

AVALIAÇÃO DE SISTEMAS DE REDUÇÃO DE TAXA DE BITS DE VÍDEO COM ENFOQUE NA PERCEPÇÃO DO USUÁRIO

Autora: Ana Eliza Faria e Silva

Orientador: Max Henrique Machado Costa

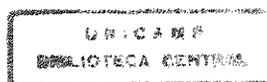
Data: fevereiro 1998

Este exemplar corresponde a redação final da tese defendida por <u>Ana Eliza Faria e Silva</u> e aprovada pela Comissão Julgada em <u>06 / 02 / 98</u> <u>Max C. M.</u> Orientador
--

Dissertação submetida à Faculdade de Engenharia Elétrica e de Computação como parte dos requisitos exigidos para a obtenção do Título de Mestre em Engenharia Elétrica



Departamento de Comunicações
Faculdade de Engenharia Elétrica e de Computação
Universidade Estadual de Campinas



9812639

UNIDADE	BC
N.º CHAMADA:	TUNICAMP
	F225a
V.	Ex.
T.º MENS	80/33998
PROJ.	395/98
C	<input type="checkbox"/>
D	<input checked="" type="checkbox"/>
PREÇO	R\$ 11,00
DATA	28/05/98
N.º CPD	

CM-00112905-6

FICHA CATALOGRÁFICA ELABORADA PELA
BIBLIOTECA DA ÁREA DE ENGENHARIA - BAE - UNICAMP

F225a Faria e Silva, Ana Eliza
Avaliação de sistemas de redução de taxa de bits de
video com enfoque na percepção do usuário. / Ana Eliza
Faria e Silva.--Campinas, SP: [s.n.], 1998.

Orientador: Max Henrique Machado Costa

Dissertação (mestrado) - Universidade Estadual de
Campinas, Faculdade de Engenharia Elétrica e de
Computação.

1. Televisão digital. 2. Comunicações digitais.

I. Costa, Max Henrique Machado II. Universidade
Estadual de Campinas. Faculdade de Engenharia Elétrica
e de Computação. III. Título.

À minha família

Agradecimentos

Agradeço ao meu orientador, professor Max Costa, o auxílio na condução desta tese ao longo destes dois anos de pesquisa.

Agradeço à equipe de processamento digital de imagens do Projeto de Pesquisa Aplicada em Processamento Digital de Sinais do CPqD/TELEBRÁS, especialmente a Antonio Cláudio França Pessoa, as orientações e sugestões na elaboração deste trabalho.

Agradeço à professora Matilde Scaramucci, do Instituto de Estudos Linguísticos da UNICAMP, o auxílio na elaboração do experimento abordando as diferenças semânticas de adjetivos para o português falado no Brasil.

Agradeço também à Central Globo de Engenharia da TV Globo o apoio recebido e a todos que participaram como avaliadores nos experimentos realizados, a inestimável contribuição.

A todos que me apoiaram nesta jornada, meu muito obrigada!

Resumo

Este trabalho trata de uma questão fundamental em compressão de vídeo, a avaliação da eficiência dos algoritmos empregados em *'esconder'* as distorções de imagens, explorando características do sistema visual humano. As degradações no vídeo digital são dificilmente mensuráveis objetivamente, embora sejam quase sempre óbvias para observadores humanos. Esta constatação aliada a alta relevância da opinião do usuário final sobre a qualidade da imagem, norteiam, nesta tese, o desenvolvimento de novas ferramentas de avaliação de qualidade de vídeo baseadas na percepção humana.

Inicialmente, as ferramentas tradicionais de avaliação subjetiva de qualidade de vídeo, e em particular, os diversos aspectos em que o projeto de um experimento pode ser otimizado são descritos. Um procedimento típico de avaliação consiste em apresentar, a um conjunto de avaliadores, geralmente não especialistas e com acuidade visual normal, uma série de sequências de vídeo em ambiente controlado e por cerca de 20 minutos. Com base nestas cenas de teste a qualidade da imagem é questionada de uma série de formas. Para a obtenção de resultados repetíveis e confiáveis faz-se uma análise sistemática de vários aspectos relacionados ao teste como seleção criteriosa de cenas, avaliadores, método de avaliação e o intervalo de confiança adequado. Consideram-se ainda as influência dos fatores psicológicos e aspectos linguísticos dos termos utilizados nos testes. Apresenta-se o resultado de um experimento real de avaliação subjetiva visando a caracterização do desempenho de um conjunto de sistemas de compressão de imagem diversos para aplicações em televisão. Este experimento permite a obtenção de um banco de dados com resultados subjetivos representativos das degradações introduzidas em cenas naturais por codecs de vídeo, bem como sua comparação à qualidade de estúdio alcançada por sistemas NTSC e PAL-M. Em seguida, estes resultados são aplicados no desenvolvimento de um método de estimação de qualidade subjetiva por meio de medidas objetivas. O modelo utiliza a diferença entre as imagens de entrada e de saída, segmentação das imagens e predição linear. Por fim, apresentam-se resultados da aplicação deste modelo.

Abstract

This thesis addresses a fundamental question in video compression, namely the evaluation of efficiency of the compression algorithms in hiding distortions by exploiting characteristics of the human visual system. Digital video degradations are usually difficult to measure objectively, even though they are often obvious for human observers. This realization, added to the relevance of final user opinion about the image quality guided, in this work, the development of new evaluation tools based on human perception.

Initially, traditional evaluation tools and various aspects in which experiment design can be improved are described. A typical evaluation procedure consists of presenting a group of observers of normal visual acuity, usually non-specialists, a selected series of test scenes under a controlled environment, for about 20 minutes. Based on these sequences, the quality of the images is questioned in different ways. Reliable and repeatable results can only be obtained after a systematic analysis is made of various aspects related to the experiment, such as careful selection of the test scenes, observers, evaluation method and adequate confidence intervals. Also considered are the influence of psychological factors and linguistic aspects. The result of a real-case experiment is presented, where a number of codecs for television applications are evaluated. Also presented is a comparison of these results with the quality obtained in NTSC and PAL-M systems. These results led to the construction of a data base of subjective perceptions which are representative of various degradations introduced in natural test scenes by video codecs, and to the development of a model for estimation of subjective image quality by means of objective measurements. The model uses the difference between input and output scenes, image segmentation, and linear prediction. Finally, application of this model is illustrated.

