4.9 pode ser re-escrita como

$$\text{SSIM}(f,h) = \frac{(2\mu_f\mu_h + C_1)(2\sigma_{fh} + C_2)}{(\mu_f^2 + \mu_h^2 + C_1)(\sigma_f^2 + \sigma_h^2 + C_2)}. \tag{4.10}$$

Esta é a forma sob a qual o descritor SSIM se apresenta na literatura. O algoritmo utiliza a técnica de janelamento, isto é, dividir a imagem em blocos menores e obter os valores do SSIM para cada bloco e, por fim, calcular o SSIM como a média aritmética do SSIM de cada bloco. Por exemplo, para duas imagens divididas em B blocos cada, o índice pode ser calculado por

$$\text{SSIM}(f,h) = \frac{1}{BN}\sum_{n=1}^{N}\sum_{i=1}^{B}\text{SSIM}(f(i,n),h(i,n)), \tag{4.11}$$

na qual i representa os *pixels*, de coordenadas x, y, de um bloco na imagem.

O SSIM tem as seguintes propriedades:

1. SSIM (f,h) = SSIM (h,f);

2. SSIM $(f, h) \leq 1$;

3. SSIM $(f, h) = 1$, se e somente se $f = h$.

Edge-Based Structural Similarity (E-SSIM)

A métrica E-SSIM foi concebida com a constatação de que o SSIM possui falhas ao mensurar a qualidade de imagens que estão contaminadas com o artefato borramento (Chen *et al.*, 2006a). Além disso, o E-SSIM pressupõe que a observação do olho humano é bastante sensível às informações de borda da imagem. As bordas são estruturas definidas por descontinuidades no nível de cinza dos *pixels* de uma imagem, isto é, são fronteiras entre duas regiões com níveis de cinza relativamente distintos (Gonzalez & Woods, 2006).

A partir destas observações, foi desenvolvido em Chen *et al.* (2006a) o E-SSIM, que modifica a componente de comparação de estrutura $s(f, h)$, na Equação 4.8, pela componente de comparação de borda $e(f, h)$.

Para a obtenção das informações de borda, foi adotado o operador de Sobel e o seu resultado é apresentado na Figura 4.2 (Chen *et al.*, 2006a).

(a) Quadro original (b) Quadro após a convolução

Figura 4.2: Exemplo da aplicação do operador de Sobel. A Figura 4.2a apresenta o quadro 1 do vídeo Foreman. A Figura 4.2b apresenta o mesmo quadro do vídeo após a convolução com o operador de Sobel.

Para cada *pixel* na posição (x, y), é associado um vetor $D_{x,y}\left(\frac{\partial f}{\partial y}, \frac{\partial f}{\partial x}\right)$ e, por meio das máscaras verticais e horizontais, são encontrados $\frac{\partial f}{\partial y}$ e $\frac{\partial f}{\partial x}$. O vetor de borda é representado por sua amplitude e direção. A amplitude pode ser estimada por

$$D_{x,y} = \left|\frac{\partial f}{\partial y}\right| + \left|\frac{\partial f}{\partial x}\right|, \tag{4.12}$$

e a direção do vetor é dada por

$$\theta = \frac{\partial f/\partial y}{\partial f/\partial x}. \tag{4.13}$$

Com isso, a cada *pixel* de borda identificado pelo operador de Sobel é associado um vetor com direção e amplitude. Esse processo é realizado para cada bloco de uma imagem, constituindo uma imagem com as direções dos vetores de cada *pixel* de borda.

Sejam $f(x, y)$ e $h(x, y)$ as imagens com as direções dos vetores de borda da imagem original e degradada, respectivamente. A comparação entre as bordas $e(f, h)$ é calculada pelo coeficiente de correlação entre $f(x, y)$ e $h(x, y)$ em cada bloco $m \times m$,

$$e(f, h) = \frac{\sigma^{'}_{fh} + C_4}{\sigma^{'}_{f}\sigma^{'}_{h} + C_4}, \tag{4.14}$$

em que C_4 é uma constante, com valor pequeno, para evitar uma divisão por zero, no cálculo caso o denominador se aproxime de zero, $\sigma^{'}_{h}$, $\sigma^{'}_{f}$ e $\sigma^{'}_{fh}$ são os desvios padrões e a covariância dos vetores de borda das imagens $f(x, y)$ e $h(x, y)$, respectivamente.

Assim, o E-SSIM relaciona todas as propriedades do SSIM, no entanto substitui o elemento de comparação da estrutura $s(f, h)$ pela comparação da borda $e(f, h)$. Sendo então calculado da mesma forma do SSIM, pela equação,

$$\text{E-SSIM}(f, h) = [l(f, h)]^{\alpha} \cdot [c(f, h)]^{\beta} \cdot [e(f, h)]^{\gamma}. \tag{4.15}$$

Na prática, calcula-se a média do E-SSIM para B blocos $m \times m$ em um vídeo com N quadros por

$$\text{MESSIM}(f, h) = \frac{1}{N \cdot B} \sum_{n=1}^{N} \sum_{i=i}^{B} \text{E-SSIM}(\mathbf{f}(i, n), \mathbf{h}(i, n)). \tag{4.16}$$

Multi-Scale Structural Similarity (MS-SSIM)

O método MS-SSIM é uma maneira de incluir os detalhes dos vídeos, presentes em diferentes situações, em uma abordagem objetiva de avaliação da qualidade (Wang *et al.*, 2003).

A partir do vídeo de referência e do vídeo distorcido, o sistema aplica um filtro

passa-baixa e reduz a resolução espacial por um fator de dois em cada iteração. As comparações de contraste $c_j(f,h)$ e estrutura $s_j(f,h)$ são calculadas até a j-ésima iteração e a comparação de luminância $l_M(f,h)$ é feita apenas na iteração de número M, isto é, M define o tamanho da escala. Esse procedimento é definido por

$$\text{MS-SSIM}(f,h) = [l_M(f,h)]^{\alpha_M} \cdot \prod_{j=1}^{M} [c_j(f,h)]^{\beta_j} [s_j(f,h)]^{\gamma_j}. \tag{4.17}$$

De modo semelhante à Equação 4.9, os expoentes α_M, β_j e γ_j são utilizados para ajustar a importância relativa dos diferentes componentes. O MS-SSIM também satisfaz as três propriedades do SSIM.

3-SSIM – *Three component Structural Similarity Index*

Atualmente, os algoritmos de avaliação da qualidade de vídeos existentes quase sempre desconsideram o conteúdo do vídeo (Li & Bovik, 2010). Entende-se por conteúdo, os objetos, contornos e a textura do vídeo.

Com o intuito de conseguir índices de qualidade mais próximos dos valores obtidos por meio de técnicas subjetivas, foi proposto o algoritmo denominado 3-SSIM, que tem por base a segmentação dos *pixels* em regiões e, além disso, as regiões significam o conteúdo do vídeo (Li & Bovik, 2010). Essas regiões são classificadas como sendo de contorno, textura e suavidade.

Em geral, o cálculo do índice 3-SSIM, segue as etapas:

1. Calcular o índice SSIM;

2. Segmentar a imagem original e a distorcida em três regiões: contornos, textura e suavidade;

   Esta segmentação é feita com a utilização do operador de Sobel, que calcula um gradiente máximo (g_{max}), que é o maior valor das amplitudes dos vetores de borda encontrados por este operador.

   A partir do gradiente são definidos dois limites, que foram determinados em Li & Bovik (2010):

   - $TH_1 = 0,12 \cdot g_{\max}$
   - $TH_2 = 0,06 \cdot g_{\max}$

Sejam $p_f(x, y)$ e $p_h(x, y)$ os gradientes da coordenada (x, y) da imagem original e degradada, respectivamente, a classificação dos *pixels* é feita da seguinte forma:

(a) Se $p_f(x, y) > TH_1$ ou $p_h(x, y) > TH_1$, então o *pixel* é considerado de contorno.

(b) Se $p_f(x, y) < TH_2$ e $p_h(x, y) \leq TH_1$, então o *pixel* é considerado da região suave.

(c) Caso nenhuma das condições citadas seja satisfeita, então o *pixel* é considerado parte da região de textura.

3. Após a classificação dos *pixels*, deve-se considerar diferentes valores de importância para os índices SSIM nas três regiões;

4. Por fim, calcular a média ponderada dos índices SSIM para formar um único índice de qualidade.

G-SSIM

Estudos do SSIM verificaram que a métrica falha em avaliar imagens desfocadas. Uma forma de melhorar a avaliação da qualidade é utilizar as informações de borda como o dado mais importante da estrutura da imagem. Essa métrica é chamada *Gradient-based Structural Similarity* (G-SSIM) (Chen *et al.*, 2006b).

O olho humano é muito sensível à informação de borda e de contorno de uma imagem, isto é, a informação de borda e de contorno pode ser a informação mais importante da estrutura de uma imagem da qual o ser humano captura a cena. Com base nisso, o G-SSIM foi desenvolvido com o intuito de melhorar o SSIM, comparando a informação de borda entre o bloco da imagem distorcida e a original, e substituindo a comparação de contraste $c(f, h)$ e de estrutura $s(f, h)$ das Equações 4.7 e 4.8, respectivamente, pela comparação do gradiente do contraste $c_g(f, h)$ e da estrutura $s_g(f, h)$,

$$c_g(f, h) = \frac{2\sigma_{f'}\sigma_{h'} + C_2}{\sigma_{f'}^2 + \sigma_{h'}^2 + C_5}, \tag{4.18}$$

$$s(f, h) = \frac{\sigma_{f'h'} + C_3}{\sigma_{f'}\sigma_{h'} + C_6}. \tag{4.19}$$

na qual $f^{'}$ e $h^{'}$ são as imagens após a operação do gradiente, $\sigma_{f'}$ e $\sigma_{h'}$ são os desvios padrões do vetor $f^{'}$ e $h^{'}$, respectivamente, $\sigma_{f'h'}$ é vetor covariância do vetor $f^{'}$ e $h^{'}$, e C_5 e C_6 são constantes. Assim o *Gradient-based Structural Similarity* é dado por

$$\text{G-SSIM}(f,h) = [l(f,h)]^{\alpha} \cdot [c_g(f,h)]^{\beta} \cdot [s_g(f,h)]^{\gamma}, \tag{4.20}$$

A semelhança estrutura global da imagem (G-SSIM) é calculada como a média dos blocos (B) em todos os quadros (N).

$$\text{G-SSIM}(f,h) = \frac{1}{BN} \sum_{n=1}^{N} \sum_{i=1}^{B} \text{GSSIM}(f(i,n), h(i,n)). \tag{4.21}$$

4.1.3 Métricas para vídeos 3D

Com o crescente uso de imagens estereoscópicas há a necessidade de se avaliar a qualidade de vídeos 3D, para melhorar a satisfação do usuário e a eficiência de uso dos recursos. A qualidade do efeito 3D em vídeos estereoscópicos pode ser avaliada levando em consideração a qualidade dos mapas de profundidade da imagem.

Os métodos de avaliação da qualidade de imagem 2D são pouco adequados para medir a qualidade da imagem em 3D, uma vez que a profundidade (o fator mais importante em um sistema 3D) e as distorções típicas da estereoscopia (por exemplo, *crosstalk*) não são incorporadas (Meesters *et al.*, 2004). Assim, em comparação com o vídeo 2D, a avaliação objetiva de vídeo 3D é mais complexa, uma vez que, a opinião do observador pode ser considerada multidimensional, incluindo fatores como a fadiga visual e a percepção de profundidade, além de aspectos do HVS que precisam ser abordados, por exemplo, a supressão binocular[1].

Na cadeia de transmissão em 3D, artefatos visíveis ocorrem em vários locais. Na captura da câmera ou nas etapas de conversão e renderização tem-se a introdução das degradações geométricas. Além dos artefatos, o conteúdo do vídeo em si tem um impacto maior sobre a qualidade visual percebida em um vídeo 3D do que em um vídeo 2D (Huynh-Thu *et al.*, 2010).

[1] A supressão binocular indica que um ponto de vista do par estéreo pode ser transmitido com uma qualidade visual pior do que outro. (Ozbek & Tekalp, 2008)

Até agora, poucos métodos objetivos de avaliação de imagens estéreo foram apresentados, que usam o mapa de profundidade para avaliar o sentido estéreo. Um algoritmo de avaliação objetiva, que utiliza o mapa de profundidade, bem como a visão estereoscópica, é proposto em Shao *et al.* (2009). Ele inclui as partes do índice de similaridade estrutural (SSIM) e a detecção de degradações de borda e cor.

A aplicabilidade da PSNR e de modelos de vídeo 2D (SSIM e VQM) para o 3D foi investigada em um conjunto de dados pequeno, tanto para o caso de vídeo estereoscópico, quanto para vídeo com informações de profundidade monoscópica. Os resultados mostram que a qualidade de vídeo 3D pode ser estimada a partir da avaliação separada de cada vista, enquanto modelos para 2D também poderiam ser usados para estimar a qualidade de percepção de profundidade (Yasakethu *et al.*, 2008).

Para o desenvolvimento de novas métricas, pesquisas têm sido realizadas em busca de medidas objetivas que possam realizar a avaliação para vídeos em 3D. Uma dessas pesquisas sugere que os observadores não percebem uma mudança de profundidade até um certo nível (Silva *et al.*, 2010). Assim, as mudanças de *pixels* na imagem de profundidade não causam distorções na percepção de profundidade, a menos que essas mudanças excedam um limite particular. Este limiar corresponde ao valor JNDD (*Just Noticeable Difference in Depth*), medido em unidades de níveis de profundidade.

Nesta seção são apresentadas algumas métricas de avaliação de vídeo 3D. As métricas foram escolhidas por terem baixa complexidade computacional. Aquelas que usam mapas de profundidade mais complexos têm um custo computacional elevado e assim não serão abordadas.

Avaliação de qualidade da imagem e do sentido estéreo

A avaliação segue o método da PSNR. E define-se o valor do IQA (*Image Quality Assessment*) como a média aritmética da avaliação da PSNR nas imagens esquerda e direita,

$$\mathrm{IQA} = \frac{(\mathrm{PSNR}_L + \mathrm{PSNR}_R)}{2}. \tag{4.22}$$

A avaliação objetiva do sentido estéreo segue o método da PSNR (Yang *et al.*, 2009). O modelo de avaliação da qualidade da imagem estéreo é realizada da seguinte forma: Obtém-se as disparidades absolutas da imagem original e processada, $d_f(x, y)$ e $d_h(x, y)$ (Equações 4.23 e 4.24).

$$d_f(x,y) = |f_l(x,y) - f_r(x,y)| \tag{4.23}$$

$$d_h(x,y) = |h_l(x,y) - h_r(x,y)| \tag{4.24}$$

Depois remove-se os ruídos suaves e os sinais com baixa magnitude a partir da disparidade absoluta da imagem original, $d_f(x,y)$, de modo a diminuir a interferência. Na visão estéreo, a disparidade com baixa magnitude pode ser suprimida ou aumentada. Devido a isso, o sinal de baixa magnitude não pode servir como avaliação. Então pode-se calcular a condição de distribuição da disparidade da imagem original (Yang *et al.*, 2009).

Em terceiro lugar, combina-se a condição de distribuição de disparidade da imagem original $d_f(x,y)$ e $d_h(x,y)$, e é calculada a SSA na posição da disparidade dos dois olhos, dada por:

$$\mathrm{SSA} = 10\log_{10}\frac{(2^b-1)^2}{\mathrm{MSE}_d}, \tag{4.25}$$

$$\mathrm{MSE}_d = \frac{1}{NYX}\sum_{n=1}^{N}\sum_{y=1}^{Y}\sum_{x=1}^{X}(d_f(x,y,n) - d_h(x,y,n))^2. \tag{4.26}$$

Os resultados apresentados em Yang *et al.* (2009), concluem que o IQA é adequado para a avaliação da qualidade da imagem e o SSA é adequado para a qualidade estéreo sentida, então a combinação das duas métricas pode ser definida como o modelo de avaliação.

Métrica da Qualidade Percebida

A métrica VQM considera uma ampla gama de características perceptivas tais como borramento, movimento não natural, ruído, distorção de cor e distorção de bloco. O VQM correlaciona-se com o HVS e pode ser usado para determinar objetivamente a qualidade dos vídeos em 3D pela média das imagens 2D esquerda e direita.

No entanto, uma métrica atualmente desenvolvida, a PQM (*Perceived Quality Metric*), apresenta melhores resultados para a avaliação de vídeo 3D de qualidade e supera a métrica VQM, pois é sensível a pequenas mudanças na degradação da imagem e a

quantificação de erro no nível do *pixel* (Joveluro *et al.*, 2010).

A PQM quantifica a distorção na luminosidade e a distorção de contraste usando uma aproximação (variância) ponderada pela média de cada bloco de *pixels* para obter a distorção de uma imagem. A componente de luminância só é considerada porque o HVS é mais sensível à componente de luminância (forma a estrutura da imagem) do que à componente de crominância.

$$\delta_{(x,y)}(i) = \begin{cases} 0, & \mu_f(i) \leq 1, \quad \mu_h(i) \leq 1 \\ 1, & \mu_f(i) \leq 1, \quad \mu_h(i) > 1 \\ \left(\dfrac{(f(x,y) - h(x,y))^2}{\mu_f(i)}\right)^2, & \text{caso contrário.} \end{cases} \tag{4.27}$$

em que $\delta_{(x,y)}$ é a distorção entre a luminância ao nível de *pixel* no bloco i, $\mu_f(i)$ e $\mu_h(i)$ são as médias da luminância do bloco i, no bloco original e no distorcido. A distorção mínima ocorre se a média de $f(x,y)$ e $h(x,y)$ for menor do que 1, porque a distorção no bloco torna-se desprezível.

A função que considera a distorção no contraste é dada por

$$C(i) = 1 + \frac{(\sigma_f^2 + \sigma_h^2)^2 + K}{(\sigma_f^4) + (\sigma_h^4) - 2(\sigma_{fh})^2 + K}, \tag{4.28}$$

na qual K é uma constante de valor 255.

A distorção perceptual no nível do bloco (PDM(i)) é calculada como

$$\text{PDM}(i) = C(i) \cdot \frac{\sum_{y=1}^{m} \sum_{x=1}^{m} \delta_{(x,y)}(i)}{mm}. \tag{4.29}$$

O PDM(i) é a distorção no nível de bloco, e é ponderado pelo fator $W(i)$ para obter o nível do quadro da Distorção Perceptual. O fator de ponderação é calculado como segue

$$W(i) = \begin{cases} 1, & \mu_f(t) = 0, \\ \dfrac{255}{\mu_f(i)}, & \text{caso contrário.} \end{cases} \tag{4.30}$$

Para obter a PQM, é subtraído de 1 o valor de ponderação realizada no PDM(i) e os valores menores que 0 são equiparados a zero como a gama da métrica é entre 0 e 1, representando as qualidades piores e melhores respectivamente. A PQM é calculada por

$$\text{PQM} = \begin{cases} 0, & \text{PDM}(i) < 0 \\ 1 - \left(\dfrac{\sum_{i=1}^{V} W(i)\text{PDM}(i)}{\sum_{i=V}^{V} W(i)} \right), & \text{caso contrário.} \end{cases} \tag{4.31}$$

4.2 Métricas de Avaliação Subjetiva

A avaliação subjetiva consiste em quantificar a qualidade de uma sequência de vídeo por meio da visualização por um observador humano, que atribui uma nota de acordo com o seu critério de qualidade.

Esse modelo é considerado ideal para atribuir um valor de qualidade a uma amostra de vídeo, entretanto, em sua metodologia tem desvantagens que não estão presentes em métodos objetivos (Reckwerdt, 2012).

Pode-se citar como principal desvantagem das avaliações subjetivas a demanda de tempo para obtenção e aplicação dos resultados adquiridos. Além desse fator, o custo com os procedimentos de preparação de ambiente, a execução dos testes e a requisição de recursos humanos são desvantagens significativas inerentes às metodologias subjetivas.

Contudo, os resultados de testes subjetivos são fundamentais para a validação de propostas de métricas objetivas, visto que é por meio da correlação entre esses resultados que se pode afirmar o quanto uma métrica objetiva consegue prever a qualidade percebida por um grupo de avaliadores.

Em experiências subjetivas, um grupo de avaliadores (tipicamente 15 a 30) são

convidados a assistir a um conjunto de vídeos com diferentes níveis de qualidade. A média de todas as notas dos espectadores de um vídeo é conhecida como MOS (*Mean Opinion Score*).

Uma vez que os indivíduos têm diferentes interesses e expectativas para o vídeo, a subjetividade e a variabilidade dos índices de audiência não podem ser completamente eliminados. Experiências subjetivas tentam minimizar esses fatores por meio de instruções precisas, treinamento e ambientes controlados. No entanto, é importante lembrar que um índice de qualidade é uma medida aleatória que é definida por uma distribuição estatística, em vez de um número exato.

Há uma grande variedade de métodos de ensaio subjetivos. A ITU (*International Telecommunications Union*) formalizou métodos em diversas recomendações (ITU-R, 2002), (ITU-T, 1999). Os métodos sugerem condições de visualização padrão, critérios de seleção dos observadores e dos materiais de teste, procedimentos de avaliação e métodos de análise de dados. Procedimentos de teste recomendados incluem comparações implícitas, como DSCQS (*Double Stimulus Continuous Quality Scale*), comparações explícitas, como DSIS (*Double Stimulus Impairment Scale*) ou classificações absolutas, como SSCQE (*Single Stimulus Continuous Quality Evaluation*) ou ACR (*Absolute Category Rating*). O procedimento utilizado para uma dada experiência é geralmente selecionado como uma função da aplicação, do intervalo de qualidade, e das tarefas dos telespectadores.

A Recomendação ITU-R BT. 500-11 trata das metodologias subjetivas para avaliação da qualidade de imagens e vídeos em televisão. A Recomendação ITU-T P.910, por sua vez, traz resoluções e orientações sobre os métodos subjetivos de avaliação da qualidade em aplicações multimídia.

Entre as principais metodologias previstas nessas recomendações, podem-se citar os métodos ACR (*Absolute Categorical Rating*), DCR (*Degradation Category Rating*), PC (*Pair Comparison*), DSCQS (*Double-Stimulus Continuous Quality-Scale*) e SS (*Single-Stimulus*).

4.2.1 Método ACR – *Absolute Categorical Rating*

O método ACR é uma metodologia de estímulo único, isto é, as amostras de vídeo degradadas são apresentadas uma por vez e, após cada apresentação, o avaliador atribui uma nota, segundo uma escala discreta (Tabela 4.1), à amostra sob teste. Essa nota

representa o nível de satisfação do avaliador em relação ao serviço visualizado.