Índice

1. Introdução	12
2. Ferramentas de codificação digital de imagens	15
2.1 Introdução: digitalização e compressão	15
2.2 Aplicação dos sistemas de compressão em televisão: da geração ao usuário	16
2.3 Fundamentos da compressão de imagens	19
2.4 Padrão MPEG-2: vídeo	21
2.5 Caracterização de artefatos associados à compressão	24
2.5.1 <i>Blocagem</i>	25
2.5.2 <i>“Blurring”</i>	25
2.5.3 <i>“Edge Business”</i>	26
2.5.4 <i>Ruído</i>	26
2.5.5 <i>“Jerkiness”</i>	26
2.5.6 <i>Blocos Errados</i>	27
3. Avaliação de sistemas de compressão	28
3.1 Insuficiência de testes analógicos	28
3.2 Escopo e objetividade dos métodos de avaliação	29
4. Avaliação subjetiva: Projeto de experimento	31
4.1 Fundamentos da metodologia de avaliação	32
4.2 Impacto das condições de observação	33
4.2.1 <i>Condições de visualização padronizadas</i>	33
4.2.2 <i>Influência da distância de observação e do tamanho da imagem</i>	34
4.3 Análise de variabilidade entre observadores	35
4.4 Seleção do material de teste	36
4.4.1 <i>Critério de ‘criticalidade’ de cenas</i>	38
4.4.2 <i>Resultados de simulação de criticalidade</i>	39
4.5 Forma de apresentação das seqüências	43
4.5.1 <i>Utilização de cenas de referência</i>	43
4.5.1.1 <i>Descrição e quantificação do efeito de adaptação</i>	44
4.5.2 <i>Técnicas de comparação</i>	45
4.5.2.1 <i>Ferramenta de análise de transitividade: método Zeta</i>	45
4.5.3 <i>Influência do número de apresentações</i>	47
4.5.4 <i>Duração dos trechos de avaliação</i>	47
4.5.4.1 <i>Capacidade de avaliação contínua</i>	48
4.5.4.2 <i>Caracterização do efeito de esquecimento</i>	49
4.6 Escalas	49

4.6.1	<i>Caracterização das escalas</i>	50
4.6.1.1	Escala de categorias.....	50
4.6.1.2	Escala numérica linear.....	50
4.6.1.3	Escala racional.....	51
4.6.1.4	Escala gráfica.....	52
4.6.1.5	Escala multidimensional.....	52
4.6.2	<i>Comparações e análise de desempenho</i>	53
4.6.2.1	Escala de categorias: qualidade vs degradação.....	53
4.6.2.2	Comparação entre escalas racionais e ordinais com relação ao efeito de adaptação.....	54
4.6.2.3	Análise de intervalos nas escalas de categorias por meio de escalas gráficas.....	55
4.6.3	<i>Análise de intervalos da escala de categorias de qualidade padronizada para o português falado no Brasil</i>	56
4.6.3.1	Descrição do teste.....	56
4.6.3.2	Apresentação dos resultados.....	56
4.6.3.3	Comparação dos intervalos de qualidade em diversas línguas.....	58
4.7	Ferramentas de análise de resultados.....	59
4.7.1	<i>Procedimentos estatísticos de descarte de notas e avaliadores</i>	59
4.7.2	<i>Estimativa do intervalo de confiança associado ao teste</i>	60
4.8	Novos problemas: novas metodologias.....	62
4.8.1	<i>Escalonamento multidimensional de imagens</i>	62
4.8.1.1	Princípios da análise em múltiplas dimensões.....	63
4.8.1.2	Consideração de projeto em Múltiplas Dimensões.....	64
4.8.1.3	Dimensões de qualidade: Resultados de estudos preliminares.....	65
4.8.2	<i>Refinamento da metodologia de “escala de degradação a estímulo duplo”</i>	66
5.	Estimação de qualidade subjetiva com base em parâmetros objetivos	68
5.1	Considerações gerais: O que medir?.....	69
5.2	Norma ANSI de parâmetros objetivos.....	70
5.3	Linhas de pesquisa.....	71
5.3.1	<i>Modelo do sistema visual humano</i>	71
5.3.2	<i>Medidas no “bitstream”</i>	73
5.3.3	<i>Medidas em cenas naturais</i>	74
5.3.3.1	Método de estimação empregado: Casamento do banco de dados subjetivo.....	75
5.4	Segmentação como ferramenta de avaliação.....	75
5.5	Seleção de estimadores objetivos.....	76
5.5.1	<i>Seleção de medidas objetivas</i>	77
5.5.1.1	Erro quadrático médio.....	78
5.5.1.2	Filtragem Sobel.....	78
5.5.2	<i>Correlação entre estimadores e degradações</i>	79
5.6	Introdução de funções logísticas como modelo para descrição de degradações.....	80
5.6.1	<i>Efeito de saturação e assimetria da percepção</i>	80
5.6.2	<i>Modelo para descrição subjetiva de uma única degradação</i>	81
5.6.3	<i>Efeito subjetivo de múltiplas fontes de degradação</i>	82
6.	Estudo de caso	83

6.1	Material de teste.....	83
6.1.1	<i>Cenas escolhidas</i>	83
6.1.2	<i>Processamento</i>	84
6.1.2.1	Cenas originais.....	84
6.1.2.2	Subamostragem temporal e espacial.....	84
6.1.2.3	Codificação MPEG-2.....	85
6.1.2.4	Codificação MPEG-2 e degradação adicional.....	85
6.1.2.5	Sistema de codificação com qualidade constante e taxa variável.....	86
6.1.2.6	Codecs para sinal composto.....	87
6.1.3	<i>Preparação do material</i>	87
6.1.3.1	Arranjo das Sequências.....	87
6.1.3.2	Ordenação.....	87
6.1.4	<i>Forma de apresentação</i>	88
6.2	Detalhes do teste.....	88
6.2.1	<i>Avaliadores</i>	88
6.2.2	<i>Votação</i>	89
6.3	Resultados e conclusões.....	90
6.3.1	<i>Médias gerais</i>	90
6.3.1.1	Descarte de observadores e de notas.....	90
6.3.1.2	Ajuste de coerência por tipo de processamento.....	90
6.3.2	<i>Diferenças perceptuais entre cenas e entre grupos</i>	92
6.3.3	<i>Análise de resultados dos subsistemas em teste</i>	93
6.3.3.1	Codec C4 (YCbCr422 18 Mbits/s).....	93
6.3.3.2	Subamostragem espacial e temporal.....	94
6.3.3.3	Sistema de codificação com qualidade constante.....	94
6.3.3.4	Codecs MPEG-2.....	96
6.3.3.5	Codecs para sinal composto.....	98
6.3.3.6	Comparação entre NTSC e codecs de vídeo.....	99
6.4	Aplicação dos Resultados na Estimação de Qualidade Subjetiva através de Parâmetros Objetivos.....	100
6.4.1	<i>Cálculo dos Parâmetros Objetivos</i>	100
6.4.2	<i>Caracterização da Base de Dados e da Base de Treinamento</i>	102
6.4.3	<i>Apresentação dos Resultados</i>	103
7.	Conclusões	106
8.	Referências Bibliográficas	109
9.	Anexos	114
9.1	Tabela de Resultados: Análise Gráfica dos Descritores de Qualidade.....	114
9.2	Base de Dados Subjetivos.....	115
9.3	Base de Dados Objetivos.....	120
9.4	Resultados de Estimação de Qualidade Subjetiva com base em Medidas Subjetivas.....	126

Índice de Figuras

Figura 1: Diagrama de blocos de um sistema de comunicações genérico.....	16
Figura 2: Cadeia de Broadcast baseada em aplicações integradas de compressão.....	17
Figura 3: Estrutura do codificador de vídeo.....	19
Figura 4: Estrutura do decodificador de vídeo.....	19
Figura 5: Estrutura de um codec MPEG-2: codificador.....	22
Figura 6: Estrutura de um codec MPEG-2: decodificador.....	22
Figura 7: Distribuição de criticalidade em programas de TV (Q=6).....	39
Figura 8: Medida de criticalidade - Variação do parâmetro de quantização.....	42
Figura 9: Medidas de criticalidade - Variância da Programação.....	42
Figura 10: Resultados de Análise Gráfica para Descritores em Português.....	57
Figura 11: Análise Gráfica dos Termos Padronizados.....	57
Figura 12: Comparação dos intervalos adotados para os descritores de Qualidade.....	58
Figura 13: Curvas com o erro máximo esperado em testes subjetivos.....	61
Figura 14: Fator de correção do erro máximo com base na média geral do teste.....	62
Figura 15: Hierarquia de desenvolvimento de parâmetros objetivos.....	69
Figura 16: Algoritmo de estimação de qualidade subjetiva com base em cenas naturais.....	75
Figura 17: Ilustração do resultado de segmentação.....	76
Figura 18: Caracterização do medidor “erro quadrático médio” (MSE).....	77
Figura 19: Descrição do filtro Sobel.....	78
Figura 20: Determinação da Diferença Sobel Positiva - PSD.....	78
Figura 21: Determinação da Diferença Sobel Negativa - NSD.....	79
Figura 22: Caracterização da filtragem Sobel.....	79
Figura 23: Sala de avaliação subjetiva.....	83
Figura 24: Forma de apresentação das seqüências.....	88
Figura 25: Relação entre a escala de degradação e o valor numérico das notas.....	90
Figura 26: Notas médias obtidas por cada seqüência.....	92
Figura 27: Resultado subjetivo do codec C4 a 18 Mbits/s - Média geral.....	93
Figura 28: Resultado subjetivo da subamostragem espacial e temporal - Médias por cena.....	94
Figura 29: Resultado subjetivo da subamostragem espacial e temporal - Médias por grupo.....	94
Figura 30: Desempenho subjetivo do sistema de qualidade constante - Média por cena.....	95
Figura 31: Desempenho subjetivo do sistema de qualidade constante - Média por grupo.....	95

Figura 32: Desempenho subjetivo de codecs MPEG-2 - Média por cena.	96
Figura 33: Desempenho subjetivo de codecs MPEG-2 - Média por grupo.	96
Figura 34: Comparação entre o desempenho com relação ao parâmetro M.	97
Figura 35: Comparação de desempenho subjetivo entre codecs MPEG-2, com M=2.	98
Figura 36: Desempenho Subjetivo dos Codecs - Média por Cena.	98
Figura 37: Desempenho Subjetivo dos Codecs - Média por Grupo.	99
Figura 38: Comparação entre o desempenho subjetivo de codecs MPEG-2 e NTSC/PAL-M.	99
Figura 39: Comparação entre o desempenho subjetivo do sistema de qualidade constante e NTSC/PAL-M.	100
Figura 40: Ilustração de Confiabilidade dos parâmetros para a cena <i>Flower Garden</i>	101
Figura 41: Algoritmo de combinação dos preditores.	102
Figura 42: Distribuição dos pesos dos parâmetros objetivos - Resultado por componente.	103
Figura 43: Distribuição dos pesos dos parâmetros objetivos - Resultado por região.	104
Figura 44: Erros de estimação de qualidade subjetiva.	104
Figura 45: Comparação de desempenho - Ganho com Segmentação.	105
Figura 46: Gráfico de dispersão - Resultados objetivos x subjetivos.	105