Tabela 4.1: Escala discreta de votação da metodologia ACR.

5	Excelente
4	Bom
3	Regular
2	Ruim
1	Péssimo

A Recomendação ITU-T P.910 (ITU-T, 1999) prevê que, caso haja necessidade de uma maior discriminação em relação à qualidade atribuída pelos avaliadores, pode-se utilizar uma escala com nove ou onze níveis de votação.

4.2.2 Método DCR – *Degradation Category Rating*

Diferentemente da metodologia ACR, o modelo empregado no método DCR é de estímulo duplo, isto é, as amostras são apresentadas em pares, uma por vez, sendo a primeira amostra sempre a original e a segunda amostra a sob teste. Após a visualização de cada par, o avaliador atribui uma nota, segundo uma escala de níveis discretos, que corresponde ao nível de degradação percebido na amostra sob teste.

Tabela 4.2: Níveis de classificação da metodologia DCR.

5	Imperceptível
4	Perceptível, mas não incômodo
3	Levemente incômodo
2	Incômodo
1	Muito incômodo

A Recomendação ITU-T P.910 (ITU-T, 1999) prevê para o método DCR que, caso as amostras possuam resolução espacial reduzida (QCIF, CIF ou SIF, por exemplo), pode-se reproduzir o vídeo de referência e o vídeo sob teste simultaneamente no mesmo monitor.

4.2.3 Método PC – *Pair Comparison*

No método PC, as sequências de teste de uma mesma cena, em diferentes condições de degradação, são exibidas em pares em todas as $c(c-1)$ combinações, dessa forma todos os pares de sequências devem ser exibidos nas duas ordens possíveis (AB e BA, por

exemplo). Após a apresentação de cada par, os avaliadores definem a sua preferência em relação a uma das sequências apresentadas. Essa metologia permite uma discriminação de qualidade muito eficiente entre as sequências.

Da mesma forma do método DCR, a Recomendação ITU-T P.910 (ITU-T, 1999) prevê que, caso as amostras possuam resolução espacial reduzida (QCIF, CIF ou SIF, por exemplo), pode-se reproduzir o vídeo de referência e o vídeo sob teste simultaneamente no mesmo monitor.

4.2.4 Método ACR-HR – *Absolute Category Rating with Hidden Reference*

A metodologia ACR é classificada como de estímulo único, na qual as amostras de vídeo são apresentadas uma por vez e, após o término de cada sequência, o avaliador atribui uma nota segundo uma escala discreta, como a apresentada na Tabela 4.1, à amostra de vídeo sob teste. A nota atribuída representa o nível de satisfação do avaliador em relação à qualidade da amostra visualizada.

O método ACR com referência oculta (ACR-HR), por sua vez, apresenta a sequência de vídeo de referência para ser avaliada como qualquer outra amostra, isto é, sem que exista o conhecimento do avaliador sobre a condição da amostra. Deste modo, a diferença entre os escores médio de opinião de uma amostra processada, e sua correspondente amostra de referência é calculado como

$$\mathrm{DMOS}_k = u_k - \hat{u}_k, \tag{4.32}$$

em que u_k e $\hat{u}_k$ são as notas referentes às k-ésimas amostras de referência e processada, respectivamente. Na Fórmula 4.32, um valor de DMOS_k próximo a zero indica que a qualidade da amostra processada é 'Excelente', enquanto um valor de DMOS_k próximo a cinco indica que a qualidade da amostra processada é 'Péssima'.

4.2.5 Montagem Experimental

O ambiente que se deseja utilizar para realização da avaliação subjetiva dos vídeos precisa ser devidamente preparado. Fatores externos, tais como, iluminação, distância da tela e posição do observador, acabam por interferir no resultado final. Os critérios para preparação do ambiente normatizados pela ITU-R BT.500, para a utilização estão

a seguir.

- **Relação entre contraste e luminância** – Instrumentos ópticos adicionais para a visualização em 3D, por exemplo, óculos e filtros, causam uma redução de luminosidade. Desta forma, é sugerido que a luminância mínima para *displays* 3D deve ser de pelo menos 30 cd/m^2 para sustentar a profundidade do foco e garantir a sensação de profundidade básica.

- **Iluminação ambiente** – A profundidade de fundo real e a profundidade percebida podem levar a conflitos se a tela está muito perto da parede. Neste caso, os objetos podem parecer dentro da parede. Além disso, a iluminação do ambiente pode precisar ser ajustada para melhor visualização 3D. Por exemplo, uma fonte de iluminação neon, possivelmente, pode causar cintilação grave, induzindo desconforto visual para os espectadores.

- **Resolução do monitor** – A resolução do monitor em geral e a resolução estereoscópica devem ser consideradas como aspectos da resolução do monitor. *Displays* 3D espacialmente multiplexados podem não possuir uma distribuição uniforme dos *pixels*, o que compromete a correta exibição do vídeo. Além disso, técnicas de multiplexação no tempo mostram que a resolução espacial completa pode ser mantida, mas as técnicas de multiplexação temporais degradam a visão devido a assimetrias temporais e da distribuição de luminosidade temporal.

- **Distância de visualização** – Três vezes a altura da tela para HDTV e seis vezes para SDTV foram as distâncias aprovadas na recomendação nos padrões ITU BT.710 e BT.500. Além disso, a distância de visualização preferida (*Prefered Viewing Distance*, PVD) foi recomendada na norma BT-500 para a visualização 2D em ambientes domésticos. A avaliação subjetiva mostra que PVD é uma função de diferentes parâmetros, como acuidade visual humana, a quantidade de movimento, tamanho da resolução da tela de imagem, etc. Como explicado em Patterson (2007), a disparidade binocular deve ser escalada visualmente pela distância de visualização para que a portanto percepção de profundidade binocular possa ocorrer, a percepção de profundidade deve ser adicionada como uma nova componente para a função de PVD.

- **Posição de visualização** – Distorções na geometria 3D, por exemplo, distorções que são causadas por um movimento lateral do observador (cisalhamento)

influenciam a posição de visualização (Woods *et al.*, 1993). A redução da luminosidade torna-se mais grave quando o ângulo de observação aumenta. Isto também aplica-se à paralaxe do movimento que é visto em *multiview* (múltiplas visões) de exibição auto-estereoscópica. O ângulo de visão se relaciona com a correta percepção da imagem do olho esquerdo e direito.

4.3 Métricas Estatísticas

As opiniões a respeito das cenas originais e processadas resultam em um par de notas. A nota média e o desvio padrão são calculados para cada nota, resultando em uma variável denominada nota média na opinião dos observadores, *Mean Opinion Score* (MOS). A média, o desvio padrão e o coeficiente de variação (CV) são medidas estatísticas. A média é dada por

$$\mu_\alpha = \frac{1}{A}\sum_{i=1}^{A}\alpha_i, \tag{4.33}$$

em que α_i são as amostras e A o número total de amostras. O desvio padrão, que é uma medida de dispersão, é dado por

$$\sigma = \sqrt{\frac{1}{A-1}\sum_{i=1}^{A}(\alpha_i - \mu_\alpha)^2}. \tag{4.34}$$

O CV é uma média relativa à dispersão, útil para comparação e observação em termos relativos do grau de concentração em torno da média de séries distintas. É independente da unidade de medida utilizada, e é usado quando se deseja comparar a variação de conjuntos de observações que diferem na média ou são medidos em grandezas diferentes (unidades de medição diferentes). O CV é o desvio padrão expresso como uma porcentagem média,

$$\mathrm{CV} = 100\left(\frac{\sigma}{\mu_\alpha}\right)(\%). \tag{4.35}$$

Com o CV pode-se classificar a dispersão das séries, por (Shiguti & da S. C. Shiguti, 2006):

- Dispersão baixa: CV $\leq 15\%$;

- Dispersão média: $15\% < \text{CV} < 30\%$;
- Dispersão alta: $\text{CV} \geq 30\%$.

4.3.1 Correlação Entre as Medidas

Objetivo do estudo correlacional é a determinação da intensidade do relacionamento entre duas observações emparelhadas. A correlação indica até que ponto os valores de uma variável estão relacionados com os de outra.

As seguintes medidas podem ser utilizadas como indicadores de desempenho dos algoritmos: Coeficiente de Correlação Linear de Pearson (PLCC – *Pearson Linear Correlation Coefficient*), Coeficiente de Correlação dos Postos de Spearman (SROCC – *Spearman Rank-Order Correlation Coefficient*), Coeficiente de Correlação de Kendall (KROCC – *Kendall Rank-Order Correlation Coefficient*) e a Raiz do Erro Médio Quadrático (RMSE – *Root Mean Squared Error*).

O PLCC permite identificar a linearidade entre dois conjuntos de medidas e quanto maior o valor do PLCC, em módulo, mais correlacionados são as medidas. O PLCC entre dois conjuntos de medidas ($A = \{a_k\}$ e $B = \{b_k\} \quad | \quad k = 1, 2, \ldots, K$) é dado por

$$\text{PLCC} = \frac{\sum_{k=1}^{K}(a_k - \mu_a) \cdot (b_k - \mu_b)}{\sqrt{\sum_{k=1}^{K}(a_k - \mu_a)^2} \cdot \sqrt{\sum_{k=1}^{K}(b_k - \mu_b)^2}}, \tag{4.36}$$

em que μ_a e μ_b significam as médias das medidas, respectivamente.

O SROCC indica a monotonicidade entre dois conjuntos ordenados de medidas A e B, sendo matematicamente expressa por

$$\text{SROCC} = 1 - \frac{6 \cdot \sum_{k=1}^{K} {d_k}^2}{K(K-1)} \tag{4.37}$$

em que K é o número de medidas e d_k é a diferença entre os postos das medidas ordenadas $a_k \in A$ e $b_k \in B$.

Assim como o SROCC, o KROCC utiliza os postos das medidas no indicador de desempenho. Especificamente, tomadas aleatoriamente duas amostras, o KROCC pode

ser interpretado como a diferença entre a probabilidade as duas amostras serem consecutivas e a probabilidade de não serem. O KROCC entre dois conjuntos de medidas é calculado como

$$\text{KROCC} = \frac{n_c - n_d}{\frac{1}{2} \cdot K(K-1)}, \tag{4.38}$$

em que n_c é número de pares concordantes e n_d é o número de pares discordantes. Um par de medidas entre os conjuntos A e B, (a_k, b_k) e (a_{k+1}, b_{k+1}), por exemplo, é concordante se $a_k > a_{k+1}$ e $b_k > b_{k+1}$ ou se $a_k < a_{k+1}$ e $b_k < b_{k+1}$, caso contrário o par é discordante.

Os valores limites do coeficiente de correlação, r, são -1 e $+1$. Assim (Roger, 2007; Shmildt, 2007):

a) se não há correlação entre as variáveis, então $r = 0$.

b) se a correlação é muito fraca entre as variáveis, então $0 < |r| < 0{,}2$;

c) se a correlação é fraca entre as variáveis, então $0{,}2 \leq |r| < 0{,}4$;

d) se a correlação é média entre as variáveis, então $0{,}4 \leq |r| < 0{,}7$;

e) se a correlação é forte entre as variáveis, então $0{,}7 \leq |r| < 0{,}9$;

f) se a correlação é muito forte entre as variáveis, então $0{,}9 \leq |r| < 1$;

g) se a correlação entre duas variáveis é perfeita e positiva, então $r = 1$;

h) se a correlação entre duas variáveis é perfeita e negativa, então $r =$ -1.

4.3.2 Construção do Intervalo de Confiança

O intervalo de confiança é calculado a partir da estatística de Fisher, na qual o coeficiente de correlação é convertido em uma variável aleatória z, que segue a distribuição Z de Fisher, que é normal com desvio padrão dado por

$$\sigma_Z = \frac{1}{\sqrt{n_a - 3}}, \tag{4.39}$$

em que n_a é o número de amostras.

O cálculo de z é

$$z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right). \tag{4.40}$$

E o intervalo de confiança (IC) é dado por

$$\mathrm{IC}(z, 1-\alpha) = (z - z_{1-\alpha} \cdot \sigma_Z, z + z_{1-\alpha} \cdot \sigma_Z), \tag{4.41}$$

em que $z_{1-\alpha}$ é um valor tabelado correspondente a uma precisão de $1-\alpha$. Neste trabalho, utilizou-se um intervalo com 95% de confiança, isto é, $\alpha = 0,05$ e $z_{0,95} = 1,96$.

4.4 Considerações Finais

Existem várias métricas de avaliação objetiva. A métrica SSIM é a que apresenta mais variações, como discutido nesse capítulo. Para avaliar vídeos 3D poucas métricas foram propostas.

As métricas de avaliação de vídeo são definidas pelas recomendações da ITU-T e ITU-R. Para aplicações multimídia, existem os métodos ACR, DCR e PC. O método ACR é de estímulo único, o DCR e PC de estímulo duplo.

Para quantificar as métricas subjetivas pode-se usar as medidas estatísticas, como a média, o desvio padrão e o coeficiente de variação, que são utilizadas para verificar o valor da avaliação subjetiva e identificar a confiabilidade da avaliação. Para realizar a comparação entre as métricas subjetivas pode ser utilizado o coeficiente de correlação.

Capítulo 5

Métricas Propostas

Experimentos têm indicado que a atenção visual humana não é igualmente distribuída por toda a imagem, mas somente para algumas regiões em que há maior interesse (Itti & Koch, 2001). Em aplicações de vídeo conferência, por exemplo, a atenção do usuário geralmente está focada na face do indivíduo e não nas regiões do fundo da imagem.

Dessa forma, a atenção visual é uma característica importante do HVS, que pode ser explorada por modelos objetivos de avaliação da qualidade de vídeo e imagem, com a possibilidade de obter índices quantitativos mais correlacionados com a qualidade percebida pelo HVS.

Uma outra forma de analisar o vídeo é por meio da informação da percepção espacial (*Spatial Perceptual Information* – SI), que é uma medida que indica a quantidade de detalhes espaciais percebidos por um observador humano em uma sequência de vídeo (ITU-T, 1999; Webster *et al.*, 1993). Quanto mais complexa for esta sequência, maior será a quantidade de informação espacial.

A estimação da informação da percepção espacial é realizada em cada quadro de uma dada sequência de vídeo por meio da aplicação do operador gradiente de Sobel, que é definido pelas seguintes máscaras:

$$\begin{bmatrix} -1 & 0 & +1 \\ -2 & 0 & +2 \\ -1 & 0 & +1 \end{bmatrix} \qquad \begin{bmatrix} -1 & -2 & -1 \\ 0 & 0 & 0 \\ +1 & +2 & +1 \end{bmatrix}$$

A aplicação do operador de Sobel resulta em uma imagem $G = g(x, y)$ que contém as bordas dos objetos de uma imagem $F = f(x, y)$ e, além disso, G indica a quantidade

de informação espacial presente na imagem (ITU-T, 1999; Gonzalez & Woods, 2006).

As bordas são estruturas definidas por descontinuidades no nível de cinza dos *pixels* de uma imagem, isto é, as bordas definem um limite entre duas regiões com níveis de cinza relativamente distintos.

A aplicação da primeira máscara fornece uma imagem G_y, comumente denominada na literatura como gradiente vertical, que contém as bordas no sentido vertical e a segunda máscara fornece uma imagem G_x denominada gradiente horizontal, que contém as bordas no sentido horizontal.

A imagem resultante, Sobel(F_n), é definida por

$$\text{Sobel}(F_n) = \sqrt{{G_x}^2 + {G_y}^2}. \tag{5.1}$$

A partir da convolução de Sobel, calculam-se os desvios padrões dos níveis de luminância dos *pixels* em cada quadro. O valor máximo entre os desvios padrões é então escolhido para representar o nível de informação espacial da sequência de vídeo. Esse procedimento é representando por

$$\text{SI}(F_n) = \text{std}[\text{Sobel}(F_n)], \tag{5.2}$$

$$\text{SI}(F) = \max\{\text{SI}(\text{F}_\text{n})\}, \tag{5.3}$$

em que std é o conjunto cujos elementos são os desvios padrões associados a cada quadro, Sobel(F_n) é o quadro convolucionado com o operador de Sobel e max é o operador que retorna o elemento de maior valor do conjunto std.

A informação da percepção temporal (*Temporal Perceptual Information* – TI) é uma medida baseada na diferença de luminosidade dos *pixels* localizados em uma mesma posição espacial, mas em quadros subsequentes. A TI indica a quantidade de mudanças espaciais em uma sequência de vídeo, que pode ser imaginada como a taxa de movimentação dos objetos. Logo, esta medida é diretamente proporcional à movimentação da cena.

Para estimar as cenas com grande movimento se faz a diferença da luminância entre dois *pixels* de coordenadas espaciais idênticas e iguais a (x, y) presentes em quadros

subsequentes de um vídeo F e é definido por

$$M_n(x, y) = F_n(x, y) - F_{n-1}(x, y). \tag{5.4}$$

Analogamente à SI, a medida de informação temporal é calculada como o valor máximo entre os desvios padrões sobre a operação $M_n(x, y)$. Esse processo é representado por

$$\mathrm{TI}(F_n) = \mathrm{std}[M_n(x, y)], \tag{5.5}$$

e

$$\mathrm{TI}(F) = \max\{\mathrm{TI}(\mathrm{F_n})\}. \tag{5.6}$$

Os valores das medidas TI e SI são calculados individualmente para cada quadro dos vídeos e, após o cálculo dos dois valores para todos os quadros, o valor máximo de cada sequência é considerado como o valor do TI e do SI do vídeo, respectivamente.

A Figura 5.1 mostra valores de informação espacial e temporal dos vídeos. Próximo ao eixo TI = 0, encontram-se as amostras de vídeo cuja variação de movimento é baixa, como por exemplo: Miss America, Mother Daughter e Akiyo. Por outro lado, as sequências Glasgow, Walk e Highway, apresentam altas taxas de informação temporal. Próximo ao eixo SI = 0, encontram-se as sequências que apresentam pouca complexidade espacial, como por exemplo: Miss America, Suzie e Mother Daughter. As cenas Mobile e Paris, apresentam vários detalhes espaciais.

A partir da informação espacial e temporal e dos experimentos do sistema visual novas métricas são propostas. Essas métricas foram avaliadas e comparadas com os resultados das avaliações subjetivas.

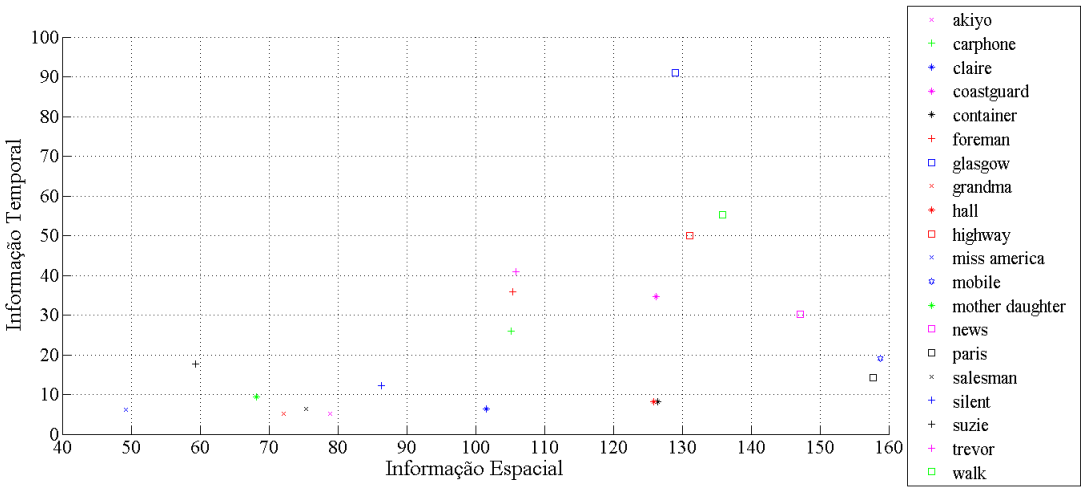


Figura 5.1: Amostras de vídeos agrupadas segundo os valores de SI e TI, vídeo QCIF.

5.1 Avaliação da Qualidade de Vídeo nas Áreas de Interesse

Esta métrica tem como principal característica encontrar as áreas perceptualmente importantes da sequência de vídeo. Para esta tarefa é empregado o operador de Sobel, que estima as principais características que atraem a atenção visual: os contornos de objetos e a quantidade de detalhes espaciais da cena. A medição dessas propriedades determina quais são as áreas de maior interesse.

Após a determinação das bordas, o quadro que contém os vetores de borda é subdividido em V blocos de 8×8 *pixels*. Para cada bloco é calculada a média aritmética da amplitude dos vetores de borda, e em seguida, o algoritmo calcula a maior média entre todos os blocos. Os blocos classificados como de interesse perceptual são aqueles cujas médias (μ_d) obedecem a seguinte condição

$$\mu_d \geq \frac{\mu_{max}}{I}, \tag{5.7}$$

esse processo é realizado para todos os quadros do vídeo. A divisão da média máxima μ_{max} pela constante I é realizada para garantir que somente os blocos relevantes serão selecionados na avaliação objetiva. Neste trabalho, I foi otimizado para a base de dados utilizada, com valor 2,1, que foi obtido após variar o I de 0 a 5, com passo de 0,1, e escolhido a partir dos resultados obtidos da correlação com os resultados da avaliação subjetiva.

Na Figura 5.2 é ilustrado o resultado desse processo. Os quadrados brancos na Figura 5.2c representam os blocos para os quais os modelos objetivos são calculados. O restante da imagem é ignorado.

(a) Quadro original (b) Bordas detectadas pelo operador de Sobel (c) Blocos segmentados

Figura 5.2: Resultado da segmentação realizada em um quadro do vídeo Glasgow.

A partir das regiões de interesse encontradas, a métrica MSE é definida por

$$\mathrm{MSE}_e = \frac{1}{64 \cdot B} \sum_{i=1}^{B} \sum_{x=1}^{8} \sum_{y=1}^{8} (f_{d_{xy}} - h_{d_{xy}})^2, \tag{5.8}$$

em que $f_{d_{xy}}$ e $h_{d_{xy}}$ são as coordenadas espaciais dos *pixels* do i-ésimo bloco.