Índice de Tabelas

Tabela 1: Cenas utilizadas na simulação de criticalidade - CLASSE 1	39
Tabela 2: Cenas utilizadas na simulação de criticalidade - CLASSE 2	40
Tabela 3: Especificação dos parâmetros de codificação para medidas de criticalidade	41
Tabela 4: Categorias para descrição da imagem para o português falado no Brasil	50
Tabela 5: Definição do critério de descarte de notas	59
Tabela 6: Associação entre medidor e degradação	80
Tabela 7: Descrição das condições de degradação avaliadas	84
Tabela 8: Especificação dos parâmetros de codificação em teste	85
Tabela 9: Descrição das Degradações adicionadas por simulação	86
Tabela 10: Configuração do sistema de codificação com qualidade constante	86
Tabela 11: Especificação do encoder e decoder	87
Tabela 12: Descrição dos grupos de proximidade no experimento	88
Tabela 13: Descrição dos avaliadores do teste subjetivo	89
Tabela 14: Discriminação do percentual de descarte de notas	90
Tabela 15: Erro absoluto médio resultante do ajuste de notas	91
Tabela 16: Diferenças perceptuais entre notas por grupos	93
Tabela 17: Definição das classes de sistemas	103

1. Introdução

Uma imagem pode ser descrita através de diferentes parâmetros físicos, como tamanho, brilho, resolução, conteúdo espectral, etc. A qualidade da imagem, entretanto, é uma noção subjetiva, pois é a sensação da pessoa que olha a imagem e deve, portanto, ser descrita em termos subjetivos. A caracterização da imagem em termos subjetivos é feita através de dimensões subjetivas de qualidade. A imagem é digitalizada em função das facilidades de manipulação e da qualidade passível de ser obtida, mas deve ser comprimida para que possa ser eficientemente transmitida e armazenada. A função da codificação digital é a redução da taxa necessária para representar uma seqüência de imagens garantindo mínima perda na qualidade da imagem.

Hoje é possível prover recepção doméstica de diversos canais de televisão digitais graças ao desenvolvimento de técnicas de transmissão bastante efetivas e a padronização das técnicas de compressão de áudio e vídeo. A fim de explorar esta nova tecnologia, diversos tipos de novos serviços foram planejados, como vídeo-sob-demanda, canais temáticos e “pay-per-view”. Embora a factibilidade destes novos serviços digitais já estejam indubitavelmente demonstradas, a maior questão neste cenário ainda permanece e consiste na aceitação do usuário deste novo tipo de serviço. Diversos fatores irão determinar seu sucesso, incluindo o preço inicial do serviço, interesse do usuário no conteúdo do canal e competitividade entre diferentes operadores. Costuma-se pensar que apenas o que se transmite e a que preço é que são fundamentais, e não os aspectos técnicos envolvendo a qualidade da imagem. Entretanto, embora o fator econômico tenha um peso considerável, a importância da qualidade da imagem não deve ser menosprezada, já que o usuário certamente irá levá-la em consideração ao fazer sua escolha pelos novos serviços disponíveis.

A economia em termos de meios de transmissão e capacidade de armazenamento, possibilitada pelo uso de compressão, depende da quantidade de informação presente no sinal original (não comprimido) e de quanto da qualidade original se está disposto a sacrificar. As degradações aparecem quando a quantidade de informação presente no vídeo é maior do que a capacidade de transmissão do canal. Os usuários de alguns tipos de serviço, entretanto, se dispõem a sacrificar parte da qualidade para economizar nos custos. É questão fundamental, portanto, a eficiência da compressão. Os equipamentos de codificação conseguem reduzir o número de bits para a representação da imagem, extraindo a redundância estatística da imagem, tanto quanto possível. Se uma redução adicional de bits é necessária, e isto é sempre verdade em aplicações para entretenimento, então alguma distorção deverá ser introduzida na imagem. Um dos objetivos deste processo de compressão é ‘esconder’ estas distorções explorando características do sistema visual humano. No processo, são geralmente usados métodos baseados na transformada discreta do cosseno e em codificação com palavras de tamanho variável. O aumento de eficiência obtido está relacionado à flexibilidade na alocação dinâmica de

bits entre as diversas regiões da imagem, de forma a preservar sua qualidade, e é justamente o que dificulta sua caracterização e avaliação. Este trabalho trata, no Capítulo 2, de algoritmos padrões para a redução da taxa de bits e faz considerações sobre seu desempenho em termos de qualidade da imagem, capacidade de processamento e robustez na presença de erros de transmissão. É dada atenção à caracterização dos artefatos tipicamente presentes em imagens digitais comprimidas.

A aplicação crescente de codificação digital e redução de taxa de bits para transmissão aumentou a relevância das medidas das degradações causadas pela codificação. No Capítulo 3 aprofunda-se a discussão sobre a importância e as dificuldades em se avaliar este tipo de sistema. A dificuldade de avaliação está fundamentalmente relacionada à adaptabilidade dos algoritmos utilizados. Em especial, analisa-se a inviabilidade de utilização de sinais de teste analógicos, já que as degradações introduzidas são próprias dos algoritmos digitais de compressão e diferentes das distorções características de sistemas analógicos.

No contexto de avaliação de imagem, os procedimentos de testes subjetivos são parte importante, pois o critério definitivo na avaliação de um sistema é a opinião do usuário final. As degradações não são, em geral, facilmente medidas objetivamente, mas são normalmente óbvias para observadores humanos. A avaliação subjetiva como definida na Recomendação ITU-R BT.500 [02], foi usada durante vários anos, com métodos de avaliação relativamente estáveis e satisfatórios até o advento da compressão digital. Resumidamente, para um conjunto de avaliadores, geralmente não especialistas com acuidade visual normal, é apresentada uma série de cenas de teste em ambiente controlado e por cerca de 20 minutos. Com base nestas cenas de teste a qualidade da imagem é questionada de uma série de formas. Em sistemas de televisão que incorporam compressão, a qualidade da cena não é constante no tempo, mas sim função da complexidade da programação. Considerando a natureza variante no tempo das degradações e o número de novas degradações, novos métodos foram propostos. Adicionalmente à seleção da metodologia de avaliação, há inúmeros outros elementos da avaliação para os quais abordagens alternativas são possíveis, tais como, condições de visualização, escolha dos avaliadores, condições de referência, sinal fonte do teste, duração dos trechos de vídeo em teste, seleção do "range" de degradação e maneiras de analisar os resultados obtidos. A seleção destes parâmetros está relacionada a aplicação e gera uma grande gama de possibilidades. É parte do Capítulo 4 desta tese discutir e analisar cada um destes aspectos.

A grande vantagem da avaliação subjetiva é a obtenção de resultados válidos para os sistemas convencionais e sistemas de compressão sobre larga gama de aplicações para cenas paradas e cenas em movimento. Em contrapartida, uma grande variedade de métodos e parâmetros de teste precisam ser considerados e diversos avaliadores devem ser envolvidos, elevando os custos de avaliação. O quarto Capítulo 4 enfatiza a metodologia de análise subjetiva desses sistemas de compressão nos diversos contextos de aplicação. Esse assunto não se esgota dentro dos limites demarcados pela recomendação do ITU. Ao contrário, são analisadas diversas outras técnicas, todas elas visando definir a qualidade da imagem da forma mais precisa possível.

Os testes subjetivos são fundamentais em aplicações de desenvolvimento, mas não se prestam diretamente para monitoração. É clara a necessidade de métodos objetivos de avaliação.

Os testes subjetivos são complexos e a variabilidade dos resultados alta. Entretanto, já que é a opinião do avaliador sobre a qualidade da imagem o que realmente conta, qualquer medida objetiva deve ter boa correlação com o resultado subjetivo do mesmo sistema para a mesma cena. Da mesma forma que as medidas subjetivas, muitas medidas objetivas não medem a qualidade da imagem diretamente, mas provêem uma indicação de quão degradada está uma imagem em relação a uma imagem de referência. É esta a abordagem escolhida no Capítulo 5 para a estimação de qualidade subjetiva com base em medidas objetivas.

A última parte do trabalho visa uma definição criteriosa de procedimentos para tratar a avaliação de codecs de vídeo. O Capítulo 6 traz uma análise detalhada dos aspectos relevantes na elaboração de testes subjetivos, análise de resultados, preparação do material e do ambiente de teste e descrição da infra-estrutura necessária para a condução do experimento. A análise está baseada em um experimento realizado no CPqD TELEBRÁS para a classificação subjetiva de um conjunto de sistemas e parâmetros usuais no contexto de televisão digital. A base de dados subjetivos obtida é usada para o desenvolvimento empírico de um sistema de estimação de qualidade subjetiva.

A atividade de avaliação de sistemas digitais de transmissão de imagens e vídeo ainda não atingiu sua maturidade no sentido do estabelecimento de padrões abrangentes capazes de prever a qualidade subjetiva de sistemas digitais para a larga gama de aplicações a que se destinam. As conclusões obtidas por meio deste trabalho estão detalhadas no Capítulo 7. Este trabalho analisa e documenta diversos trabalhos relacionados a atividade de avaliação nos últimos anos e fornece importantes contribuições no tocante a avaliação de sistemas de redução de taxa de bits de vídeo com enfoque na percepção do usuário.

2. Ferramentas de codificação digital de imagens

Este capítulo procura caracterizar os principais elementos em um sistema de transmissão de vídeo digital, com enfoque nas ferramentas de codificação MPEG-2, “Motion Picture Experts Group” [07]. Esta análise é imprescindível para se verificar o potencial de aplicações e adquirir sensibilidade sobre os problemas e dificuldades na implementação desta nova tecnologia. O problema da compressão é abordado, primeiramente do ponto de vista da aplicação em ambiente de televisão e em seguida é feita uma descrição do codificador, do decodificador e dos artefatos presentes em imagens comprimidas.

2.1 Introdução: digitalização e compressão

São muitas as vantagens da digitalização e da compressão de imagens. A digitalização permite maior flexibilidade na manipulação dos dados e enriquecimento das possibilidades de tratamento do sinal. Os circuitos são mais estáveis e confiáveis, a imunidade ao ruído é maior e o transporte de sinais de áudio, vídeo e dados pode ser feito conjuntamente através de multiplexagem. Além disso, há possibilidade de detecção e correção de erros pela utilização de códigos e melhores alternativas de criptografia e implementação de acesso condicional. A TV digital adapta-se mais facilmente a automação pois pode ser controlada por meio de computadores e, uma vez que carrega as informações dos componentes de cor separadamente, facilita também a troca de programas entre países.

A digitalização de imagens permite um aumento da robustez do sistema frente às distorções introduzidas pelo canal de transmissão e a integração de mídias e serviços. Além disso, em sua forma digital as imagens têm melhor qualidade, a manipulação e edição são mais fáceis e, quando associadas a técnicas de compressão, pode-se obter um aumento da eficiência espectral.

Compressão de imagens é a redução do número de bits requeridos para se descrever uma imagem ou seqüência de imagens, garantindo um nível de fidelidade em relação às imagens originais que depende fortemente do serviço e da aplicação. Estas técnicas sofreram considerável avanço na última década, motivada pela redução de custos que se pode obter por economia dos meios de transmissão e diminuição dos requisitos de armazenamento. Existem duas categorias de compressão. A compressão sem perdas, ou compactação, na qual o sinal reconstruído é exatamente igual ao sinal original, e a compressão com perdas, na qual o sinal reconstruído pode ser diferente do sinal original, mantido porém certo nível de fidelidade objetiva ou subjetiva.