A PSNR para um sistema de 8 *bits* é calculada por

$$\mathrm{PSNR}_e \;=\; 10 \cdot \log_{10} \left[\frac{255^2}{\mathrm{MSE}_e} \right]. \tag{5.9}$$

Da mesma forma para o SSIM, os valores de média (μ), desvio padrão (σ) e covariância (σ_{fh}) são calculados somente para os blocos selecionados, resultando em

$$\mathrm{SSIM}_e = \frac{1}{B} \sum_{d=1}^{B} \mathrm{SSIM}(f_d, h_d). \tag{5.10}$$

5.2 B-SSIM: *Structural SIMilarity Index for Blurred Videos*

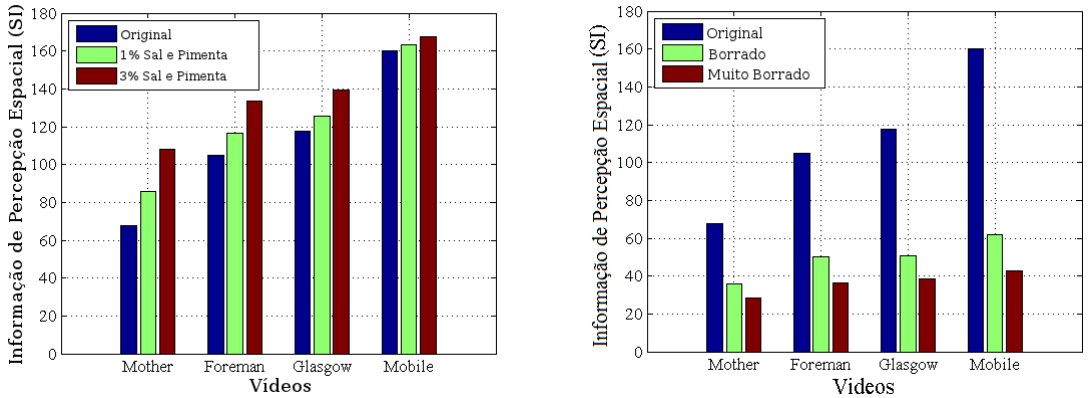
O borramento é apresentado como uma redução da nitidez nos contornos e uma perda de detalhe espacial. Em aplicações reais, esta degradação se deve à exclusão de coeficientes de alta frequência uma vez que no processo de transmissão esses coeficientes

(a) Quadro Original (b) Quadro Borrado 2 (c) Quadro Borrado 4

Figura 5.3: Vídeo Foreman com diferentes níveis de borramento.

podem ser filtrados (Winkler, 2005). Na Figura 5.3 é apresentado um exemplo dessa distorção, na qual o Borrado 2 representa que um vídeo passou por duas aplicações do filtro da média 3×3 e o Borrado 4 representa um vídeo que passou por quatro aplicações do filtro da média.

A investigação que resultou no método proposto iniciou a partir da observação de que o SI das amostras de vídeo está intimamente relacionado à distorção do borramento. A Figura 5.4b apresenta o aumento na quantidade do borramento, e uma redução na informação espacial perceptual.



(a) Informação espacial em vídeos com ruído sal & pimenta.

(b) Informação espacial em vídeos com ruído de borramento.

Figura 5.4: Relação entre: a) Ruído Sal & Pimenta e a Informação Perceptual Espacial; b) Borramento e a Informação Perceptual Espacial.

Uma comparação entre a informação espacial perceptiva dos vídeos originais e processados fornece uma boa aproximação para a qualidade percebida pelos HVS sobre degradação do borramento.

Assim, foi proposta uma modificação na métrica SSIM, considerando a informação

espacial perceptual para avaliar a qualidade de vídeos que apresentam este tipo de degradação, sendo chamado de Blur-SSIM (B-SSIM).

Inicialmente, foi definida a função de comparação das informações espaciais perceptuais, $b(f,h)$ como

$$b(f,h) = \frac{2\mathrm{SI}_f\mathrm{SI}_h}{\mathrm{SI}_f^2 + \mathrm{SI}_h^2}, \tag{5.11}$$

na qual SI_f and SI_h são as informações espaciais perceptuais do vídeo original e processado, respectivamente. Assim o B-SSIM é encontrado a partir da ponderação da métrica SSIM pela função $b(f,h)$ e dado por

$$\text{B-SSIM}(f,h) = b(f,h) \cdot \frac{1}{B}\sum_{i=1}^{B}\mathrm{SSIM}(f_i,h_i). \tag{5.12}$$

5.3 PW-SSIM (*Perceptual Weighted Video Quality Approach*)

Alguns modelos que utilizam métodos de atenção visual em métricas de qualidade de imagem têm sido propostos (Akamine & Farias, 2012), (You *et al.*, 2010). Esses modelos, como os mapas de saliência, as regiões de interesse e foco visual, usam a ponderação para avaliar a qualidade da imagem, dando mais importância às regiões visualmente mais importantes.

O método proposto utiliza a informação espacial perceptiva como forma de ponderação das regiões visualmente mais importantes. Esta ponderação é obtida do seguinte modo: em primeiro lugar é calculada a magnitude dos vetores de gradiente no vídeo original, por meio das máscaras de Sobel, depois é gerado um quadro no qual os valores dos *pixels* são as magnitudes dos gradientes. Em seguida, esse quadro é particionado em blocos 8×8 *pixels* e para cada bloco é calculado o SI,

$$\mathrm{SI}_i = \left(\frac{1}{P}\sum_{j=1}^{P}(\mu_i \ - \ \nabla f_j)^2\right)^{\frac{1}{2}}, \tag{5.13}$$

em que, μ_i representa o valor médio da magnitude do gradiente em um bloco e P é o número de *pixels* no bloco.

Baseado nessa consideração, o valor do SI foi incorporado ao SSIM, levando ao modelo denominado Índice de Semelhança Estrutural com Ponderação Perceptual (PW-SSIM – *Structural Similarity Index with Perceptual Weighting*),

$$\text{PW-SSIM}(f,h) = \frac{1}{N}\sum_{n=1}^{N}\frac{\sum_{i=1}^{B}\text{SSIM}_i(f_n,h_n)\cdot\text{SI}_{in}}{\sum_{i=1}^{B}\text{SI}_{in}}. \tag{5.14}$$

O SI é calculado apenas para o vídeo original, considerando que este é o melhor valor, desta forma, sendo o melhor valor para a ponderação. Além disso, como mostra a Figura 5.4, a informação da percepção espacial apresenta uma variação considerável para vídeos degradados, principalmente com o borramento, que compromete a identificação das regiões de maior importância no vídeo.

5.4 TPW-SSIM (*Temporal Perceptual Weighted Video Quality Approach*)

A partir dos testes realizados com as três métricas propostas no Capítulo 6, foi verificado que a métrica PW-SSIM obteve o melhor resultado, considerando a análise da correlação com os testes subjetivos realizados. Assim, nesta seção é proposta uma modificação nesta métrica.

A modificação surge da observação de que o vídeo apresenta uma componente temporal que não é considerada nos algoritmos de avaliação. Essa componente tem uma correlação satisfatória com as pontuações de opinião, obtidas a partir das avaliações subjetivas.

A mudança temporal de um vídeo é estimada pela diferença de luminância de *pixels* que estão na mesma posição espacial em quadros sucessivos (ITU-T, 1999). Uma abordagem semelhante, proposta por Vu & Deshpande (2012), foi utilizada para estimar a qualidade da componente temporal e se baseou no MS-SSIM. Sendo calculado como segue:

$$\begin{aligned} \mathcal{D}_{f,n} &= |f_{n+1} - f_n|, \\ \mathcal{D}_{h,n} &= |h_{n+1} - f_n|. \end{aligned} \tag{5.15}$$

No algoritmo proposto, a qualidade temporal é estimada por meio do índice de PW-SSIM entre as diferenças dos quadros ($\mathcal{D}_{f,n}$ e $\mathcal{D}_{h,n}$),

$$\text{TP-VQI} = \frac{1}{N-1} \sum_{n=0}^{N-2} \text{PW-SSIM}\left(\mathcal{D}_{f,n}, \mathcal{D}_{h,n}\right). \tag{5.16}$$

O índice PW-SSIM usa regiões com grandes mudanças perceptivas e apresenta uma melhor correlação, com os testes subjetivos, do que o MS-SSIM (Regis *et al.*, 2012b). Assim, o índice de qualidade global é a média entre o índice de qualidade espacial (PW-SSIM) e temporal (TP-VQI), dada por

$$\text{TPW-SSIM} = \frac{\text{PW-SSIM} + \text{TP-VQI}}{2}. \tag{5.17}$$

5.5 DTPW-SSIM (*Disparity Temporal Perceptual Weighted Video Quality Approach*)

Todas as propostas de métricas apresentadas foram desenvolvidas para vídeos em duas dimensões. Para avaliar os vídeos 3D propõe-se uma modificação nas métricas PW-SSIM e TPW-SSIM.

Uma sequência de vídeo estereoscópico é definida como $f_{3D} = (f_l(x, y, n), f_r(x, y, n))$, na qual as funções escalares f_l e f_r correspondem às vistas esquerda e direita, respectivamente.

Considere f_{3D} e h_{3D} duas sequências de vídeos digitais estereoscópicos que correspondem ao vídeo de referência e ao vídeo degradado (em teste), respectivamente. Um algoritmo de avaliação da qualidade objetiva de vídeo 3D com referência total é uma função $g(f_{3D}, h_{3D})$, em que sua pontuação representa a qualidade de h_{3D} com relação a f_{3D}.

A disparidade presente em uma sequência de vídeo estereoscópico é uma informação relacionada ao sentido da percepção estéreo. Esta informação é calculada como a diferença entre dois *pixels* correspondentes nas vistas esquerda e direita. A disparidade

deve ser considerada no desenvolvimento de métricas objetivas, a fim de melhorar a relação entre a predição objetiva e as pontuações subjetivas de um vídeo estereoscópico.

O mapa de disparidade, $D(f(x, y, n))$, é calculado como

$$D(f(x, y, n)) = |f_l(x, y, n) - f_r(x, y, n)| \quad \forall \quad (x, y, n). \tag{5.18}$$

A introdução da informação de disparidade nas métricas objetivas 2D foi feita pela média ponderada das medições objetivas com o mapa de disparidade. A disparidade foi escolhida por ser a forma mais simples de se avaliar a profundidade. Esse critério foi escolhido uma vez que se espera um menor custo computacional para as métricas de avaliação de vídeo.

Esta abordagem foi implementada em duas métricas objetivas, PSNR e SSIM, sendo chamado DPSNR e DSSIM.

$$\mathrm{DMSE}_l(f, h) = \frac{\sum_{p=1}^{P} [f_l(p) - h_l(p)]^2 \cdot D(f(p))}{\sum_{p=1}^{P} D(f(p))}, \tag{5.19}$$

na qual p é uma simplificação para a posição (x, y, n), que representa a posição do p-ésimo *pixel*. O DPSNR_l é calculado como

$$\mathrm{DPSNR}_l(f, h) = 10 \cdot \log_{10} \left[\frac{255^2}{\mathrm{DMSE}_l(f, h)} \right] \mathrm{dB}, \tag{5.20}$$

e o DMSE e o DPSNR para a vista direita (DMSE_r e DPSNR_r) são calculados da mesma maneira, para *pixels* definidos com 8 *bits*. Em seguida, o DPSNR é dado pela média entre o DPSNR_l e o DPSNR_r.

O DSSIM para um quadro é calculado da seguinte maneira

$$\mathrm{DSSIM}(f, h) = \frac{\sum_{i=1}^{B} \mathrm{SSIM}(f_i, h_i) \cdot D_i(f)}{\sum_{i=1}^{B} D_i(f)}, \tag{5.21}$$

no qual $D_i(f)$ é a disparidade média contido em um bloco i.

Além da PSNR e da SSIM, foi acrescentada a disparidade nas métricas PW-SSIM

e TPW-SSIM. O DPW-SSIM para um quadro é calculado da seguinte maneira:

$$\text{DPW-SSIM}(f,h) = \frac{\sum_{i=1}^{B} \text{SSIM}(f_i,h_i)\cdot \text{SI}_i(f)\cdot D_i(f)}{\sum_{i=1}^{B} [\text{SI}_i(f)\cdot D_i(f)]}. \tag{5.22}$$

e o DTPW-SSIM é dado por

$$\text{DTPW-SSIM} = \frac{\text{DPW-SSIM} + \text{DTP-VQI}}{2}, \tag{5.23}$$

na qual o DTP-VQI é dado por,

$$\text{DTP-VQI} = \frac{1}{N-2}\frac{\sum_{n=0}^{N-2} \text{PW-SSIM}\left(\mathcal{D}_{f,n},\mathcal{D}_{h,n}\right)\cdot D_n(f)}{\sum_{n=0}^{N-2} D_n(f)}. \tag{5.24}$$

5.6 Considerações Finais

As métricas propostas para a tese de doutorado foram apresentadas neste capítulo. A métrica B-SSIM é proposta para uma degradação específica, o borramento, e foi obtida pela identificação da relação entre este efeito e a informação espacial perceptual. Ela acrescenta à métrica SSIM a informação espacial perceptual.

A métrica PW-SSIM avalia os vídeos usando a informação perceptual, para fazer uma ponderação em cada região da imagem, considerando que as regiões mais importantes em um vídeo são as que têm a maior informação espacial perceptual.

A essa métrica foram incorporadas mais duas características. A primeira foi a avaliação da informação temporal e a segunda foi a avaliação da disparidade para avaliar vídeos 3D.

Capítulo 6

Apresentação e Análise dos Resultados

Neste capítulo serão apresentados os resultados obtidos com métricas propostas e as comparações com as métricas existentes. Para realizar esta avaliação, foram feitos testes subjetivos e usadas bases de dados projetados para testes de métricas de avaliação de vídeo.

6.1 Avaliação Subjetiva de Vídeos com Degradações

A fim de obter vídeos que apresentem artefatos, que possam se assemelhar aos encontrados em cenas reais, foi desenvolvido um *software* que permite controlar o tipo e a intensidade do ruído desejado. Esse programa foi construído tomando como base o trabalho de Albini (2009).

O *software* que gera os artefatos recebe os vídeos originais e permite oito saídas:

- o artefato de borramento, formado pelo efeito do filtro da média no vídeo com duas ou quatro passagens;
- o artefato sal & pimenta, cuja degradação pode ser de 1 ou 3% do vídeo;
- o artefato de blocagem, cuja degradação pode ser de 1 ou 3% do vídeo.

Na Figura 6.1 é ilustrado o resultado da aplicação das degradações nos vídeos "Foreman", "Glasgow" e "Mobile & Calendar". Nas Figuras 6.1b e 6.1c são representados os artefatos de blocagem com 1% e 3% de probabilidade de ocorrência, respectivamente. Nas Figuras 6.1e e 6.1f é ilustrado o artefato borramento com duas e quatro aplicações

do filtro da média. E, por fim, as Figuras 6.1h e 6.1i contêm o ruído sal & pimenta com 1% e 3% de probabilidade de ocorrência.

(a) Foreman.

(b) 1% de Blocagem.

(c) 3% de Blocagem.

(d) Glasgow.

(e) 2 aplicações do filtro da média.

(f) 4 aplicações do filtro da média.

(g) Mobile & Calendar.

(h) 1% de Sal & Pimenta.

(i) 3% de Sal & Pimenta.

Figura 6.1: Amostras de vídeo degradadas pelo simulador.

Com o *software* pronto, foram inseridos artefatos nos vídeos, escolhidos na Seção 6.1.1, com diferentes níveis de intensidade. A partir desses vídeos gerados foram realizadas avaliações objetivas e subjetivas.

6.1.1 Escolha das Amostras de Vídeo

De acordo com a Recomendação ITU-T P.910 (ITU-T, 1999), os valores de SI e TI dos vídeos devem ser utilizados para decidir se uma amostra de vídeo fará, ou não, parte de um teste subjetivo. Por exemplo, caso seja feito um teste subjetivo com

quatro sequências, é interessante escolher aquelas cujos valores de informação espacial e temporal estejam mais dispersos.

A fim de constituir um conjunto de vídeos para integrar uma avaliação subjetiva, o experimento realizado considerou vinte vídeos no formato QCIF, comumente citados na literatura e disponíveis na Internet (Trace, 2008). Para esses vídeos foram calculados os valores de SI e TI, e gerado um gráfico de caracterização da informação espacial e temporal, representado na Figura 6.2.

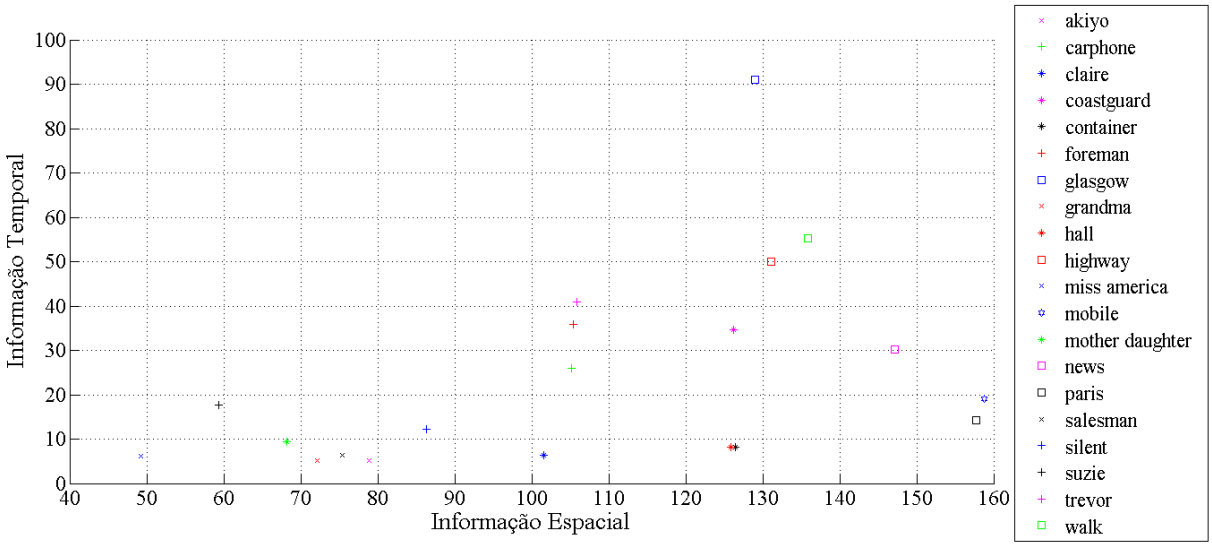


Figura 6.2: Gráfico de dispersão entre a informação espacial (SI) e a informação temporal (TI) de vídeos no formato QCIF.

As amostras escolhidas para a avaliação subjetiva foram: "Mobile & Calendar", por ter o maior valor de informação espacial dentre todas as amostras; "Glasgow", por possuir o maior valor de informação temporal; "Mother and Daughter", por ter baixo valor de informação temporal e espacial; e "Foreman" por possuir valores medianos de informação temporal e espacial (Figura 6.3).

6.1.2 Análise dos Resultados da Avaliação Subjetiva

O objetivo da realização da avaliação subjetiva foi obter valores médios de opinião (MOS) para diversos vídeos degradados, com o intuito de possibilitar a comparação com os valores das métricas objetivas e efetuar a avaliação da eficiência de cada uma. Como se desejava obter os resultados de forma rápida, os testes subjetivos foram realizados pela Internet.

(a) Foreman. (b) Glasgow. (c) Mobile & Calendar. (d) Mother-Daughter.

Figura 6.3: Amostras de vídeos Foreman, Glasgow, Mobile & Calendar e Mother-Daughter utilizadas na avaliação subjetiva.

Para a realização da avaliação subjetiva, foram inseridas nos vídeos as degradações blocagem, borramento e ruído sal & pimenta, com dois níveis para cada degradação. Para a blocagem e sal & pimenta os níveis de degradação foram 1 e 3% do vídeo, respectivamente. Para a degradação borramento, o filtro da média 3×3 foi usado duas e quatro vezes.

A avaliação subjetiva realizada foi composta por cinquenta e dois participantes de 15 a 60 anos e utilizou o método ACR com uma escala de votação discreta de cinco níveis. Os vídeos foram disponibilizados em *site*, na qual os participantes assistiam aos vídeos e davam a nota em seguida. O *site* foi programado para funcionar de acordo com o método ACR.

Essa maneira de avaliação permite que os resultados sejam obtidos de forma rápida, dando um panorama do que está acontecendo no experimento, porém esse método não gera bons indicadores de confiabilidade, ocorrendo pela falta de controle na visualização pelos participantes (distância do participante para a tela, tipo de monitor utilizado, luminosidade da sala e etc).

Por esses motivos, os resultados encontrados serviram para obtenção de um comportamento inicial das métricas propostas e logo depois as métricas foram avaliadas em bases de vídeo com indicadores melhores de confiabilidade. Os resultados obtidos pela avaliação subjetiva para as diferentes degradações são apresentados nas Figuras 6.4, 6.5 e 6.6.

Para os vídeos, as degradações afetam a qualidade visual de forma distinta. Para o ruído de blocagem, os vídeos que tiveram sua qualidade mais afetada foram o Foreman e o Glasgow, que têm mais informações temporais. O Glasgow e o Mobile foram os mais afetados pelo borramento, e são os que têm mais informações espaciais. Para o ruído sal & pimenta, os vídeos que tiveram suas qualidades mais afetadas foram o

Vídeos	Nome	MOS
1	Mother 1%	2,76
2	Mother 3%	1,64
3	Glasgow 1%	2,06
4	Glasgow 3%	1,46
5	Mobile 1%	2,26
6	Mobile 3%	1,56
7	Foreman 1%	1,97
8	Foreman 3%	1,62

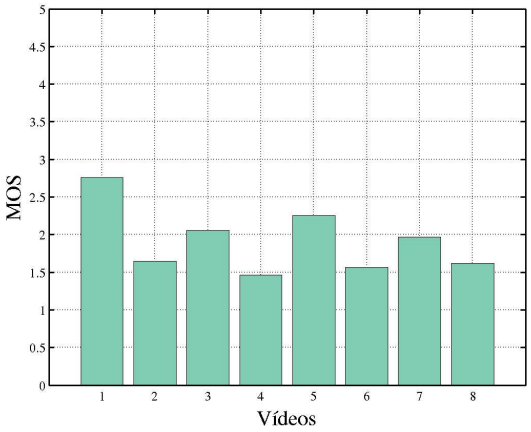


Figura 6.4: MOS obtidos para cada uma das sequências sob teste para a blocagem.

Vídeos	Nome	MOS
1	Mother 2	2,05
2	Mother 4	1,55
3	Glasgow 2	1,77
4	Glasgow 4	1,41
5	Mobile 2	1,77
6	Mobile 4	1,31
7	Foreman 2	2,05
8	Foreman 4	1,53

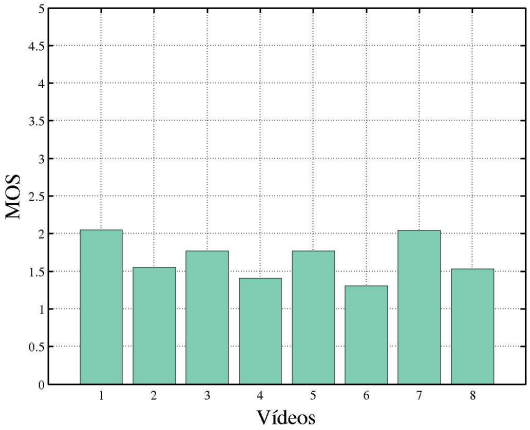


Figura 6.5: MOS obtidos para cada uma das sequências sob teste para o borramento.