Alguns aspectos da tecnologia de compressão de vídeo que podem implicar em perda de qualidade devem ser levados em consideração. O vídeo comprimido é fortemente susceptível a erros de canal, aumentando os requisitos por códigos corretores de erros, algoritmos de

compressão mais robustos e técnicas de cancelamento eficientes. A compressão dificulta a concatenação de “streams” de vídeo de diversas fontes em tempo real, como por exemplo, a inserção de comerciais. A codificação e o empacotamento do vídeo para a transmissão implicam em um atraso que pode ser crítico dependendo da aplicação. Além disso, em aplicações em que há cascadeamento de diversas etapas de compressão é preciso cuidar para que os estágios iniciais sejam comprimidos com alta qualidade, como forma de garantir que não haja excessiva degradação de qualidade para o usuário.

Um sistema de comunicações digital pode ser representado genericamente pelo diagrama de blocos da Figura 1.

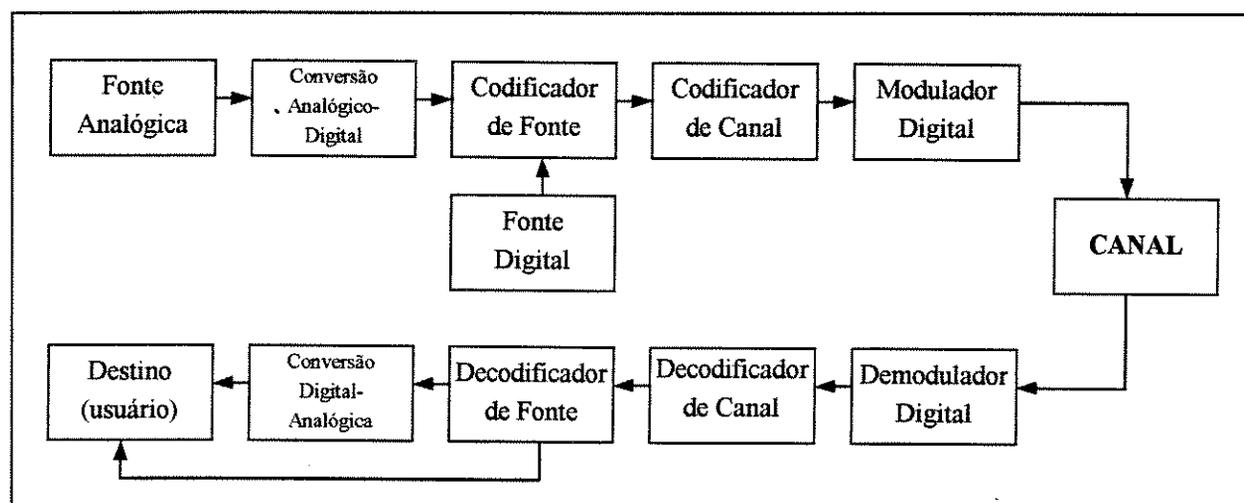


Figura 1: Diagrama de blocos de um sistema de comunicações genérico

2.2 Aplicação dos sistemas de compressão em televisão: da geração ao usuário

A compressão digital de vídeo incorpora diferentes algoritmos, técnicas e taxas de bits que permeiam todo o ambiente de televisão. Esta nova realidade traz novos problemas e preocupações entre o corpo de gerenciamento técnico e os profissionais de criação que servem a indústria da televisão. O fato é que a compressão de vídeo é uma realidade e vai continuar a se expandir e afetar a produção, pós-produção e os processos de distribuição e transmissão; e esta talvez seja a melhor maneira de apresentar necessidades, restrições e problemas na transmissão de vídeo digital em toda sua amplitude.

A televisão é colocada na forma digital pela qualidade inerente deste formato e pela ausência de degradação na replicação e na transmissão. O sinal é comprimido para que possa ser eficientemente transmitido e armazenado. O sistema de compressão ideal é o que permite que se tenha a melhor qualidade de imagem com a menor taxa possível. Os sinais de vídeo contém redundâncias que podem ser eliminadas para que se tenham transmissão e armazenamento eficientes. Imagens e movimentos complexos requerem técnicas sofisticadas para que se obtenha esta redução. A introdução de métodos de compressão mais eficientes têm permitido a

As primeiras aplicações de compressão se limitavam à edição “off-line” e processos similares utilizando sistemas de compressão JPEG, “Joint Photographic Expert Group” [10]. O sistema MPEG-2, por sua vez, foi projetado de forma a cobrir toda a gama de sistemas de vídeo para televisão convencional, incluindo aplicações de transmissão de vídeo digital por satélite, cabo e transmissões terrestres através da especificação de uma interface geral para o “bitstream” comprimido. Há interesse, portanto, em verificar como integrar esta nova tecnologia na emissora, fornecendo alternativas para a diminuição das etapas de conversão e concatenação dos codecs.

Para ilustrar a aplicabilidade deste sistema de compressão consideremos o sistema apresentado na Figura 6, representativo de uma grande rede digital, da fonte ao transmissor que usa a tecnologia de compressão digital e interface digital para a transmissão dos dados em formato comprimido. Na Figura 6 [28], denotam-se por C_n subsistemas em que os produtos com compressão podem substituir os sistemas digitais e convencionais; e n indica o nível de compressão, ou seja, quanto maior é n , menor é a taxa.

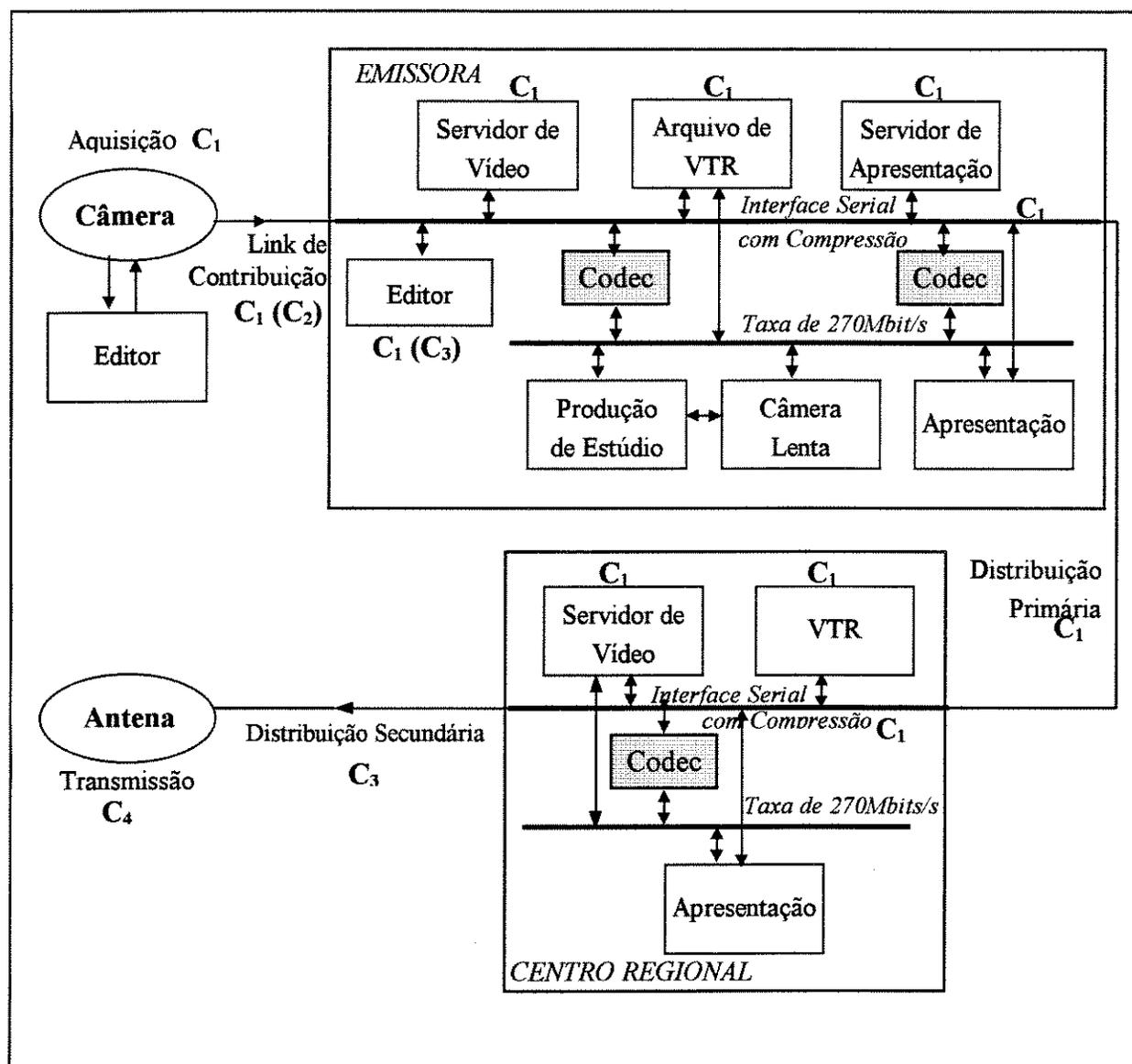


Figura 2: Cadeia de Broadcast baseada em aplicações integradas de compressão

A câmera de aquisição e o sistema de edição devem preferencialmente usar a configuração MPEG-2 422P, na qual não há subamostragem dos sinais de crominância na codificação. Idealmente todo o "link" de contribuição (C_1) deveria acomodar a taxa associada ao sistema 422P de 20 a 50 Mbit/s. Entretanto, se contribuição em tempo real for essencial e a capacidade do link for limitada o sinal deverá passar por mais uma etapa de compressão (C_2). Na emissora, o sinal restaurado ao nível de compressão C_1 , se necessário, é conectado via rede a um servidor de vídeo. Este servidor pode lidar também com materiais de outras fontes restaurados ou convertidos à taxa C_1 . As etapas subsequentes de edição podem ser conduzidas em estações de trabalho, possivelmente operando internamente com níveis de compressão elevados (C_3), porém com estágio final à taxa de compressão principal C_1 .

A cadeia de produção convencional da emissora opera a taxa de 270 Mbits/s para dar maior flexibilidade aos processamentos. Um método possível para atingir tal flexibilidade nos processamentos da emissora é via um barramento SDI (Interface Serial Digital) e um número de codecs transferindo dados entre os barramentos com e sem compressão. O número de codecs pode ser minimizado pela implantação de gravadores ou servidores de vídeo. O material de arquivo também pode ser comprimido ou não e deve ser automaticamente roteado através da rede. Da mesma forma, a área de apresentação que seleciona os programas, insere comerciais e promoções para distribuição pela rede nacional, pode também operar com material comprimido ou não, com acesso ao servidor de vídeo comprimido ou não comprimido. A saída da emissora para os centros regionais é também digital com taxa de compressão C_1 , que é factível na maioria dos meios de distribuição.

Nos centros regionais a área de apresentação é bastante similar a área da emissora. Redes locais trafegando material comprimido ou não comprimido estão disponíveis para inserção de programação regional e comerciais. A conversão dos sinais regionais de saída do formato C_1 para o formato de transmissão pode se dar na saída da central regional ou no transmissor, dependendo da capacidade do "link" de transmissão secundária e do número de transmissores. A transmissão para o usuário final (C_4) se faz utilizando níveis de codificação na faixa de 2 a 15Mbits/s, dependendo do meio de transmissão utilizado.

Este exemplo, ilustra as diversas atividades em que uma emissora pode se beneficiar pela utilização de um mesmo sistema de compressão unificado e integrado, que garanta alta qualidade de imagem e que facilite e racionalize as atividades de manutenção e treinamento. A concatenação de diferentes sistemas na cadeia de transmissão, tornou-se um dos principais problemas das redes atuais neste período de transição e convivência de múltiplas soluções digitais. A qualidade do sinal é reduzida devido à concatenação de diferentes tipos de codecs operando sobre diferentes taxas de compressão. Cada codec faz aproximações e introduz ruído que se acumula com a concatenação de etapas de codificação. O erro final será dominado pelo codec de mais alta taxa de compressão, mas, este pode ser agravado por etapas anteriores à compressão, podendo inclusive haver falhas de codificação dependendo do tipo de imagem.