Mother and Daughter e o Foreman, que são os vídeos com menor informação espacial.

Também foram avaliados nos testes subjetivos o desvio padrão e o coeficiente de variação, como pode ser observado nas Tabelas 6.1, 6.2 e 6.3. O desvio padrão variou de 0,53 a 0,84, enquanto o coeficiente de variação variou entre 24% e 46%.

Os coeficientes de variação foram altos, caracterizando assim uma elevada dispersão entre os dados. Os coeficientes para os vídeos borrados ficaram, para todos os vídeos, acima de 40%, mostrando a dificuldade de se quantificar a qualidade de vídeos com essa degradação.

Vídeos	Nome	MOS
1	Mother 1%	2,32
2	Mother 3%	1,65
3	Glasgow 1%	2,43
4	Glasgow 3%	1,94
5	Mobile 1%	2,72
6	Mobile 3%	1,97
7	Foreman 1%	2,14
8	Foreman 3%	1,74

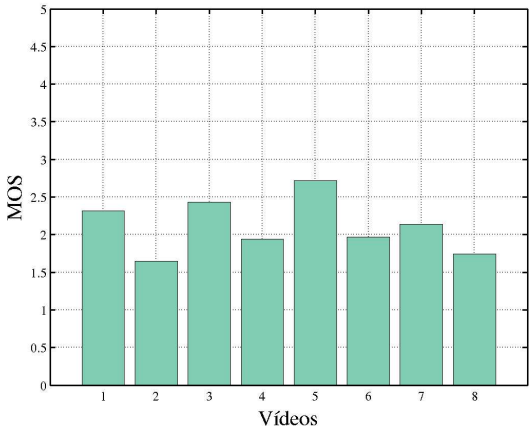


Figura 6.6: MOS obtidos para cada uma das sequências sob teste para o ruído sal & pimenta.

Tabela 6.1: Desvio padrão e coeficiente de variação para os vídeos com blocagem.

Vídeo	Desvio padrão	CV (%)
Mother 1%	0,6834	24,37
Mother 3%	0,6369	38,74
Glasgow 1%	0,7446	36,21
Glasgow 3%	0,5395	36,88
Mobile 1%	0,7708	34,14
Mobile 3%	0,5616	35,9
Foreman 1%	0,8033	40,78
Foreman 3%	0,6017	37,11

Tabela 6.2: Desvio padrão e coeficiente de variação para os vídeos com borrado.

Vídeo	Desvio padrão	CV (%)
Mother 2	0,8392	40,92
Mother 4	0,6693	45,11
Glasgow 2	0,8072	45,62
Glasgow 4	0,6059	42,91
Mobile 2	0,8248	46,62
Mobile 4	0,5335	40,68
Foreman 2	0,8491	41,51
Foreman 4	0,6132	40,07

6.2 Comparação entre Métricas Objetivas Existentes usando os Resultados da Avaliação Subjetiva

Os coeficientes de correlação de Pearson foram calculados entre a MOS (Seção 6.1.2), que é o valor médio das votações subjetivas, e os valores das métricas objetivas, para

Tabela 6.3: Desvio padrão e coeficiente de variação para os vídeos com ruído sal & pimenta.

Vídeo	**Desvio padrão**	**CV (%)**
Mother 1%	0,6589	28,45
Mother 3%	0,6191	37,56
Glasgow 1%	0,6358	26,12
Glasgow 3%	0,7253	37,34
Mobile 1%	0,6911	25,44
Mobile 3%	0,5862	29,82
Foreman 1%	0,8573	40,13
Foreman 3%	0,6155	35,32

cada amostra de vídeo contaminada com um tipo e nível de ruído. Os coeficientes estão dispostos na Tabela 6.4. Em termos de qualidade, quanto maior o módulo do valor do coeficiente de correlação de Pearson, mais próxima está uma métrica objetiva de representar o conceito de qualidade para um grupo de espectadores.

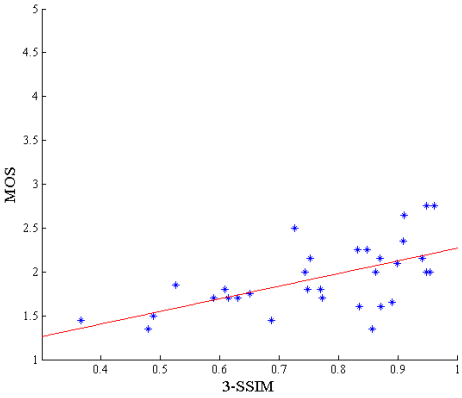
Para realizar a comparação entre métricas existentes, foram analisadas as medidas objetivas MSE, PSNR, SSIM, MS-SSIM e 3-SSIM. A comparação foi realizada com relação aos valores de MOS das 20 primeiras pessoas que realizaram os testes subjetivos.

A observação da Tabela 6.4 indica que as métricas baseadas no índice de similaridade estrutural são mais adequadas para predizer a qualidade percebida pelo HVS, pois obtiveram coeficientes de correlação significativamente superiores se comparadas à abordagem de sensibilidade ao erro. Especificamente, a métrica 3-SSIM obteve os três coeficientes de correlação de valores mais altos nos quatro ensaios realizados.

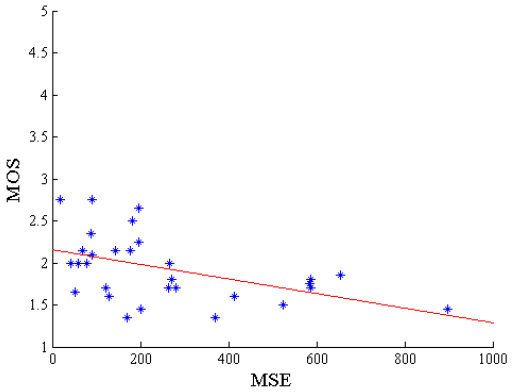
Tabela 6.4: Avaliação das métricas existentes usando o coeficiente de correlação de Pearson.

Métrica	Sal & Pimenta	Ruído Gaussiano Branco	Blocagem	Borramento
MSE	-0,8224	-0,8637	-0,5157	-0,5702
PSNR	0,8284	0,8580	0,6971	0,6067
SSIM	**0,9022**	0,9313	0,7918	0,7762
MS-SSIM	0,8835	0,920	0,764	0,7635
3-SSIM	0,8935	**0,934**	**0,8444**	**0,8651**

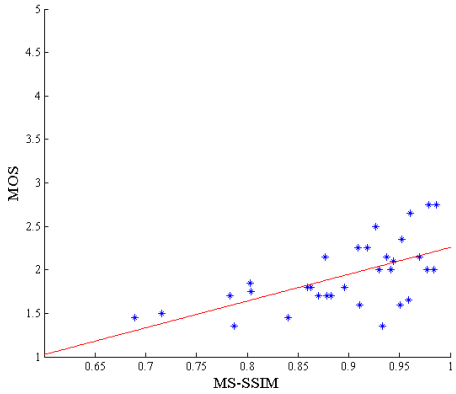
Na Figura 6.7 são ilustrados os gráficos de comparação entre os valores subjetivos (MOS) e o resultado das métricas objetivas para os 32 vídeos considerados no experimento.



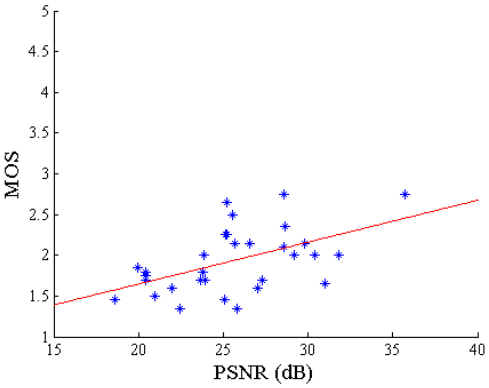
(a) MOS versus 3-SSIM.



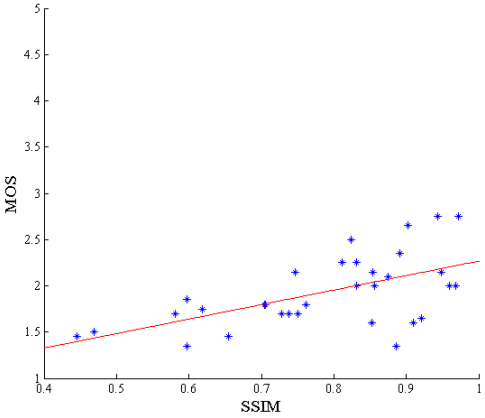
(b) MOS versus MSE.



(c) MOS versus MS-SSIM.



(d) MOS versus PSNR.



(e) MOS versus SSIM.

Figura 6.7: Gráficos da MOS versus a métrica objetiva sob teste.

6.3 Comparação entre os Modelos Propostos usando os Resultados da Avaliação Subjetiva

As primeiras métricas sugeridas neste trabalho, PSNR_e e SSIM_e (Seção 5.1), B-SSIM (Seção 5.2) e PW-SSIM (Seção 5.3) foram comparadas, usando o coeficiente de correlação de Pearson, com as métricas que obtiveram os melhores resultados, na Tabela 6.4, e apresentadas na Tabela 6.5.

Tabela 6.5: Avaliação das métricas propostas usando os resultados da avaliação subjetiva.

Modelo	Ruído Sal & Pimenta	Borramento	Blocagem
PSNR	$0,8284$	$0,6067$	$0,6971$
PSNR_e	$0,8331$	$0,6248$	$0,7523$
SSIM	$0,9022$	$0,7762$	$0,7918$
SSIM_e	$0,9030$	$0,872$	$0,8367$
3-SSIM	$0,8953$	$0,8651$	**0,8444**
B-SSIM		**0,8753**	
PW-SSIM	**0,9195**	$0,8662$	$0,8336$

Para os vídeos com blocagem, a métrica que obteve maior coeficiente de correlação foi a 3-SSIM, no entanto, a diferença em relação à métrica PW-SSIM foi de apenas $0,011$. Para os vídeos degradados com o ruído sal & pimenta, o melhor coeficiente de correlação foi obtido pela métrica PW-SSIM.

Para o artefato de borramento a métrica que obteve o melhor resultado foi o B-SSIM, como era esperado, com a correlação de $0,905$. A diferença entre o B-SSIM e o SSIM foi de $0,15$ e menos de $0,01$ para a métrica PW-SSIM. O B-SSIM e o SSIM_e obtiveram resultados semelhantes, porém o SSIM_e obteve vantagem sobre B-SSIM em relação ao tempo de processamento, como visto na Tabela 6.6. Para a realização do cálculo do tempo de processamento foi usado um computador do Labcom 4 que possui as seguintes configurações: Processador: AMD Athlon(tm) 64 X2 Dual Core Processor 4400+ 2.3GHz HyperTransport 1000MHz, Placa-mãe: ASUS M2N-X, Disco-rígido: SATA II SAMSUNG 160 GB e Memória RAM: Kingston DDR2 1 GB 667MHz.

Para as métricas PSNR_e e SSIM_e houve uma melhoria em todos os coeficientes de correlação em comparação aos modelos PSNR e MSE e superou o 3-SSIM para as degradações borramento e sal e pimenta.

Tabela 6.6: Tempo de processamento.

Modelo	Tempo (ms)	Desvio Padrão
PSNR	$235,15$	$17,40$
SSIM	$366,55$	$7,83$
$SSIM_e$	$1489,25$	$66,77$
B-SSIM	$2887,25$	$20,64$
PW-SSIM	$1544,50$	$10,14$

A métrica que se apresenta com os melhores coeficientes de correlação é a PW-SSIM, que proporcionou o melhor resultado para o ruído sal e pimenta e sempre está entre os melhores para os demais ruídos, sendo a diferença entre a PW-SSIM e a melhor métrica para cada degradação não superior a $0,01$.

6.4 Avaliação do Efeito do Operador Diferencial usando os Resultados da Avaliação Subjetiva

O PW-SSIM utiliza um operador diferencial, assim nesta seção é apresentada uma avaliação do efeito de diferentes operadores na métrica PW-SSIM. Para essa avaliação foram usados os operadores de Sobel (S-PW-SSIM), Prewitt (P-PW-SSIM), Roberts (R-PW-SSIM) e Laplaciano (L-PW-SSIM) e na Figura 6.8 é apresentado um dos quadros do vídeo Mother obtidos pelo uso de cada um dos operadores.

(a) Original. (b) Sobel. (c) Prewitt. (d) Roberts. (e) Laplace.

Figura 6.8: Um quadro do vídeo Mother após a operação dos diferentes gradientes.

Para avaliar o desempenho do operador diferencial foi utilizado o Coeficiente de Correlação de Pearson linear (PLCC), de maneira a se obter a comparação entre as medidas objetivas e MOS.

A Tabela 6.7 apresenta o PLCC, entre o MOS e medida objetiva, considerando as distorções. De um modo geral, o melhor desempenho das métricas foi obtido com a utilização do operador diferencial Sobel. Especificamente, o S-PW-SSIM, no cenário

de vídeos com a degradação sal e pimenta, e o S-PW-GSSIM, para vídeos que apresentaram as degradações de borramento e blocagem, foram os que obtiveram os maiores valores de correlação.

O operador de segunda ordem Laplaciano apresenta uma correlação maior para a degradação blocagem. Também é importante observar que nenhum dos modelos propostos apresentou melhora na correlação para o ruído gaussiano, o que pode ser justificado pela baixa correlação entre o ruído gaussiano e SI.

Tabela 6.7: Avaliação dos operadores gradientes nas métricas PW-SSIM e G-SSIM (Regis *et al.*, 2012a).

Modelo	**Sal & Pimenta**	**Borramento**	**Blocagem**	**Gaussiano**
PSNR	0,828	0,607	0,697	0,858
SSIM	0,902	0,776	0,792	**0,931**
S-PW-SSIM	**0,920**	0,866	0,834	0,918
R-PW-SSIM	0,883	0,870	0,831	0,903
P-PW-SSIM	0,914	0,874	0,833	0,912
L-PW-SSIM	0,907	0,882	0,840	0,915
G-SSIM	0,864	0,972	0,867	0,822
S-PW-GSSIM	0,914	**0,984**	**0,876**	0,913
R-PW-GSSIM	0,865	0,983	0,872	0,905
P-PW-GSSIM	0,911	0,984	0,875	0,911
L-PW-GSSIM	0,894	0,977	**0,876**	0,917

No que diz respeito ao custo computacional, os operadores de primeira ordem Sobel e Prewitt (S-PW-SSIM e P-PW-SSIM), apresentam praticamente o mesmo incremento no tempo de execução (450 ms), comparando com o SSIM. Os operadores de Roberts e Laplaciano (R-PW-SSIM e L-PW-SSIM) apresentaram tempos de processamento mais curtos, em relação ao SSIM, a diferença de tempo foi de 331 ms e 210 ms, respectivamente, e em relação ao GSSIM, o R-PW-GSSIM e L-PW-GSSIM tiveram os tempos 370 ms e 286 ms, respectivamente. Para obter esses resultados foi utilizado a mesma máquina da Seção anterior.

6.5 Avaliação das Métricas Objetivas Usando a Base de Dados LIVE

Para avaliar as métricas PW-SSIM e TPW-SSIM foi utilizado a base de dados LIVE (Apêndice A). A escolha dessa base de dado ocorreu por conta da variedade de degra-

dações disponibilizadas, por ele conter a avaliação com várias métricas objetivas e por ter uma alta confiabilidade entre os pesquisadores. Além das métricas já apresentadas foi acrescentada a avaliação das métricas ViMSSIM (*MOtion-based Video Integrity Evaluation*), S-ViMSSIM e T-ViMSSIM apresentadas em Vu & Deshpande (2012).

Os resultados das avaliações podem ser observados nas Tabelas 6.8 e 6.9. Os coeficientes de correlação indicam que o algoritmo proposto é o melhor para as transmissões sobre redes IP e para o codificação MPEG-2. Para a transmissão usando redes sem fio o algoritmo proposto, o T-MOVIE (*Temporal Video quality Metric Structural Similarity Index*), e o MOVIE apresentaram os melhores resultados. Para as distorções geradas com a codificação H.264 o Temporal-ViMSSIM (T-ViMSSIM) é o melhor.

Tabela 6.8: Avaliação das métricas usando o coeficiente de correlação de Pearson na base de dados LIVE.

Algoritmos	**H.264**	**IP**	**Sem fio**	**MPEG-2**	**Todos**
PSNR	0,5492	0,4645	0,6690	0,3891	0,5621
SSIM	0,6656	0,5119	0,5401	0,5491	0,5444
VQM	0,6459	0,6480	0,7325	0,7860	0,7236
S-MOVIE	0,7252	0,7378	0,7883	0,6587	0,7451
T-MOVIE	0,7920	0,7383	**0,8371**	**0,8252**	**0,8217**
MOVIE	0,7902	**0,7622**	**0,8386**	0,7595	0,8116
S-ViMSSIM	0,7834	0,7503	0,7837	0,7515	0,7796
T-ViMSSIM	**0,8810**	0,6890	0,8219	0,7909	0,8122
ViMSSIM	0,8117	0,7322	0,8327	0,7978	**0,8260**
B-SSIM	0, 7212	0, 5934	0, 5993	0, 6445	0, 6164
PW-SSIM	0, 7263	0, 6331	0, 6376	0, 6665	0, 6370
TPW-SSIM	**0,8180**	**0,8061**	0,8298	**0,8144**	0,7601

A métrica TPW-SSIM teve uma melhoria de mais de 22,89% para a degradação da codificação H.264, mais de 24,3 % para a transmissão IP, mais de 13,28 % para a transmissão sem fio e mais de 3,61 % para a codificação MPEG-2 considerando as métricas PSNR, SSIM e VQM e o uso da correlação de Pearson.

Para todas as degradações testadas a métrica TPW-SSIM obteve resultados bons, uma vez que está sempre entre as três melhores métricas, de acordo com a Tabela 6.8, resultado esse que nenhuma outra métrica conseguiu. Para a transmissão IP a métrica foi superior ao melhor resultado obtido (MOVIE) em 5,76 %. Para a codificação H.264 e para a codificação MPEG-2 o resultado obtido foi inferior apenas ao da métrica T-ViMSSIM em 7,15 % e ao da métrica T-MOVIE em 1,31 %, respectivamente. Para

a transmissão sem fio as métricas MOVIE e T-MOVIE foram superiores a métrica TPW-SSIM em 0,87 %.

Tabela 6.9: Avaliação das métricas usando o coeficiente de correlação de Spearman na base de dados LIVE.

Algoritmos	**H.264**	**IP**	**Sem fio**	**MPEG-2**	**Todos**
PSNR	0,4585	0,4167	0,6574	0,3862	0,5398
SSIM	0,6514	0,4550	0,5233	0,5545	0,5257
VQM	0,6520	0,6383	0,7214	0,7810	0,7026
S-MOVIE	0,7066	0,7046	0,7927	0,6911	0,7270
T-MOVIE	0,7797	**0,7192**	**0,8114**	**0,8170**	**0,8055**
MOVIE	0,7664	0,7157	0,8109	0,7733	0,7890
S-ViMSSIM	0,7713	0,6521	0,7340	0,7694	0,7690
T-ViMSSIM	**0,8580**	0,6650	0,7951	0,7499	0,7984
ViMSSIM	**0,8559**	0,6774	0,8111	0,7630	**0,8211**
B-SSIM	0,7103	0,4994	0,5790	0,6254	0,6062
PW-SSIM	0,7229	0,5573	0,6161	0,6311	0,6263
TPW-SSIM	0,7938	**0,7758**	**0,8158**	**0,7767**	0,7553

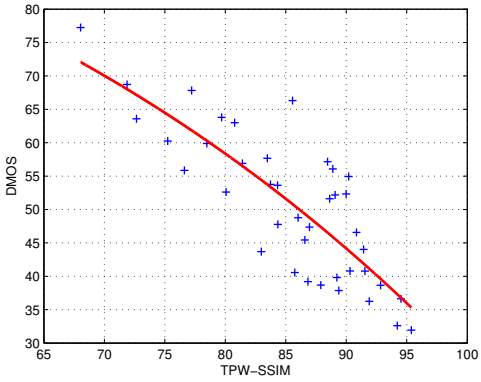
Para a correlação de Sperman, a métrica TPW-SSIM foi melhor do que as métricas PSNR, SSIM e VQM em mais de 21,75% para a degradação da codificação H.264, mais de 21,54 % para a transmissão IP, mais de 13,08 % para a transmissão sem fio e uma perda para a métrica VQM de 0,55 % com a codificação MPEG-2.

Como ocorreu para a correlação de Pearson, os resultados da métrica TPW-SSIM para a correlação de Spearman ficaram sempre entre as três melhores métricas testadas, Tabela 6.9. Para a transmissão IP e sem fio a métrica foi superior ao melhor resultado obtido (T-MOVIE) em 7,87 % e 0,54 %, respectivamente. Para a codificação MPEG-2 o resultado obtido foi inferior apenas a métrica T-MOVIE em 5,4 % e para a codificação H.264, as métricas T-ViMSSIM e ViMSSIM foram superiores em 7,48 %.

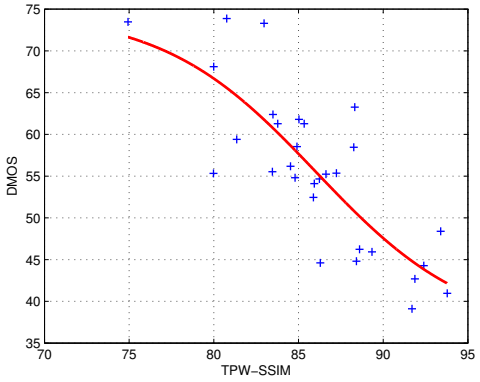
Os gráficos de dispersão da Figura 6.9 ilustram o comportamento não linear das medidas do algoritmo proposto com relação ao conceito de qualidade dos observadores nas experiências subjetivas realizadas no LIVE *Video Quality Database*.

6.6 Avaliação Objetiva de Vídeos 3D

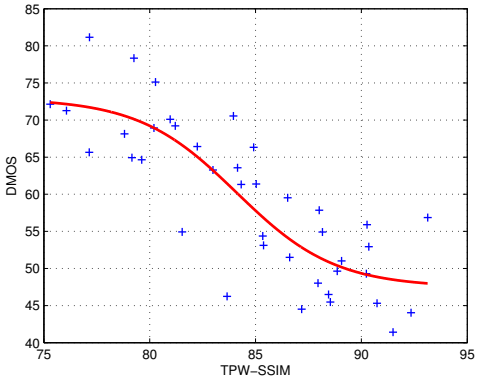
Para avaliar as métricas propostas para vídeos 3D foi utilizado a base de dados NAMA3DS1-COSPAD1 (Apêndice B). Essa base de dados disponibiliza os vídeos



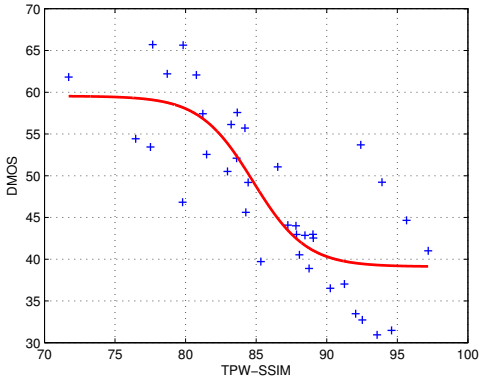
(a) H.264.



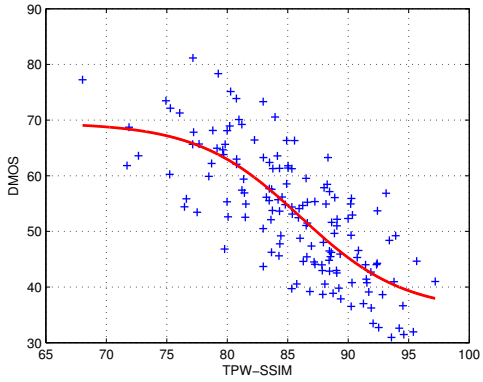
(b) IP.



(c) Sem fio.



(d) MPEG-2.



(e) Todos os dados.

Figura 6.9: O comportamento entre as medidas do algoritmo proposto e os escores das experiências subjetivas realizadas na LIVE *Video Quality Database*.

utilizados na avaliação subjetiva e os resultados do teste.

Os vídeos utilizados foram degradados utilizando os codificadores H.264 e JPEG, a transcodificação espacial e a *image sharpening*. Para os vídeos codificados com o H.264

foram usados três taxas de *bits* diferentes, enquanto para o JPEG foram usadas quatro taxas de *bits*. As outras degradações foram analisadas apenas uma vez, o que resulta em apenas dez sequências a serem avaliadas para cada vídeo de referência.

Considerando a quantidade de vídeos com a mesma degradação foram escolhidos apenas os vídeos que contêm as codificações H.264 e JPEG para avaliar as métricas 3D. Os resultados para os vídeos codificados com H.264 podem ser observados na Tabela 6.10, e para os codificados com JPEG na Tabela 6.11.

Tabela 6.10: Medidas de desempenho dos algoritmos objetivos para a codificação H.264.

Algorithm	**PLCC**	**SROCC**	**KROCC**	**RMSE**
PSNR	0,774946	0,721424	0,533869	0,689299
SSIM	0,730523	0,716222	0,555117	0,744770
IQSSA	0,762415	0,721424	0,533869	0,705725
PQM	0,765747	0,756468	0,597614	0,701421
PW-SSIM	**0,915983**	**0,906776**	**0,756978**	**0,437573**
TPW-SSIM	0,863014	0,804381	0,624175	0,550957
DPSNR	0,863640	0,838604	0,640111	0,549789
DSSIM	0,901635	0,892266	0,746354	0,471688
DPW-SSIM	**0,954403**	**0,937166**	**0,815412**	**0,325572**
DTPW-SSIM	0,868858	0,797536	0,597614	0,539923

As seguintes medidas foram utilizadas como indicadores de desempenho dos algoritmos objetivos: Coeficiente de Correlação Linear de Pearson (PLCC – *Pearson Linear Correlation Coefficient*), Coeficiente de Correlação de Spearman (SROCC – *Spearman Rank-Order Correlation Coefficient*), Coeficiente de Correlação de Kendall (KROCC – *Kendall Rank-Order Correlation Coefficient*) e a Raiz Quadrada do Erro Médio (RMSE – *Root Mean Squared Error*).

Para comparar as métricas propostas nesta tese foram também avaliadas as métricas PSNR e SSIM, para vídeos 2D e as métricas PQM e IQSSA, para vídeos 3D. Para obter a medida IQSSA foi calculada a média das medidas IQA e SSA, Seção 4.1.3.

Para a codificação H.264, as métricas existentes na literatura não proporcionaram bons resultados, como mostrado na Tabela 6.10, considerando as medidas de correlação apresentadas. As métricas propostas para vídeos 2D nesta tese (PW-SSIM e TPW-SSIM) obtiveram resultados melhores, com destaque para a métrica PW-SSIM que obteve valores de correlação acima de 0,9 com o PLCC e o SROCC.

Uma das possíveis causas da métrica PW-SSIM ter resultados melhores que a métrica TPW-SSIM é a correlação entre a informação espacial e temporal dos vídeos que

foram usados nas bases de dados LIVE e NAMA3DS1-COSPAD1, uma vez que a métrica usa pesos iguais para a informação espacial e temporal. Para a base de dados LIVE, a correlação entre a informação espacial e a temporal é de 0,8566 e na base NAMA3DS1-COSPAD1 a correlação é de -0,00594.

Após inserir a disparidade nas métricas, as correlações de todas as métricas melhoraram, com a métrica DPW-SSIM obtendo os melhores resultados. Mais uma vez a métrica com a informação temporal não obteve resultado melhor do que a métrica que usa apenas a informação espacial.

A métrica DPW-SSIM teve uma correlação superior às demais métricas, para a degradação H.264, para todos os tipos de correlação avaliados. Em relação às métricas existentes (PSNR, SSIM, IQSSA e PQM) essa métrica foi superior em mais de 23,16 % para a correlação de Pearson, 23,89 % para a correlação Sperman, 36,45 % para a correlação de Kendall e 52,76 % para a correlação RMSE.

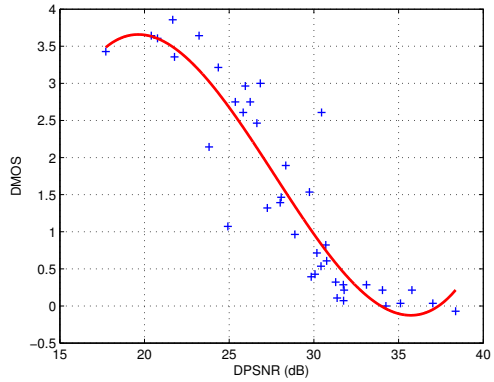
Os gráficos de dispersão da Figura 6.10 ilustram o comportamento não linear das medidas dos algoritmos com relação às experiências subjetivas realizadas na base de dados NAMA3DS1-COSPAD1, para a codificação H.264.

Para a codificação JPEG, os resultados foram de acordo com os obtidos para a codificação H.264. Porém, em termos de correlação, todas as métricas obtiveram resultados melhores para a codificação JPEG, Tabela 6.11.

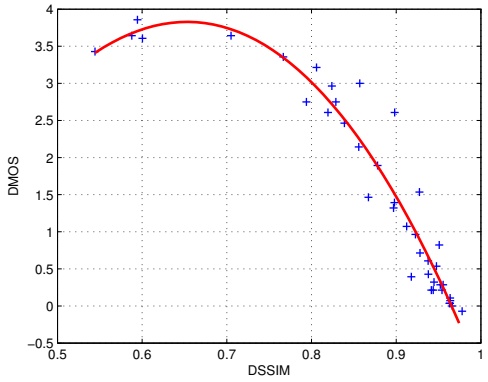
Tabela 6.11: Medidas de desempenho dos algoritmos objetivos JPEG.

Algoritimos	**PLCC**	**SROCC**	**KROCC**	**RMSE**
PSNR	0,828049	0,825865	0,662380	0,734844
SSIM	0,896314	0,907419	0,750010	0,581185
IQSSA	0,857576	0,880766	0,713927	0,674162
PQM	0,884372	0,903946	0,760319	0,611806
PW-SSIM	**0,972477**	**0,965980**	**0,860836**	**0,305388**
TPW-SSIM	0,908694	0,914551	0,760319	0,547162
DPSNR	0,914034	0,927596	0,770629	0,531663
DSSIM	0,969310	0,962132	0,853104	0,322222
DPW-SSIM	**0,975911**	**0,971048**	**0,865991**	**0,285951**
DTPW-SSIM	0,917419	0,917648	0,768051	0,521553

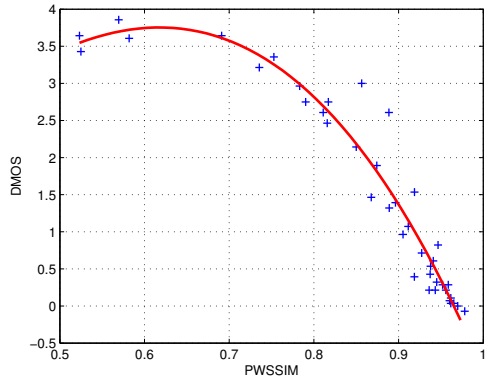
Com a inserção da disparidade nas métricas os resultados melhoram, mas não muito para as métricas PW-SSIM e TPW-SSIM, mas mesmo assim a métrica DPW-SSIM foi a que obteve os melhores resultados. Em relação às métricas já existentes (PSNR, SSIM, IQSSA e PQM) essa métrica foi superior em mais de 8,88 % para a correlação



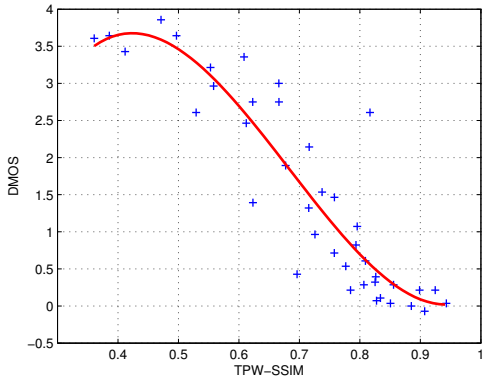
(a) DPSNR



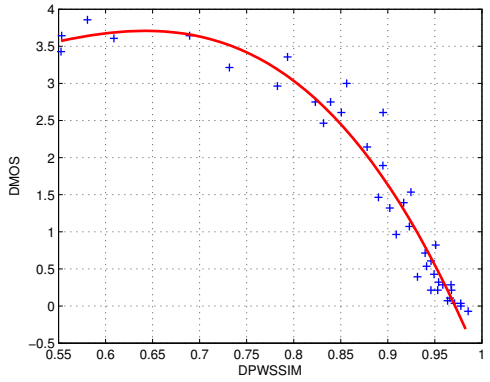
(b) DSSIM



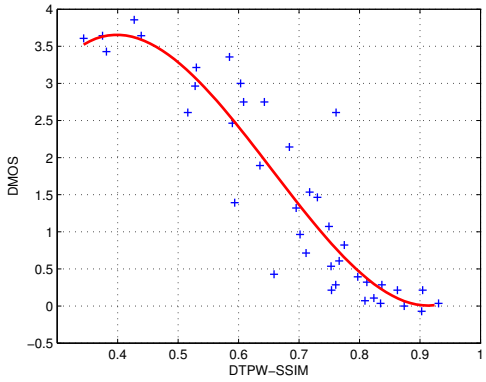
(c) PW-SSIM



(d) TPW-SSIM



(e) DPW-SSIM



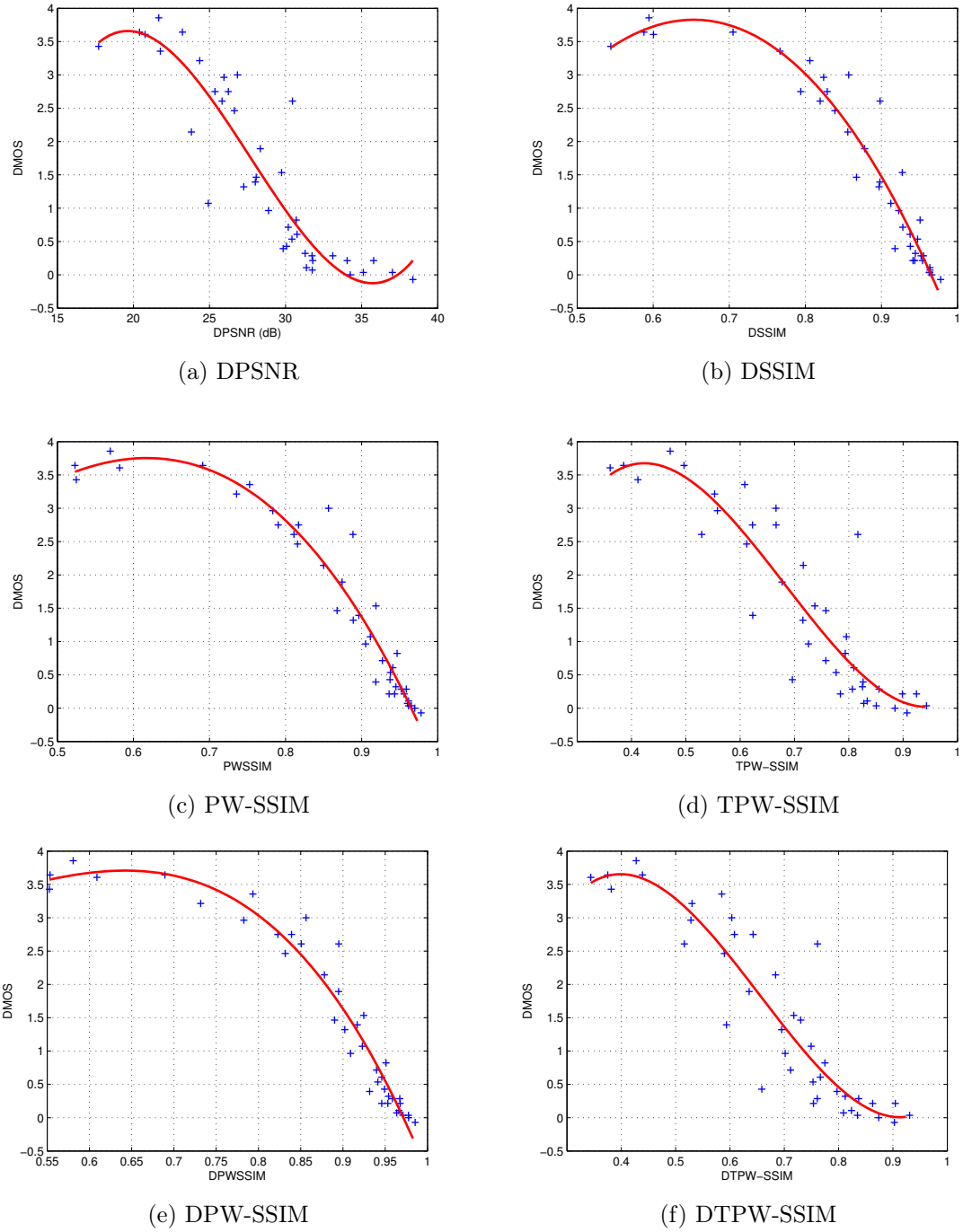
(f) DTPW-SSIM

Figura 6.10: O comportamento entre as medidas do algoritmo proposto e os escores das experiências subjetivas realizadas no NAMA3DS1-COSPAD1, para a codificação H.264.

de Pearson, 7 % para a correlação Sperman, 13,9 % para a correlação de Kendall e 50,81 % para a correlação RMSE.

Os gráficos de dispersão, da Figura 6.11, ilustram o comportamento não-linear das

medidas dos algoritmos com relação às experiências subjetivas realizadas na base de dados NAMA3DS1-COSPAD1, para a codificação JPEG.



(a) DPSNR (b) DSSIM

(c) PW-SSIM (d) TPW-SSIM

(e) DPW-SSIM (f) DTPW-SSIM

Figura 6.11: O comportamento entre as medidas do algoritmo proposto e os escores das experiências subjetivas descritas no NAMA3DS1-COSPAD1 para a codificação JPEG.

Foi analisada também a correlação das medidas para todos os vídeos codificados,

como mostrado na Tabela 6.12. Os resultados apresentados indicam mais uma vez que a métrica DPW-SSIM obteve os melhores resultados, com ganho em relação as métricas PSNR, SSIM, IQSSA e PQM de mais 16,16% para a correlação de Pearson, 13,54 % para a correlação Sperman, 25,34 % para a correlação de Kendall e 54,02 % para a correlação RMSE.

Foi importante a inserção da disparidade na métrica SSIM, produzindo a medida DSSIM, o que permitiu um ganho mais alto que o das outras métricas, fazendo com que a métrica DSSIM tenha resultados próximos ao da métrica DPW-SSIM.

Tabela 6.12: Medidas de desempenho dos algoritmos objetivos para os vídeos codificados.

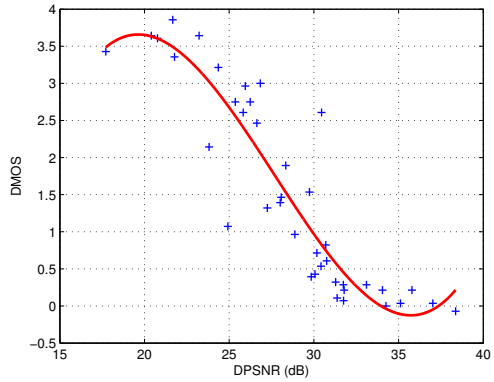
Algorithm	**PLCC**	**SROCC**	**KROCC**	**RMSE**
PSNR	0,790152	0,766721	0,588923	0,750780
SSIM	0,832476	0,841566	0,658728	0,678694
IQSSA	0,815261	0,810482	0,629569	0,709356
PQM	0,828756	0,841508	0,662262	0,685490
PW-SSIM	0,951992	0,943427	0,800988	0,374981
TPW-SSIM	0,889386	0,867494	0,696723	0,559991
DPSNR	0,875461	0,858578	0,678167	0,592001
DSSIM	0,944039	0,942530	0,801872	0,404026
DPW-SSIM	0,967001	0,955609	0,830147	0,312082
DTPW-SSIM	0,892946	0,864401	0,692305	0,551425

Os gráficos de dispersão da Figura 6.12 ilustram o comportamento não linear das medidas dos algoritmos com relação ao conceito de qualidade dos observadores nas experiências subjetivas realizadas para montar o NAMA3DS1-COSPAD1 para os vídeos testados.

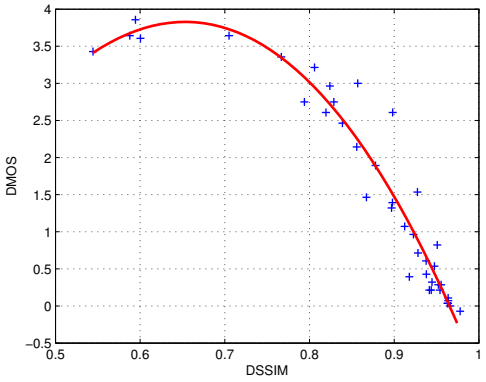
6.7 Considerações Finais

Este capítulo apresentou as avaliações objetivas para vídeos em duas e três dimensões. Para alguns vídeos 2D foram gerados artefatos, para que se pudesse comparar as métricas objetivas com as subjetivas, e foi utilizado a base de dados LIVE.

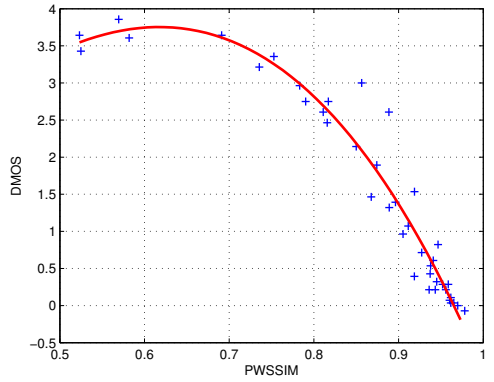
Neste trabalho, foi realizada uma avaliação subjetiva utilizando o método ACR e os resultados foram comparados com os resultados das métricas objetivas PSNR, PSNR_e, SSIM, SSIM_e, 3-SSIM, MS-SSIM, B-SSIM e PW-SSIM. As métricas que proporcionaram melhores resultados na comparação com a métrica subjetiva foram a PW-SSIM



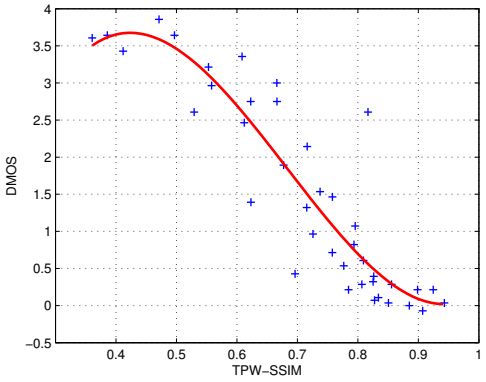
(a) DPSNR



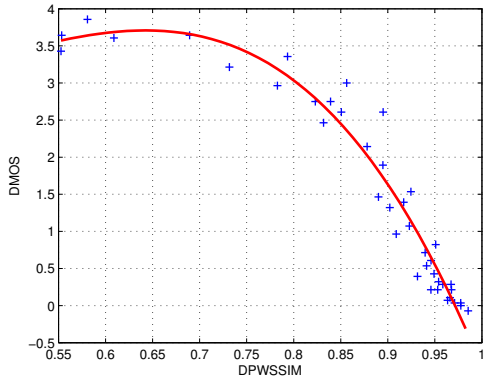
(b) DSSIM



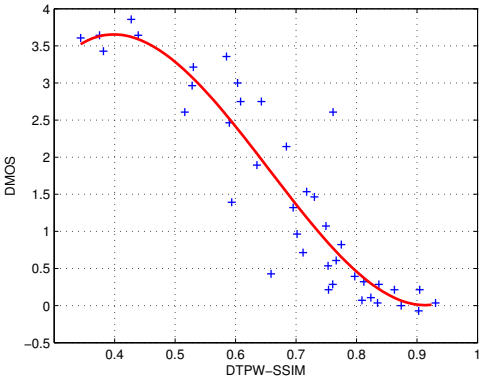
(c) PW-SSIM



(d) TPW-SSIM



(e) DPW-SSIM



(f) DTPW-SSIM

Figura 6.12: O comportamento entre as medidas do algoritmo proposto e os escores das experiências subjetivas descritas no NAMA3DS1-COSPAD1, para os vídeos codificados.

para a degradação Sal & Pimenta, B-SSIM, para a degradação por borramento, e a 3-SSIM, para a degradação de blocagem.

A segunda etapa avaliou a métrica PW-SSIM usando a base de dados LIVE. A partir dos testes como essa base de dados uma nova métrica foi proposta, a TPW-SSIM. Essa métrica obteve resultados melhores ou próximos dos resultados das melhores métricas encontradas na literatura.

Para vídeos 3D foi utilizado a base de dados NAMA3DS1-COSPAD1. Essa base utiliza vídeos codificados com H.264 e JPEG. Com os vídeos dessa base foram avaliados os algoritmos PSNR, SSIM, PW-SSIM, TPW-SSIM, IQSSA e PQM, e o melhor resultado foi do algoritmo PW-SSIM. Também foram avaliadas as métricas com a disparidade incluída em seu algoritmo, o que gerou uma melhoria na avaliação dos vídeos. Assim, a métrica que obteve o melhor resultado para essa base de dados foi a DPW-SSIM.

Capítulo 7

Conclusão e Propostas para Trabalhos Futuros

Esta tese abordou a avaliação de vídeo por meio de métricas de qualidade objetiva. A escolha dessas métricas se deve ao fato delas fornecerem os resultados das avaliações de forma mais rápida e com menor custo computacional em relação às métricas subjetivas. Um outro fator importante para o estudo desse tema reside no fato de que na última década várias pesquisas foram realizadas em busca da métrica objetiva que obtenha os melhores resultados, levando em consideração o sistema visual humano.

As métricas de avaliação objetiva são classificadas de três formas: com referência total, referência parcial e sem referência. Nesta tese foram propostas métricas com referência total para vídeos 2D e 3D, partindo da ideia de avaliar os vídeos colocar um peso maior nas suas áreas mais importantes.

Foram usadas nas novas métricas a informação da percepção espacial e a temporal para vídeos 2D, e para vídeos 3D, foi acrescentada a disparidade, para que se conseguisse escolher as regiões de maior interesse nos diferentes vídeos.

Para vídeos 2D foram apresentadas cinco métricas, PSNR_e, SSIM_e, B-SSIM, PW-SSIM e TPW-SSIM. As métricas PSNR_e e SSIM_e classificam blocos, 8×8, pelo nível de importância e avaliam apenas os blocos considerados mais importantes, usando as métricas PSNR e SSIM.

A métrica B-SSIM é proposta a partir de uma modificação da métrica SSIM. Essa modificação permite que vídeos borrados possam ser avaliados de forma mais condizente com o procedimento observado pelo sistema visual humano.

As métricas PW-SSIM e TPW-SSIM diferem uma da outra porque a TPW-SSIM

avalia, além da informação espacial, a informação temporal. Essas duas métricas obtiveram bons resultados, com um destaque maior para a TPW-SSIM que obteve resultados melhores do que a maioria das métricas testadas nesta tese.

Para avaliar vídeos 3D foi sugerido que se acrescentasse às métricas PSNR, SSIM, PW-SSIM e TPW-SSIM a avaliação da disparidade. O objetivo é poder avaliar melhor os vídeos 3D, uma vez que estes se destacam mais quando têm maior profundidade.

Os resultados obtidos para vídeos 3D indicam que as métricas IQSSA e PQM têm resultados próximos aos obtidos pela SSIM e PSNR e a métrica que mais se destacou foi a DPW-SSIM, que avalia a informação espacial e a disparidade.

7.1 Contribuições Mais Relevantes

As seguintes contribuições podem ser destacadas:

1. Os expectadores focalizam a visão em partes de um vídeo. Assim, avaliar apenas as partes de maior destaque aumenta a correlação das medidas, como foi provado pelos resultados obtidos pelas métricas PSNR_e e SSIM_e.

2. A métrica B-SSIM usa a informação espacial para avaliar os vídeos borrados, uma vez que vídeos com o artefato do borramento geram um aumento da informação espacial. Para esse artefato esta métrica obteve o melhor resultado.

3. A avaliação de vídeos usando a informação temporal, como apresentado em Vu & Deshpande (2012), tem um bom resultado quando a informação espacial do vídeo tem uma forte correlação com a informação temporal, como é apresentado na base de dados LIVE. Para bases de dados nos quais os vídeos têm baixa correlação entre a informação espacial e temporal, esta forma de avaliar não apresenta bons resultados.

4. A métrica PW-SSIM apresenta ganho sobre aquelas mais utilizadas nos dias atuais (PSNR, SSIM e VQM) para o efeito da codificação com H.264. Essa métrica também obteve o melhor resultado para vídeos 3D, em relação a todas as métricas testadas.

5. A métrica TPW-SSIM produziu resultados melhores que as métricas MOVIE e a desenvolvida por Vu & Deshpande (2012), para a base de dados LIVE.

6. Na avaliação de vídeos 3D, a adição da disparidade melhorou as métricas para vídeos 2D em até 0,1 na correlação. Todas as métricas que incluíram a disparidade obtiveram resultados melhores que as métricas originais, o que representa uma solução fácil de ser implementada e com ganhos satisfatórios.

7.2 Propostas de Trabalhos Futuros

Podem ser apontadas as seguintes propostas de continuação do trabalho:

1. Avaliar o efeito da informação espacial e temporal em métricas de avaliação de vídeo objetiva sem referência. As métricas sem referência podem ajudar a melhorar a qualidade de sistemas em tempo real. Uma vez que, se for possível avaliar o vídeo em tempo real, pode-se modificar os codificadores para que o vídeo sempre seja recebido com a qualidade esperada.

2. Encontrar uma relação entre a informação espacial e temporal dos vídeos, para que se possa avaliar vídeos 3D usando ambas as informações.

3. Os pesquisadores que trabalham com codificadores de vídeo 3D, na sua maioria, utilizam a PSNR para avaliar seus codificadores. Isso pode ocorrer por dois motivos: o primeiro porque a PSNR é a métrica mais simples e segundo pela falta de aplicativos que realizem a avaliação dos vídeos com diferentes métricas. Assim, preparar um aplicativos com as melhores métricas pode facilitar as pesquisas na área.

4. Uma outra proposta seria gerar uma métrica de avaliação de vídeos que sofreram fusão de imagens. Para a realização dessa métrica, as métricas PSNR_e, SSIM_e podem ser a base para avaliar apenas as partes dos vídeos.

5. A avaliação da complexidade computacional das métricas pode contribuir também para a escolha da melhor métrica a ser utilizada, umas vez que alguns sistemas têm limitação de *hardware* ou limitação de tempo.

7.3 Lista de Publicações Geradas

Capítulo de livro:

1. REGIS, C. D. M., OLIVEIRA, J. F. Televisão em Três Dimensões. Livro Televisão Digital, Autor: Marcelo Sampaio de Alencar, segunda edição, Editora Érica, 2012.

Artigos aceitos para publicação em periódicos:

1. REGIS, C. D., CARDOSO, J. V., OLIVEIRA, I.P., ALENCAR, M. S. Fuzzy Logic and Temporal Information Applied to Video Quality Assessment. Journal of Mobile Multimedia, 2013.

Artigos publicados em periódicos:

1. CARDOSO, J. V., MARIANO, A. C. S., REGIS, C. D. M., ALENCAR, M. S. Comparação das Métricas Objetivas de Qualidade de Vídeos Baseadas na Similaridade Estrutural e na Sensibilidade ao Erro. Revista de Tecnologia da Informação e Comunicação (RTIC), ISSN: 2237-5112, 2012.

2. CRUZ, J. N., COSTA, L.A.C., FERREIRA, M.W.F., SOUSA, V.F., REGIS, C. D. M.Comparação de Diferentes Métodos de Transcodificação Espacial de Vídeo Digital Utilizando Wavelets. Revista Principia, Ano 13, N 19, 2011.

3. REGIS, C. D. M., ROCHA, R. B., FARIAS, M. C. Q., ALENCAR, M. S. Objective and Subjective Evaluation of Space-Transcoded Videos for Mobile Receivers. Journal of Communications Software and Systems (JCOMMS), v. 6, p. 49-55, 2010.

Artigos publicados em anais de eventos:

1. REGIS, C. D., CARDOSO, J. V., ALENCAR, M. S. Effect of Visual Attention Areas on the Objective Video Quality Assessment. XVIII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia 2012), 2012

2. REGIS, C. D., CARDOSO, J. V., OLIVEIRA, I.P., ALENCAR, M. S. Performance of Objective Video Quality Metrics with Perceptual Weighting Considering First and Second Order Differential Operators. XVIII Simpósio Brasileiro de Sistemas Multimídia e Web (WebMedia 2012), 2012

3. REGIS, C. D., MOTA, M., LOPES, R. F., ALENCAR, M. S. Performance Analysis of a 64-QAM Rotated Constellation in Rice Fading Channels. XXX Simpósio Brasileiro de Telecomunicações (SBrT'12), 2012.

4. CARDOSO, J. V. REGIS, C. D., ALENCAR, M. S. B-SSIM: Structural Similarity Index for Blurred Videos. XXX Simpósio Brasileiro de Telecomunicações (SBrT'12), 2012.

5. REGIS, C. D., CARDOSO, J. V., ALENCAR, M. S. Improved Estimation of the Video Quality Based on the Effect of Spatial Perceptual Information. XXX Simpósio Brasileiro de Telecomunicações (SBrT'12), 2012.

6. LOPES, R. F., REGIS, C. D. M., LOPES, W. T. A., ALENCAR, M. S. Adapt-VoD - An Adaptive Video-on-Demand Platform for Mobile Devices. 5th FTRA *International Conference on Multimedia and Ubiquitous Engineering* (MUE 2011), 2011, Crete, Greece. p. 257-262.

7. LOPES, R. F., REGIS, C. D. M., QUEIROZ, W. J. L., LOPES, W. T. A., ALENCAR, M. S. Effect of Channel Estimation Errors on Adaptive Modulation Systems Subject to Rayleigh Fading. XXIX Simpósio Brasileiro de Telecomunicações (SBrT'11), 2011, Curitiba, PR.

8. ROCHA, R. B., REGIS, C. D. M., ALENCAR, M. S. Subjective and Objective Evaluation of Transcoded Video Quality after Transmission. IWT'11 – *International Workshop on Telecommunications*, 2011, Rio de janeiro.

9. SANTOS, M. O., ALCANTARA, E. C. S., REGIS, C. D. M. Geração e visualização de videos estereoscópicos. Encontro Anual do Iecom em Comunicações, Redes e Criptografia, 2011, Campina Grande.

10. LOPES, R. F., REGIS, C. D. M., SOUSA, M. P., LOPES, W. T. A., ALENCAR, M. S. FAST: A Fuzzy Adaptive Spatial Video Transcoding System. 13th *International Symposium on Wireless Personal Multimedia Communications* (WPMC 2010), 2010, Recife, PE.

11. LOPES, R. F., REGIS, C. D. M., LOPES, W. T. A., ALENCAR, M. S. Uma Análise Comparativa de Técnicas de Mapeamento de Símbolos em Constelações de Sinais. *International Information and Telecommunication Technologies Symposium* (I2TS 2010), 2010, Rio de Janeiro.

12. LOPES, R. F., REGIS, C. D. M., LOPES, SILVA, E. F., CORTES, O. A. C., LOPES, W. T. A., ALENCAR, M. S. VoDTV - A VoD Platform for the ISDB-

Tb Digital Television System (invited paper). 1st *Brazil Japan Symposium on Advances in Digital Television*, São Paulo, SP, 2010.

13. ROCHA, R. B., REGIS, C. D. M., LOPES, W. T. A., ALENCAR, M. S. Analysis of the Quality of Transcoded Videos After the Transmission Over a Noisy Channel. 13th *International Symposium on Wireless Personal Multimedia Communications* (WPMC 2010), 2010, Recife, PE.

14. REGIS, C. D. M., ROCHA, R. B., LOPES, W. T. A., ALENCAR, M. S. On the Effects of Spatial Transcoding in a Video Transmission System. 13th *International Symposium on Wireless Personal Multimedia Communications* (WPMC 2010), 2010, Recife, PE.

15. REGIS, C. D. M., MORAIS, D. C., FARIAS, M. C. Q., ALENCAR, M. S. Assessment of Spacial Video Transcoding Based on Structural Distortion Measurement. *International Workshop on Telecommunicatios* IWT'09, v. 1, p. 230-234, 2009.

16. REGIS, C. D. M., MORAIS, D. C., ALENCAR, M. S., FARIAS, M. C. Q. Assessment of the Error Caused by the Encoding of Spatially Transcoded Videos to Be Used in Mobile TV. The 12th *International Symposium on Wireless Personal Multimedia Communications*, v. 12, p. 1569204419, 2009.

17. REGIS, C. D. M. MORAIS, D. C. FARIAS, M. C. Q. ALENCAR, M. S. Objective and Subjective Assessment of Space-Transcoded Videos for Mobile Receivers. IEEE *International Symposium on Broadband Multimedia Systems and Broadcasting*, 2009, Bilbao.

18. REGIS, C. D. M., MORAIS, D. C., ROCHA, R. B., ALENCAR, M. S., FARIAS, M. C. Q. Video Quality Issues for Mobile Television. 15th *International Conference on Distributed Multimedia Systems*, 2009, San Francisco.

19. OLIVEIRA, J., REGIS, C. D. M., ALENCAR, M. S. An Experimental Evaluation of the Mobile Channel Performance of the Brazilian Digital Television System. 15th *International Conference on Distributed Multimedia Systems*, 2009, San Francisco.

20. REGIS, C. D. M., ROCHA, R. B., SOUSA, M. P., ALENCAR, M. S. Transmissão de Vídeos Transcodificados com o Filtro da Média. 8th *International Information*

and Telecommunication Technologies Symposium, 2009, Florianópolis.

21. OLIVEIRA, J. F., REGIS, C. D. M., ALENCAR, M. S. Avaliação Experimental do Desempenho do Canal Móvel do Sistema Brasileiro de Televisão Digital. 8th *International Information and Telecommunication Technologies Symposium*, 2009, Florianópolis.

22. SOUSA, M. P., REGIS, C. D. M., ALENCAR, M. S., LOPES, W. T. A. Diversidade Cooperativa Adaptativa Aplicada a Redes de Sensores sem Fio. 8th *International Information and Telecommunication Technologies Symposium*, 2009, Florianópolis.

Apêndice A

LIVE Video Quality Database

Essa base de dados é composto por 10 vídeos de cenas naturais sem compressão e 150 vídeos distorcidos (obtidos a partir das referências), com quatro tipos diferentes de distorções. Cada vídeo foi avaliado por 38 pessoas em um estudo de único estímulo com remoção da referência, em que os avaliadores marcaram a qualidade do vídeo em uma escala de qualidade contínua (Seshadrinathan *et al.*, 2010b), (Seshadrinathan *et al.*, 2010a).

A base de dados de vídeo LIVE inclui vídeos distorcidos pela compressão H.264 e MPEG-2, bem como vídeos resultantes da simulação da transmissão de fluxos de vídeos H.264 por meio de canais de comunicação com erros, sem fio e IP. Os vídeos do LIVE foram todos capturados em formatos de varredura progressiva, permitindo o desenvolvimento de algoritmos de avaliação de qualidade de vídeo. A base de dados abrange vários níveis de qualidade, sendo os vídeos de baixa qualidade projetados para serem de qualidade semelhante aos encontrados em aplicativos de *streaming* de vídeo na Internet (Youtube, vídeos sem fio, *streaming* ao vivo de vídeos com baixa largura de banda, etc.).

A.1 Sequências Fontes

Foram utilizados 10 vídeos descomprimidos, de alta qualidade, vídeos fontes de cenas naturais (em oposição à animação, texto e gráficos) que estão disponíveis gratuitamente para *download* a partir da Universidade Técnica de Munique (Technical University of Munich, 2012). Todos os vídeos foram filmados com equipamento profissional e convertidos para o formato digital, garantindo que os vídeos são de referência,

livres de distorções.

Só foram usados os vídeos progressivos nessa base de dados, evitando assim problemas com o *de-interlacing*. Foram usados os vídeos digitais fornecidos em Alta Definição (HD) no formato 4:2:0 YUV e nenhum dos vídeos contêm componentes de áudio. No entanto, devido a limitações de recursos na exibição desses vídeos, foi realizado o *downsampling* de todos os vídeos a uma resolução de 768 $\times$ 432 *pixels*, escolhida para assegurar que a razão de aspecto dos vídeos HD fosse mantida, minimizando assim as distorções visuais. Além disso, esta resolução assegura que o número de linhas e colunas são múltiplos de 16, como é usual para sistemas de compressão, tais como MPEG-2. O *downsampled* foi realizado em cada quadro do vídeo usando o *imresize*, função do Matlab, usando interpolação bicúbica para minimizar as distorções devido ao *aliasing*.

A Figura A.1 é apresentado um quadro de cada vídeo de referência da base de dados Live e na Figura A.2 a relação entre a informação espacial e temporal. Todos os vídeos, exceto o *blue sky*, são de 10 segundos de duração. A sequência do *blue sky* tem 8,68 segundos de duração. As primeiras sete sequências têm taxa de 25 quadros por segundo, enquanto os três restantes (*Park run, Shields, and Mobile & Calendar*) têm taxa de 50 quadros por segundo. Uma breve descrição desses vídeos é fornecida a seguir:

- *Blue Sky* – movimento de câmera circular que mostra um céu azul e algumas árvores;
- *River Bed – Still camera*, mostra um leito de rio contendo algumas pedras e água;
- *Pedestrian area – Still camera*, mostra algumas pessoas caminhando em uma intersecção de ruas;
- *Tractor – Camera pan*, mostra um trator em movimento em um campo;
- *Sunflower – Still camera*, mostra uma abelha se movendo em torno de um girassol;
- *Rush hour – Still camera*, mostra o tráfego na hora do *rush* em uma rua;
- *Station – Still camera*, mostra um trilho, um trem e algumas pessoas em pé;
- *Park run – Camera pan*, uma pessoa correndo em um parque;
- *Shields – Camera pan* no início, depois *still* e zoom; mostra uma pessoa que anda e aponta para a tela;

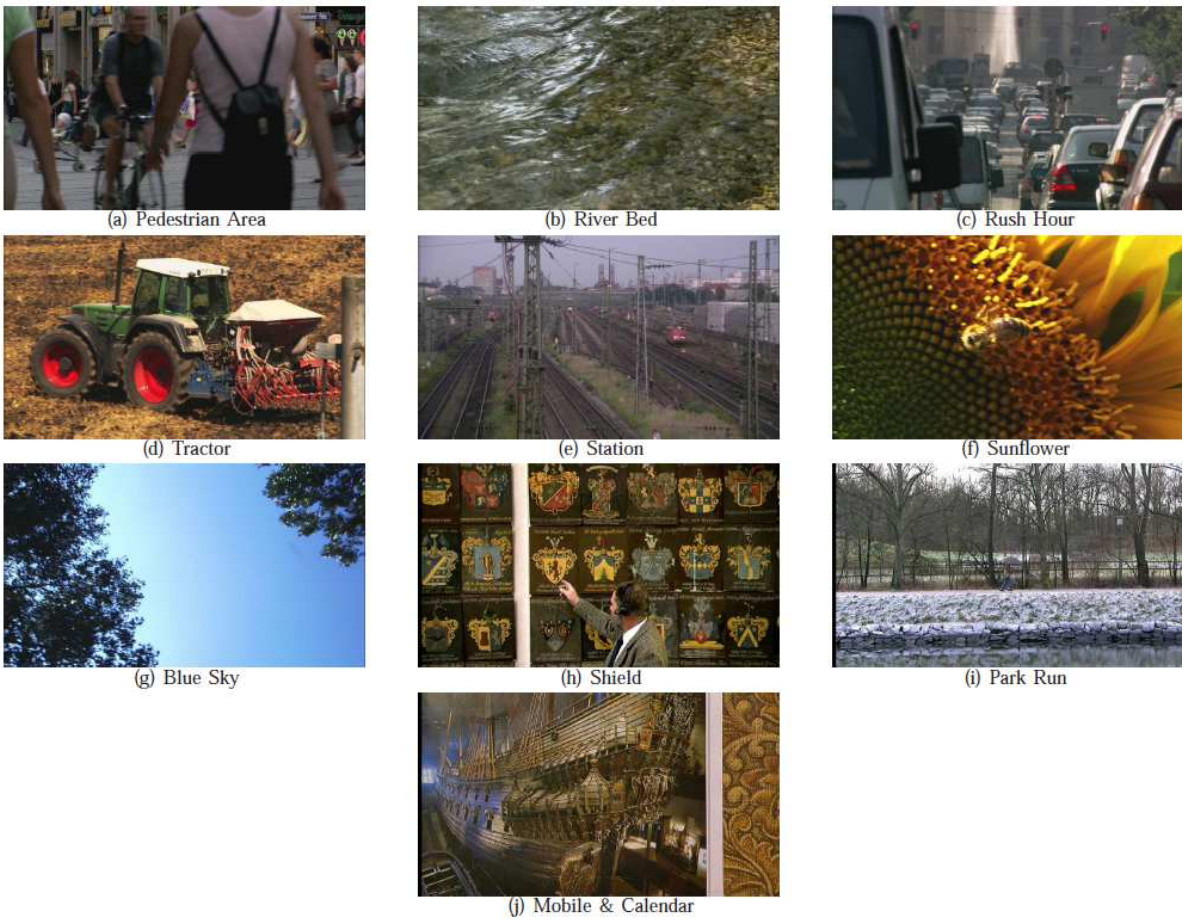


Figura A.1: Um quadro de cada um dos dez vídeos de referência utilizado no estudo (Seshadrinathan *et al.*, 2010a).

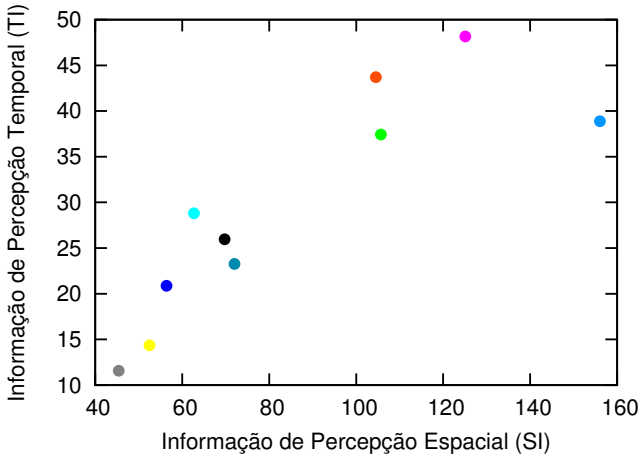


Figura A.2: Diversidade espacial e temporal das sequências de vídeos disponibilizadas na base de dados LIVE.

- *Mobile & Calendar – Camera pan*, trem de brinquedo movendo-se horizontalmente com um calendário que se move verticalmente ao fundo.

A.2 Sequências de teste

Foram criadas 15 sequências de teste de cada vídeo de referência, utilizando quatro processos diferentes de distorção – compressão MPEG-2 e H.264, transmissão simulada de *bitstreams* em vídeo comprimido com H.264 por uma rede propensa a erros IP e por outra propensa a erros de redes sem fio.

Os sistemas de compressão, como MPEG-2 e H.264, produzem distorções distribuídas uniformemente na qualidade do vídeo, tanto espacialmente quanto temporalmente (Seshadrinathan *et al.*, 2010a). Perdas na rede, no entanto, causam distorções transitórias no vídeo, tanto espacial quanto temporal. Na Figura A.3 é apresentado um quadro da sequência corrompida por cada um dos tipos de distorção da base de dados LIVE.

Figura A.3: Representação das degradações em um Quadro para a base de dados LIVE.(a) Quadro comprimido com MPEG-2 (b) Quadro comprimido com H.264 (c) Quadro gerado a partir da simulação de uma rede IP (d) Quadro gerado a partir da simulação de uma rede sem fio (Seshadrinathan *et al.*, 2010a).

Os vídeos MPEG-2 e H.264 exibem os artefatos de compressão típicas, como a blocagem, o borramento e a incompatibilidade de compensação de movimento ao redor

das bordas do objeto.

Vídeos obtidos a partir da transmissão com perdas em redes sem fio apresentam erros que são restritos a pequenas regiões de um quadro. Vídeos obtidos a partir de transmissão com perdas em redes IP exibem erros em regiões maiores do quadro. Erros em redes sem fio e IP também são transitórios e aparecem como falhas no vídeo.

As distorções foram ajustadas manualmente pelos autores da base de dados LIVE. Um conjunto grande de vídeos foi gerado e visualizado pelos autores e um subconjunto desses vídeos foi escolhido para ser incluído na base de dados LIVE. Para ilustrar este procedimento, considera-se quatro classificações para a qualidade visual ("Excelente", "Bom", "Razoável"e "Pobre") e um vídeo de referência ("Tractor"). Quatro versões do vídeo "Tractor"compactadas com MPEG-2 foram escolhidas de forma que coincidam com as quatro qualidades visuais escolhidas. Procedimento semelhante é aplicado para selecionar para as versões do H.264, sem fio e IP. Note que o vídeo "Excelente"do MPEG-2 e H.264 é projetado para ter a mesma qualidade visual aproximada, assim como para outras categorias de distorções e qualidades visuais.

Vídeos com qualidade "Excelente"foram escolhidos para estarem bem perto da referência em qualidade visual. Vídeos com qualidade "Pobre"foram escolhidos para terem qualidade semelhante a vídeos do Youtube.

O procedimento de seleção é então repetido para cada vídeo de referência. Os vídeos distorcidos têm o intuito de testar a capacidade das métricas objetivas em prever a qualidade visual de forma consistente em diferentes tipos de conteúdo e distorção.

A.2.1 MPEG-2

O padrão MPEG-2 é usado em uma ampla variedade de aplicações de vídeo, mais notavelmente em DVDs e na transmissão de televisão digital. Para a codificação dos vídeos com MPEG-2 foi utilizado o *software* de referência MPEG-2, disponível a partir da *International Organization for Standardization* (IOS) (ISO/IEC, 2005). As taxas de codificação variaram de 700 k$bits$/s a 4 M$bits$/s, dependendo da sequência de referência.

A.2.2 H.264

O H.264 está rapidamente ganhando popularidade devido à sua eficiência de compressão superior em relação ao MPEG-2. Foi Usado o *software* de referência JM (versão

12.3) disponibilizado pelo *Joint Video Team* (JVT) (Tourapis, 2012). As taxas de codificação variaram de 200 k*bits*/s a 5 Mbps.

A.2.3 Transmissão em Redes IP

Os vídeos são muitas vezes transmitidos por redes IP em aplicações tais como telefonia e vídeo conferência, IPTV e vídeo sob demanda. Os fluxos de vídeo H.264 foram criados com taxas de codificação variando de 0,5 a 7 M*bits*/s.

Um estudo aprofundado do transporte de vídeo H.264 sobre redes IP pode ser encontrado em Wenger (2003) e muitas das considerações foram realizadas com base neste estudo. Nas redes IP as perdas de pacotes ocorrem principalmente devido a um estouro no *buffer*, nos nós intermediários, em uma rede congestionada.

A.2.4 Transmissão por meio de redes sem fio

A transmissão de vídeo para terminais móveis foi concebida para ser uma grande aplicação em sistemas 3G e a eficiência de compressão H.264 torna-a mais indicada para uso na transmissão sem fio em ambientes hostis (Stockhammer *et al.*, 2003). As taxas de compressão variaram entre 0,5 a 7 M*bits*/s.

Um pacote transmitido por um canal sem fio é suscetível a erros de *bit* devido à atenuação, sombreamento, interferência e desvanecimento em canais sem fios. Foi assumido que um pacote é perdido, mesmo que ele contenha apenas um *bit* errado.

Para realizar a simulação de uma rede sem fio foi utilizado o *software* disponível no VCEG (*Video Coding Experts Group*) (VCEG, 1999). As taxas de erro de pacotes usadas variaram entre 0,5-10%. A decodificação e as técnicas de ocultação de erro para as simulações sem fio foram idênticas às simulações IP.

A.3 Projeto dos testes subjetivos

Para a realização dos testes subjetivos, foi adotado o procedimento de único estímulo contínuo para obter as avaliações subjetivas para as sequências de vídeo diferentes. Essa escolha é adequada para um grande número de aplicações multimídia, como controle de qualidade para vídeo sob demanda, IPTV, Internet e etc. Além disso, reduz o tempo necessário para conduzir o estudo em comparação com um estudo de estímulo duplo (Seshadrinathan *et al.*, 2010a).

Os avaliadores indicaram a qualidade do vídeo em uma escala contínua e observaram todos os vídeos, o que exigiu o tempo de uma hora para cada avaliador realizar o teste. Para minimizar os efeitos da fadiga, a avaliação foi realizada em duas sessões de trinta minutos cada.

Os cento e cinquenta vídeos foram apresentados sem ordem definida usando um gerador de números aleatórios. Esta lista foi então dividida em duas partes, para a realização das duas sessões. Para evitar que o avaliador seja mais crítico em uma seção do que em outra, foi incluído o vídeo de referência em ambas as sessões. Foi inserido cada um dos dez vídeos de referência, nas listas em cada sessão de forma aleatória. As pontuações DMOS foram então calculadas para cada vídeo por sessão, utilizando o índice de qualidade atribuído ao vídeo de referência nessa sessão.

A.4 Exibição dos testes subjetivos

Os participantes da base de dados LIVE desenvolveram a interface do usuário para o estudo em um PC com Windows usando Matlab, em conjunto com a caixa de ferramentas XGL para Matlab desenvolvida na Universidade do Texas em Austin (Mehlitz, 2008). A caixa de ferramentas XGL permite a apresentação precisa de estímulos psicofísicos para observadores humanos.

Os vídeos foram vistos pelos sujeitos em um monitor CRT, para evitar os efeitos de desfoque de movimento e baixa taxa de atualização do monitor de LCD. O estudo inteiro foi realizado utilizando o mesmo monitor, que foi calibrado usando o *Monaco Optix XR Pro device*.

A tela foi fixada em uma resolução de 1024×768 *pixels* e os vídeos exibidos em sua resolução nativa para evitar distorções, devido a operações de escala realizadas pelo *software* ou *hardware*. As demais áreas da tela eram escuras. No final da apresentação do vídeo, uma escala contínua para uma qualidade de vídeo foi visualizada na tela, com um cursor fixado no centro da escala de qualidade para evitar a polarização da qualidade percebida. A escala de qualidade teve cinco pontos marcados para ajudar o avaliador. A extremidade esquerda da escala foi marcada como "Péssimo"e na extremidade direita foi marcada como "Excelente". Os outros três pontos foram igualmente espaçados e marcados como "Ruim", "Regular"e "Bom", semelhante à escala ITU-R ACR.

A.5 Avaliadores e Treinamento

Todos os avaliadores que participaram do estudo foram recrutados a partir da graduação e da turma de processamento digital de vídeo da Universidade do Texas em Austin. O grupo constituiu de 38 avaliadores, sendo a amostra composta principalmente de estudantes do sexo masculino.

Cada avaliador foi individualmente informado sobre o objetivo do experimento e assistiu uma curta sessão de formação antes de iniciar o experimento. A primeira sessão foi composta por seis vídeos de treinamento, sendo que três vídeos de treinamento foram usados em sua segunda sessão. Os participantes foram solicitados a também fornecer índices de qualidade para os vídeos de treinamento para se familiarizarem com o procedimento de teste. Os vídeos de treinamento não faziam parte da base de dados e havia conteúdos diferentes. Os vídeos de treinamento foram de 10 segundos de duração e também foram degradados pelas mesmas distorções dos vídeos de teste.

A.6 Tratamento das notas subjetivas

Como o foco na base de dados LIVE é testar os algoritmos de avaliação objetiva com referência completa, que assumem um vídeo de referência, sem degradação , foi calculada a diferença na pontuação entre o vídeo de teste e o de referência correspondente, para descontar as preferências dos usuários para vídeos de referência determinados.

Assim, s_{ijk} define a pontuação atribuída pelo avaliador i no vídeo j da sessão $k = \{1, 2\}$. A diferença das pontuações d_{ijk} foi calculada por sessão, subtraindo a qualidade atribuída pelo avaliador ao vídeo de referência pela avaliação do vídeo sob teste da mesma sessão,

$$d_{ijk} = s_{ij_{ref}k} - s_{ijk}. \tag{A.1}$$

Foram removidas as avaliações quando o d_{ijk} foi igual a "0" em ambas as sessões. Os valores das diferenças por sessão são convertidas em z (Dijk *et al.*, 1995)

$$z_{ijk} = \frac{d_{ijk} - \mu_{ik}}{\sigma_{ik}}, \tag{A.2}$$

em que μ_{ik} e σ_{ik} são a média e o desvio padrão dos d_{ijk}.

O procedimento de rejeição do usuário especificado na ITU-R BT 500.11 foi usado para descartar as notas dos indivíduos não confiáveis (ITU-R, 2002). Primeiramente o padrão recomenda determinar se as notas atribuídas por um avaliador são normalmente distribuídos pelo cálculo dos valores da curtose (β_2).

As pontuações são consideradas normalmente distribuídas se o curtose cai entre os valores 2 e 4. O procedimento rejeita um avaliador sempre que mais do que 5% dos pontos atribuídos por ele estão acima de duas vezes o desvio padrão. Se os resultados não são normalmente distribuídos, o avaliador é rejeitado sempre que mais do que 5% das suas pontuações cai fora do intervalo. Nesta base de dados, nove dos trinta e oito indivíduos foram rejeitados nesta fase.

A.7 Desempenho dos Modelos Objetivos

O PLCC e SROCC foram calculados após a realização de uma regressão não linear nos valores obtidos pelos resultados da avaliação objetiva dos vídeos, utilizando uma função logística de quatro parâmetros monotônicos para se ajustar os índices da predição objetiva aos da qualidade subjetivas. Essa função é dada por (Seshadrinathan *et al.*, 2010a)

$$Q_k^{'} = \beta_2 + \frac{\beta_1 - \beta_2}{1 + e^{-\left(\frac{Q_k - \beta_3}{\mid \beta_4 \mid}\right)}}, \tag{A.3}$$

na qual Q_k representa a avaliação da qualidade objetiva do vídeo para os k-ésimos vídeos da base de dados LIVE.

A otimização é realizada utilizando a função do Matlab `nlinfit`, para encontrar o parâmetro ótimo que miniminize o erro quadrático entre a lista dos valores subjetivos (DMOS_k) e a lista dos valores de pontuação objetivos ajustados ($Q_k^{'}$). As estimativas iniciais do parâmetro foram (VQEG, 2000): $\beta_1 = \max(\mathrm{DMOS}_k)$, $\beta_2 = \min(Q_k^{'})$, $\beta_3 = \mathrm{mean}(Q_k^{'})$ e $\beta_4 = 1$. A função do Matlab `nlpredci` foi usada para obter a predição dos valores do DMOS.

Em Seshadrinathan *et al.* (2010a) são apresentados os resultados da correlação para diferentes métricas de avaliação objetiva: SSIM (Wang *et al.*, 2004), MS-SSIM (Wang *et al.*, 2003), Speed SSIM, VSNR (Chandler & Hemami, 2007), V-VIF (Sheikh & Bovik, 2004), VQM, S-MOVIE, T-MOVIE e MOVIE (*Video quality Metric Structural*

Similarity Index) (Seshadrinathan & Bovik, 2010). Os que obtiveram os melhores resultados (Tabelas A.1 e A.2) foram os S-MOVIE, T-MOVIE e MOVIE, que têm a desvantagem de levar aproximadamente cinco horas para calcular a qualidade de um vídeo com 250 quadros e resolução espacial de 768 × 432 (Vu & Deshpande, 2012).

Tabela A.1: Avaliação das métricas existentes usando o PLCC na base de dados LIVE (Seshadrinathan *et al.*, 2010b).

Algoritmos	**H.264**	**IP**	**Sem fio**	**MPEG-2**	**Todos**
PSNR	0,5492	0,4645	0,6690	0,3891	0,5621
SSIM	0,6656	0,5119	0,5401	0,5491	0,5444
MS-SSIM	0,6919	0,7219	0,7170	0,6604	0,7441
Speed SSIM	0,7206	0,5587	0,5867	0,6270	0,5962
VSNR	0,6216	0,7341	0,6992	0,5980	0,6896
VQM	0,6459	0,6480	0,7325	0,7860	0,7236
V-VIF	0,6911	0,5102	0,5488	0,6145	0,5756
S-MOVIE	0,7252	0,7378	0,7883	0,6587	0,7451
T-MOVIE	**0,7920**	0,7383	**0,8371**	**0,8252**	**0,8217**
MOVIE	0,7902	**0,7622**	**0,8386**	0,7595	0,8116

Tabela A.2: Avaliação das métricas existentes usando o SROCC na base de dados LIVE (Seshadrinathan *et al.*, 2010b).

Algoritmos	**H.264**	**IP**	**Sem fio**	**MPEG-2**	**Todos**
PSNR	0,4585	0,4167	0,6574	0,3862	0,5398
SSIM	0,6514	0,4550	0,5233	0,5545	0,5257
MS-SSIM	0,7051	0,6534	0,7285	0,6617	0,7361
Speed SSIM	0,7086	0,4727	0,5630	0,6185	0,5849
VSNR	0,6460	0,6894	0,7019	0,5915	0,6755
VQM	0,6520	0,6383	0,7214	0,7810	0,7026
V-VIF	0,6807	0,4736	0,5507	0,6116	0,5710
S-MOVIE	0,7066	0,7046	0,7927	0,6911	0,7270
T-MOVIE	**0,7797**	**0,7192**	**0,8114**	**0,8170**	**0,8055**
MOVIE	0,7664	0,7157	0,8109	0,7733	0,7890

Apêndice B

Base de dados NAMA3DS1-COSPAD1

A fim de proporcionar um conjunto de sequências de vídeo 3D degradadas para comparação, foi criada uma base de dados chamada NAMA3DS1-COSPAD1. Essa base de dados foi gerado em Nantes e Madrid e está restrito apenas a codificação e degradações espaciais, como codificação baseada em blocagem, codificação *wavelet*, redução da resolução e algoritmos de realce de borda.

Para a geração dessa base de dados foram realizadas as etapas de aquisição, codificação e avaliação subjetiva de vídeos 3D de alta qualidade. A metodologia ACR-HR (*Absolute Category Rating with Hidden Reference*) (ITU-T, 1999) foi adotada para a avaliação subjetiva, na qual os observadores foram orientados a aferir apenas a qualidade do vídeo.

Para a gravação das sequências de vídeo foram utilizados alguns critérios importantes, como a variabilidade de conteúdo e as diversidades espacial, temporal e de profundidade. Essas medidas garantem um maior interesse por parte dos espectadores em um experimento para avaliação subjetiva. Além disso, o conteúdo foi selecionado de modo que represente o conteúdo apresentado em programas televisivos convencionais, tais como cenas de esportes e jornalismo.

B.1 Sequências de Vídeo

As sequências (Figura B.1) foram capturadas com uma câmera Panasonic do modelo AG-3DA1E, que apresenta um par de lentes cujos eixos centrais estão separados de uma

(a) Barrier Gate (SRC1)

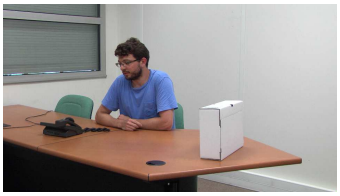
(b) Basket (SRC2)

(c) Boxers (SRC3)

(d) Hall (SRC4)

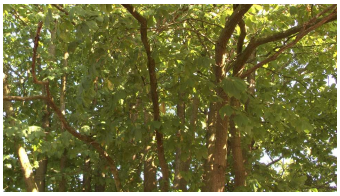
(e) Lab (SRC5)

(f) News Report (SRC6)



(g) Phone Call (SRC7)

(h) Soccer (SRC8)



(i) Tree Branches (SRC9)

(j) Umbrella (SRC10)

Figura B.1: Vista esquerda das sequências de vídeos em três dimensões disponibilizadas em Nantes-Madrid.

distância de 60 mm, que é aproximadamente a distância entre as pupilas dos olhos humanos. A distância focal proporcionada varia de $4,2$ mm até $23,5$ mm. Os vídeos foram gravados com uma resolução espacial de 1920×1080 *pixels* a uma frequência de 25 Hz. Para armazenar as sequências de vídeo, a câmera Panasonic AG-3DA1E realiza o processo de compressão utilizando o codificador H.264/AVC em configuração de High-Profile com uma taxa de *bits* de 24 Mbit/s.

As sequências de vídeos disponibilizadas foram quantificadas a partir da Informação da Percepção Espacial (SI) e Temporal (TI) como descrito na recomendação ITU-T P.910 (ITU-T, 1999). A Figura B.2 indica a heterogeneidade das fontes disponíveis na base de dados de vídeos NAMA3DS1-COSPAD1 (IRCCyN-IVC, 2012).

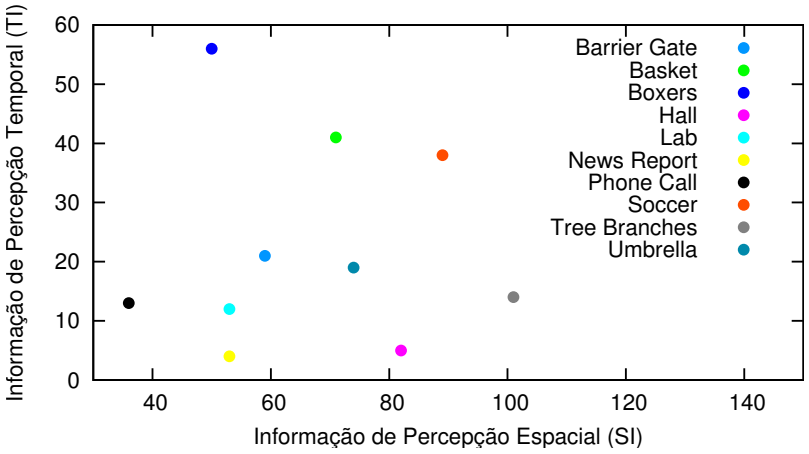


Figura B.2: Diversidade espacial e temporal das sequências de vídeos disponibilizadas em Nantes-Madrid.

B.2 Experimento Subjetivo

O experimento subjetivo realizado em Urvoy *et al.* (2012) teve como objetivo fornecer resultados em termos de Escores Médio de Opinião (*Mean Opinion Score* – MOS), para servir de suporte ao desenvolvimento de algoritmos objetivos de avaliação da qualidade de vídeos estereoscópicos.

As sequências fonte foram submetidas a dez tipos de degradações, dispostas na Tabela B.1.

Tabela B.1: Degradações utilizadas pela base de dados NAMA3DS1-COSPAD1.

Tipo	**Parâmetros**	Abreviação
Codificação da fonte (H.264)	QP32	SRC1
	QP38	SRC2
	QP44	SRC3
Codificação da fonte (JPG2K)	2 Mbit/s	SRC4
	8 Mbit/s	SRC5
	16 Mbit/s	SRC6
	32 Mbit/s	SRC7
Transcodificação Espacial	$\downarrow$ 4 *downsampling*	SRC8
Image Sharpening	Realce das bordas	SRC9
Transcodificação Espacial e *Image Sharpening*	Realce das bordas e $\downarrow$ 4 *downsampling*	SRC10

O conjunto de avaliadores foi constituído por 12 mulheres e 17 homens com idade entre 18 e 63 anos. Os avaliadores foram submetidos a testes de acuidade visual, capaci-

dade de distinção das cores e acuidade estéreo, de modo que os avaliadores selecionados foram aqueles que apresentaram aptidão em todos os testes supracitados.

Os avaliadores foram colocados em uma sala padronizada, na qual se utilizou um monitor Philips, com estereoscopia via óculos, de modelo 46PFL9705H com 46 polegadas para visualização das sequências. O brilho da tela foi ajustado em 180 cd/m^2, resultando em um brilho percebido de 56 cd/m^2 devido à atenuação gerada pelos óculos. A distância de visualização foi definida como sendo três vezes a altura do monitor, resultando em 172 cm.

Escores Médios de Opinião (MOS)

Os Escores Médios de Opinião (MOS), com um nível de confiança de 95%, para as amostras de vídeo de referência estão na Figura B.3. O intervalo de confiança em torno da média é dado por

$$(\text{MOS} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{\gamma}}, \text{MOS} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{\gamma}}), \tag{B.1}$$

em que MOS representa a média amostral, σ é o desvio padrão populacional, γ é o número de amostras e $z_{\alpha/2}$ é igual a $1,96$ para a construção de um intervalo de 95% de confiança.

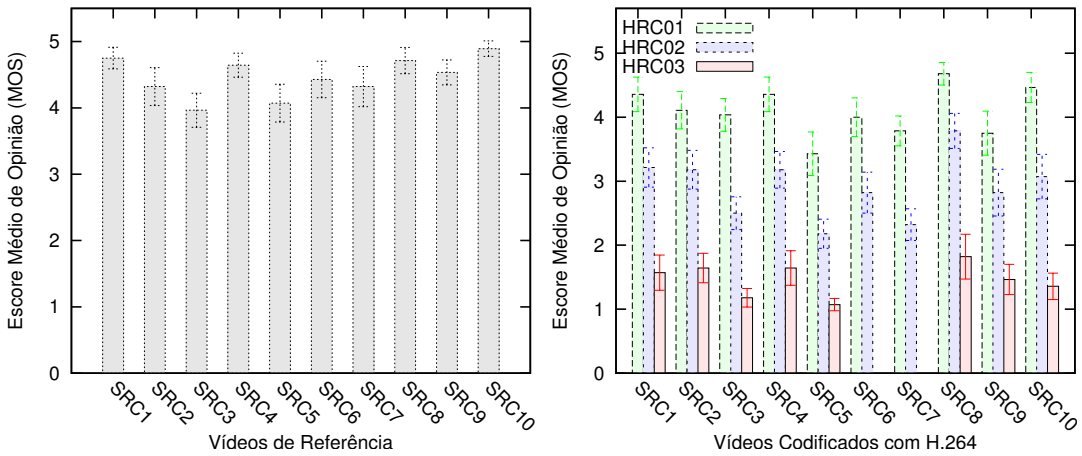


Figura B.3: MOS para as amostras de vídeos de referência com um nível de confiança de 0,95.

Bibliografia

Akamine, W. Y. L., & Farias, M. C. Q. 2012. Incorporating Visual Attention Models Into Image Quality Metrics. *In: Proceedings of the Sixth International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM).*

Alajel, K.M., & Xiang, Wei. 2012 (oct.). Color Plus Depth 3-D Video Transmission with Hierarchical 16-QAM. *Pages 1–4 of: 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON), 2012.*

Albini, F. L. P. 2009 (Março). *Geração e Avaliação de Artefatos em Vídeo Digital.* Dissertação de Mestrado, Universidade Tecnológica Federal do Paraná, Curitiba, Brasil.

Andrade, L. A., & Goularte, R. 2009. Percepção Estereoscópica Anaglífica em Vídeos Digitais Comprimidos com Perda. *XV Simpósio Brasileiro de Sistemas Multimídia e Web – WebMedia 2009.*

Arthur, Rangel. 2002 (Abril). *Avaliação Objetiva de Codecs de Vídeo.* Dissertação de Mestrado, Universidade Estadual de Campinas – UNICAMP, Campinas, Brasil.

Avcibas, I., Sankur, B., & Sayood, K. 2002. Statistical Evaluation of Image Quality Measures. *Journal of Electronic Imaging*, **11**, 206–223.

Benoit, A., Le Callet, P., Campisi, P., & Cousseau, R. 2008 (Oct.). Using Disparity for Quality Assessment of Stereoscopic Images. *Pages 389–392 of: 15th IEEE International Conference on Image Processing, ICIP 2008.*

Bernardino Júnior, Francisco Madeiro. 1998 (Março). *Quantização Vetorial Aplicada à Compressão de Sinais de Voz e Imagem.* Dissertação de Mestrado, Universidade Federal da Paraíba – Campus II, Campina Grande, Brasil.

Cardoso, J. V. M., Mariano, A. C. S., Regis, C. D. M., & Alencar, M. S. 2012. Comparação das Métricas Objetivas de Qualidade de Vídeos Baseadas na Similaridade Estrutural e na Sensibilidade ao Erro. *Pages 33–40 of: Revista de Tecnologia da Informação e Comunicação (RTIC).*

Chan, S.C., Shum, Heung-Yeung, & Ng, King-To. 2007. Image-Based Rendering and Synthesis. *IEEE Signal Processing Magazine*, **24**(6), 22–33.

Chandler, Damon M., & Hemami, Sheila S. 2007. VSNR: A Wavelet-Based Visual Signal-to-Noise Ratio for Natural Images. *IEEE Transactions on Image Processing*, **16**(9), 2284–2298.

Chen, Guan-Hao, Yang, Chun-Ling, Po, Lai-Man, & Xie, Sheng-Li. 2006a. Edge-Based Structural Similarity for Image Quality Assessment. *IEEE International Conference on Acoustics, Speech and Signal Processing.*

Chen, Guan-Hao, Yang, Chun-Ling, & Xie, Sheng-Li. 2006b (Oct.). Gradient-Based Structural Similarity for Image Quality Assessment. *Pages 2929–2932 of: IEEE International Conference on Image Processing.*

Cheolkon, J., & Jiao, L.C. 2011. Disparity-Map-Based Rendering for Mobile 3D TVs. *IEEE Transactions on Multimedia*, September.

Corbetta, M. 1998. Frontoparietal cortical networks for directing attention and the eye to visual locations: Identical, independent, or overlapping neural systems? *Proceedings of the National Academy of Sciences USA*, **95**, 831–838.

Cozman, F., & Krotkov, E. 1997 (June). Depth from scattering. *Pages 801–806 of: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997.*

Dijk, A. M. Van, Martens, J.-B., & Watson, A. B. 1995. Quality Asessment of Coded Images Using Numerical Category Scaling. *Proc. SPIE – Advanced Image and Video Communications and Storage Technologies.*

Estrada, Cassius Rodrigo Duque. 2009. *Avaliação Automática de Qualidade de Videoconferências de Alta Definição.* Dissertação de Mestrado, Universidade Federal do Rio de Janeiro, Rio de Janeiro, Brasil.

Farias, M.C.Q., & Akamine, W.Y.L. 2012. On Performance of Image Quality Metrics Enhanced with Visual Attention Computational Models. *Electronics Letters*, **48**(11), 631–633.

Fonseca, Roberto Nery. 2008. *Algoritmos para Avaliação da Qualidade de Vídeo em Sistemas de Televisão Digital*. Dissertação de Mestrado, Escola Politécnica da Universidade de São Paulo, São Paulo, Brasil.

Fragoso, Miguel, Cruz, Pedro, & Marcelino, Vasco. 2012. *Televisão 3D*. MEEC de Comunicação de Áudio e Vídeo (CAV) do Instituto Superior Técnico, Lisboa, Portugal.

Gonzalez, Rafael C., & Woods, Richard E. 2006. *Digital Image Processing*. 3rd edition edn. Upper Saddle River, NJ, USA: Prentice-Hall, Inc.

Graziosi, D., Pagliari, C. L., Rodrigues, N. M. M., Silva, E. A. B., de Faria, S. M. M., & Carvalho, M. B. 2012. Codificação de mapas de profundidade usando casamento de padrões multiescalas. *In: XXX Simpósio Brasileiro de Telecomunicações(SBrT'12)*.

Guyton, Arthur C., & Hall, John E. 2006. *Tratado de Fisiologia Médica*. 11 edn. Elsevier.

han Lu, Xiao, Wei, Fang, & min Chen, Fang. 2012 (july). Foreground-Object-Protected Depth Map Smoothing for DIBR. *Pages 339 –343 of: IEEE International Conference on Multimedia and Expo (ICME)*.

Hewage, C.T.E.R., & Martini, M.G. 2011. Reduced-reference quality assessment for 3D video compression and transmission. *IEEE Transactions on Consumer Electronics*, **57**(3), 1185–1193.

Hewage, C.T.E.R., Worrall, S.T., Dogan, S., Villette, S., & Kondoz, A.M. 2009. Quality Evaluation of Color Plus Depth Map-Based Stereoscopic Video. *IEEE Journal of Selected Topics in Signal Processing*, **3**(2), 304 –318.

Hur, Namho, Tam, W.J., Speranza, F., Ahn, Chunghyun, & Lee, Soo In. 2005 (jan.). Depth-image-based stereoscopic image rendering considering IDCT and anisotropic diffusion. *Pages 381–382 of: International Conference on Consumer Electronics, ICCE.*

Huynh-Thu, Quan, Callet, P. Le, & Barkowsky, M. 2010 (sept.). Video quality assessment: From 2D to 3D – Challenges and future trends. *Pages 4025–4028 of: 17th IEEE International Conference on Image Processing (ICIP), 2010.*

IRCCyN-IVC. 2012 (Julho). *Nantes-Madrid 3D Stereoscopic Database.* `http://www.irccyn.ec-nantes.fr/spip.php?article1052`.

ISO/IEC. 2005. *ISO/IEC TR 13818-5:2005 - Information technology – Generic coding of moving pictures and associated audio information – Part 5: Software simulation.* http://www.nhzjj.com/asp/admin/editor/newsfile/201031910151236.pdf.

Itti, L, & Koch, C. 2001. Computational modeling of visual attention. *Nature Reviews Neuroscience*, **2**(3), 194–203.

ITU-R. 2000. *ITU-R BT.1438 – Subjective Assessment of Stereoscopic Television Pictures.*

ITU-R. 2002. *ITU-R BT.500-11 – Methodology for the subjective assessment of the quality of television pictures.*

ITU-R. 2010. *ITU-R BT.500-10 – Methodology for the subjective assessment of the quality of television pictures.*

ITU-T. 1999. *ITU-T P.910 – Subjective video quality assessment methods for multimedia applications.*

Joveluro, P., Malekmohamadi, H., Fernando, W.A.C., & Kondoz, A.M. 2010 (june). Perceptual Video Quality Metric for 3D video quality assessment. *Pages 1–4 of: 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video (3DTV-CON).*

Keval, Hina Uttam. 2009. *Effective, Design, Configuration, and Use of Digital CCTV.* Ph.D. thesis, Department of Computer Science, University College London.

Kim, C. S., Jin, S. H., Seo, D. J., & Ro, Y. M. 2008. Measuring Video Quality on Full Scalability of H.264/AVC Scalable Video Coding. *IEICE Transactions on Communications*, **E91-B**(5), 1269–1278.

Kuze, J., & Ukai, K. 2008. Subjective evaluation of visual fatigue caused by motion images. *Displays*, **29**(2), 159–166. Health and Safety Aspects of Visual Displays.

Leon, G., Kalva, H., & Furht, B. 2008 (May). 3D Video Quality Evaluation with Depth Quality Variations. *Pages 301–304 of: 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video.*

Li, Chaofeng, & Bovik, Alan Conrand. 2010. Content-weighted video quality assessment using a three-component image model. *Journal of Electronic Imaging, Special Section on Image Quality, Vol: 19 No: 1.*

Li, Junli, Chen, Gang, Chi, Zheru, & Lu, Chenggang. 2004. Image coding quality assessment using fuzzy integrals with a three-component image model. *IEEE Transactions on Fuzzy Systems*, **12**(1), 99 – 106.

Lucas, L. F. R., Rodrigues, N. M. M., Pagliari, C. L., Silva, E. A. B., & Faria, S. M. M. 2012. Codificação Eficiente de Mapas de Profundidade com Base em Predição e Aproximação Linear. *In: XXX Simpósio Brasileiro de Telecomunicações (SBrT'12).*

Maiti, S., Desai, P., Patel, B., Goel, Y., Piccinelli, E.M., Aliprandi, D., Fragneto, P., & Rossi, B. 2012 (oct.). Smart 3D video coding. *Pages 1–4 of: 3DTV-Conference: The True Vision – Capture, Transmission and Display of 3D Video (3DTV-CON), 2012.*

Malard, M. L., Costa, B. M., & Cosme, V. 2008 (August). *Development of MPEG Standards for 3D and Free Viewpoint Video.*

Mancini, A. 1994. *Disparity Estimation and Intermediate View Reconstruction for Novel Applications Stereoscopic Video.* M.Phil. thesis, McHill University.

Meesters, L. M. J., IJsselsteijn, W. A., & Seuntiens, P. J. H. 2004. A survey of Perceptual Evaluations and Requirements of Three-Dimensional TV. *IEEE Transactions on Circuits and Systems for Video Technology*, **14**(3), 381–391.

Mehlitz, Peter C. 2008. *The XGL Toolbox.*

Miksicek, Frantisek. 2006. *Causes of Visual Fatigue and Its Improvements in Stereoscopy.* Tech. rept. No. DCSE/TR-2006-0, University of West Bohemia in Pilsen, Czech Republic.

Oprea, C., Pirnog, I., Paleologu, C., & Udrea, M. 2009 (may). Perceptual Video Quality Assessment Based on Salient Region Detection. *Pages 232–236 of: Fifth Advanced International Conference on Telecommunications (AICT '09).*

Ozbek, N., & Tekalp, A. M. 2008 (april). Unequal Inter-view Rate Allocation Using Scalable Stereo Video Coding and an Objective Stereo Video Quality Measure. *Pages 1113–1116 of: IEEE International Conference on Multimedia and Expo, 2008.*

Patterson, R. 2007. Human factors of 3-D displays.

Pedrini, Hélio, & Schwartz, Willian Robson. 2007. *Análise de Imagens Digitais: Princípios, Algoritmos e Aplicações.* Thomson Learning.

Pereira, Eanes Torres. 2007 (Março). *Atenção Visual Bottom-up Guiada por Otimização via Algoritmos Genéticos.* Dissertação de Mestrado, Universidade Federal de Campina Grande, Campina Grande, Brasil.

Pinson, M.H., & Wolf, S. 2004. A new standardized method for objectively measuring video quality. *IEEE Transactions on Broadcasting*, **50**(3), 312–322.

Rajashekar, U., van der Linde, I., Bovik, A.C., & Cormack, L.K. 2008. GAFFE: A Gaze-Attentive Fixation Finding Engine. *IEEE Transactions on Image Processing*, **17**(4), 564 –573.

Ramos, André. 2006. *Fisiologia da Visão: Um estudo sobre o ver e o enxergar.* Apostila PUC-RIO.

Reckwerdt, Bill. 2012. *Quantitative Picture Quality Assessment Tools.* http://www.videoclarity.com/WPUnderstandingJNDDMOSPSNR.html.

Regis, C. D. M., Cardoso, J. V. M., Oliveira, I. P., & Alencar, M. S. 2012a. Performance of the objective video quality metrics with perceptual weighting considering first and second order differential operators. *Pages 71–74 of: Proceedings of the 18th Brazilian symposium on Multimedia and the web.* WebMedia'12. New York, NY, USA: ACM.

Regis, Carlos D. M., Cardoso, José V. M., & Alencar, Marcelo S. 2012b. Video Quality Assessment Based on the Effect of the Estimation of the Spatial Perceptual Information. *In: XXX Simpósio Brasileiro de Telecomunicações (SBrT'12).*

Richardson, Ian E. 2010. *The H.264 advanced video compression standard.* John Wiley & Sons, Ltd.

Roger, Leonardo Lorenzo Bravo. 2007. *Apostila de Estatística Geral e Aplicada*. http://www.ceset. unicamp.br/webdidat/matdidat.php?cod=TT%20203&nome =Leonardo+Lorenzo+Bravo+Roger: UNICAMP.

Seshadrinathan, K., & Bovik, A.C. 2010. Motion Tuned Spatio-Temporal Quality Assessment of Natural Videos. *IEEE Transactions on Image Processing*, **19**(2), 335–350.

Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. 2010a. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 1427–1441.

Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. 2010b. A subjective study to evaluate video quality assessment algorithms. *SPIE Proceedings Human Vision and Electronic Imaging*.

Shao, Hang, Cao, Xun, & Er, Guihua. 2009 (May). Objective quality assessment of depth image based rendering in 3DTV system. *Pages 1–4 of: 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*.

Sheikh, Hamid R., & Bovik, Alan C. 2004. Image information and visual quality. *Pages 430–444 of: IEEE Transactions on Image Processing*.

Shiguti, Wanderley Akira, & da S. C. Shiguti, Valéria. 2006. *Apostila de Estatística*.

Shmildt, Edilson Romais. 2007. *Apostila de Bioestatística: Estatística Descritiva – Unidade II*.

Silva, D.V.S.X., Fernando, W.A.C., Nur, G., Ekmekcioglu, E., & Worrall, S.T. 2010 (sept.). 3D video assessment with Just Noticeable Difference in Depth evaluation. *Pages 4013–4016 of: 17th IEEE International Conference on Image Processing (ICIP)*.

Silva, V., Fernando, A., Worrall, S., Arachchi, H.K., & Kondoz, A. 2011. Sensitivity Analysis of the Human Visual System for Depth Cues in Stereoscopic 3-D Displays. *IEEE Transactions on Multimedia*, **13**(3), 498–506.

Smolic, A., Kimata, H., & Vetro, A. 2005 (October). *Development of MPEG Standards for 3D and Free Viewpoint Video*. Tech. rept. Mitsibishi Electronic Research Laboratories.

Stockhammer, T., Hannuksela, M.M., & Wiegand, T. 2003. H.264/AVC in wireless environments. *IEEE Transactions on Circuits and Systems for Video Technology*, **13**(7), 657–673.

Tam, Wa James, Vázquez, Carlos, & Speranza, Filippo. 2007. Surrogate depth maps for stereoscopic imaging:different edge types. *Proc. SPIE 6490*.

Tam, W.J., & Zhang, Liang. 2006 (July). 3D-TV Content Generation: 2D-to-3D Conversion. *Pages 1869–1872 of: IEEE International Conference on Multimedia and Expo, 2006*.

Tam, W.J., Speranza, Yano, F., S., Shimono, K., & Ono, H. 2011. Stereoscopic 3D-TV: Visual Comfort. *IEEE Transactions on Broadcasting*, **57**(2), 335–346.

Technical University of Munich. 2012. *LIVE Video Quality Database*. ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test sequences/.

Tikanmaki, A., Gotchev, A., Smolic, A., & Miller, K. 2008 (april). Quality assessment of 3D video in rate allocation experiments. *Pages 1–4 of: IEEE International Symposium on Consumer Electronics, (ISCE)*.

Torralba, A., & Oliva, A. 2002. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **24**(9), 1226 – 1238.

Tourapis, Alexis Michael. 2012 (March). *H.264/AVC Software Coordination, version 18.3*. http://iphome.hhi.de/suehring/tml/.

Trace, Website. 2008 (Abril). *YUV Video Sequences*.

Tsotsos, John K. 1990. *Analyzing vision at the complexity level*. The Behavioraland Brain Sciences.

Urvoy, M., Barkowsky, M., Cousseau, R., Koudota, Y., Ricorde, V., Callet, P. Le, Gutierrez, J., & Garcia, N. 2012 (july). NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences. *Pages 109–114 of: 2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX)*.

VCEG. 1999. *Common test conditions for RTP/IP over 3GPP/3GPP2*. http://ftp3.itu.ch/av-arch/video-site/0109 San/VCEG-N80software.zip.

VQEG. 2000 (April). *Final report from the video quality experts group on the validation of objective models of video quality assessment, phase II (FR-TV2.* available at http://www.vqeg.org/.

Vranjes, M., Rimac-Drlje, S., & Zagar, D. 2007. Objective video quality metrics. *49th International Symposium ELMAR, 2007*, Sept., 45–49.

Vu, Cuong, & Deshpande, Sachin. 2012. ViMSSIM: from image to video quality assessment. *Pages 1–6 of: Proceedings of the 4th Workshop on Mobile Video.* MoVid '12. New York, NY, USA: ACM.

Wang, Z., Simoncelli, E.P., & Bovik, A.C. 2003 (Nov.). Multiscale structural similarity for image quality assessment. *Pages 1398–1402 of: Conference Record of the Thirty-Seventh Asilomar Conference on Signals, Systems and Computers*, vol. 2.

Wang, Zhou, & Li, Qiang. 2011. Information Content Weighting for Perceptual Image Quality Assessment. *IEEE Transactions on Image Processing*, **20**(5), 1185–1198.

Wang, Zhou, & Simoncelli, E.P. 2005 (18 - 23). Translation Insensitive Image Similarity in Complex Wavelet Domain. *Pages 573–576 of: IEEE International Conference on Acoustics, Speech, and Signal Processing, (ICASSP '05)*, vol. 2.

Wang, Zhou, Bovik, Alan C., Sheikh, Hamid R., & Simoncelli, Eero P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, **13**(4), 600–612.

Webster, Arthur A., Jones, Coleen T., Pinson, Margaret H., Voran, Stephen D., & Wolf, Stephen. 1993. An Objective Video Quality Assessment System Based on Human Perception. *Pages 15–26 of: in SPIE Human Vision, Visual Processing, and Digital Display IV.*

Wenger, S. 2003. H.264/AVC over IP. *IEEE Transactions on Circuits and Systems for Video Technology*, **13**(7), 645 – 656.

Winkler, Stefan. 2005. *Digital Video Quality: Vision Models and Metrics.* Wiley.

Winkler, Stefan, & Mohandas, Praveen. 2008. The Evolution of Video Quality Measurement: From PSNR to Hybrid Metrics. *IEEE Transactions on Broadcasting*, **54**(3), 660–668.

Wolfe, Jeremy M., & Horowitz, Todd S. 2004. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, **5**(6), 495–501.

Woods, Andrew, Docherty, Tom, & Koch, Rolf. 1993. Image Distortions in Stereoscopic Video Systems. *In: Stereoscopic Displays and Applications*.

Wu, H.R., & Rao, K.R. 2006. *Digital Video Image Quality and Perceptual Coding*. Taylor & Francis.

Yang, Jiachen, Hou, Chunping, Zhou, Yuan, Zhang, Zhuoyun, & Guo, Jichang. 2009 (May). Objective quality assessment method of stereo images. *Pages 1 –4 of: 3DTV Conference: The True Vision - Capture, Transmission and Display of 3D Video, 2009*.

Yasakethu, S.L.P., Hewage, C., Fernando, W., & Kondoz, A. 2008. Quality analysis for 3D video using 2D video quality models. *IEEE Transactions on Consumer Electronics*, **54**(4), 1969–1976.

You, Junyong, Perkis, A., & Gabbouj, M. 2010 (July). Improving image quality assessment with modeling visual attention. *Pages 177–182 of: 2nd European Workshop on Visual Information Processing (EUVIP)*.

Young, Richard A. 1991. Oh say, can you see? The physiology of vision. 92–123.

Zhang, Liang, & Tam, W.J. 2005. Stereoscopic image generation based on depth images for 3D TV. *IEEE Transactions on Broadcasting*, **51**(2), 191–199.

Zhang, Liang, Vazquez, C., & Knorr, S. 2011. 3D-TV Content Creation: Automatic 2D-to-3D Video Conversion. *IEEE Transactions on Broadcasting*, **57**(2), 372–383.

Zhao, Xiaoqun, & Zhang, Qianying. 2012 (sept.). Research of unequal error protected Turbo code based on H.264 stereoscopic video transmission. *Pages 449 –452 of: 2nd International Conference on Applied Robotics for the Power Industry (CARPI)*.

Zhou Wang, Ligang Lu, & Bovik, Alan C. 2004. Video Quality Assessment Based on Structural Distortion Measurement. *Signal Processing: Image Communication*, **19**(2), 121–132.