2.3 Fundamentos da compressão de imagens

Os elementos chaves de um sistema de compressão são os codificadores, que realizam a compressão, transformando a informação visual original em uma seqüência de dados codificados (“bitstream”) e os decodificadores, responsáveis pela reconstrução do material visual a partir da “bitstream”. Em geral, os sistemas de compressão são assimétricos, de tal forma que a complexidade do codificador é muito maior do que a do decodificador, sendo a qualidade da imagem definida pelo codificador.

Os elementos de uma imagem, pixels, são em geral, altamente correlacionados espacial e temporalmente com seus vizinhos. Os valores dos pixels apresentam grande dispersão estatística [38]. Estas características fazem com que a representação do vídeo em sua forma original seja bastante ineficiente. O princípio básico de sistemas de compressão é converter esta fonte de entrada (pixels) em outra fonte de dados (descritores) com alta concentração estatística e baixa correlação. Esta mudança na forma de descrever a cena garante uma redução no número de bits necessários para representar a imagem. Este processo corresponde à etapa de análise no codificador, e a, de síntese, no decodificador conforme Figuras 3 e 4. O processo de análise envolve redução de redundância, sem perda de informação.

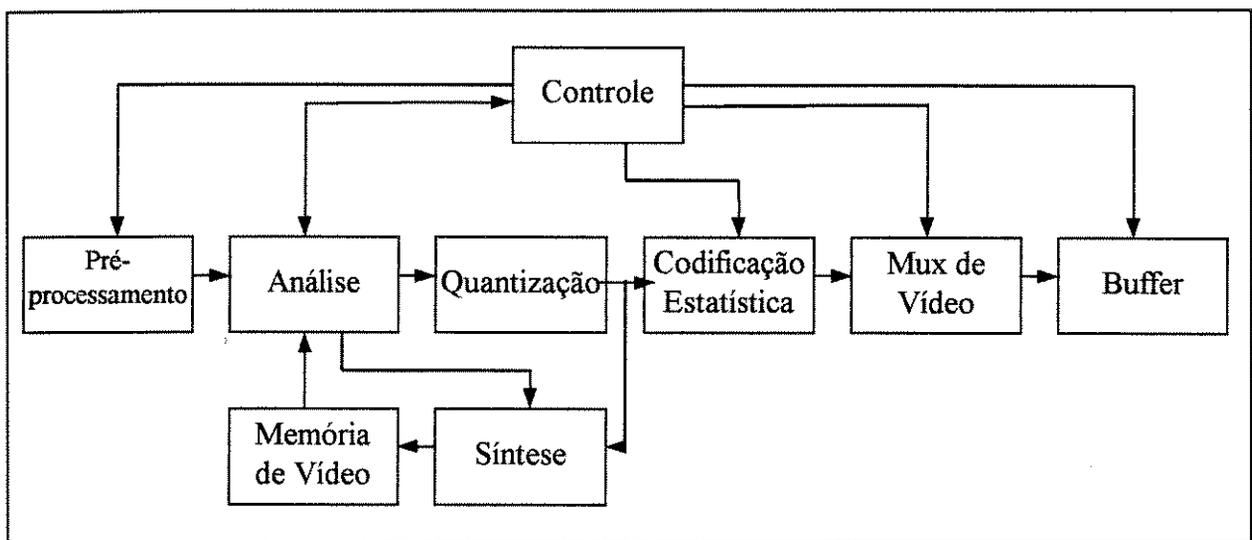


Figura 3: Estrutura do codificador de vídeo

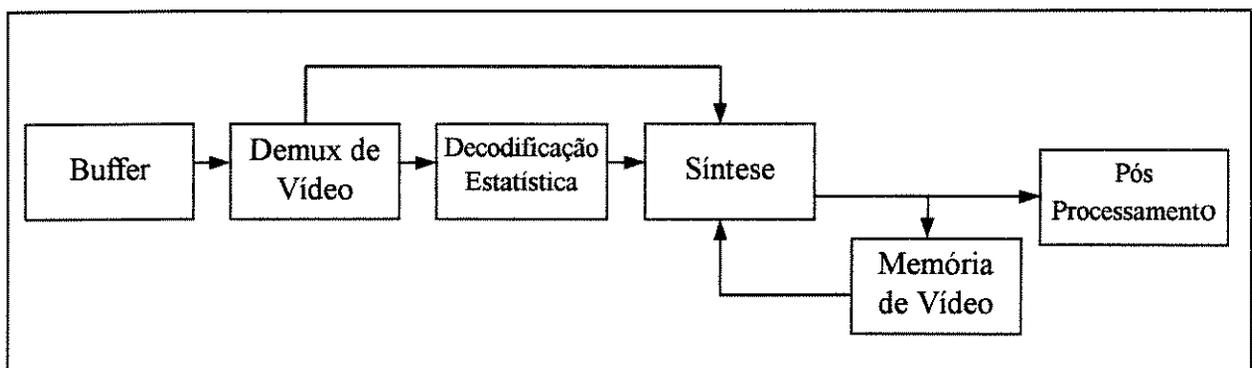


Figura 4: Estrutura do decodificador de vídeo

Uma das primeiras ferramentas de transformação da fonte foi a análise preditiva. O descritor utilizado neste caso, é o erro de predição da imagem. Esta ferramenta tem a vantagem de ser bastante simples, porém os ganhos de compressão alcançados são baixos. Em codificadores práticos, utilizam-se preditores pré-definidos, pois os preditores otimizados dependem da estatística do sinal e a diferença de desempenho entre os dois não justifica o aumento de complexidade. Os preditores de vídeo se classificam, segundo a posição espaço-temporal das amostras em preditores intra-campo, inter-campos e inter-quadros, se as amostras estão no mesmo campos, ou em campos ou em quadro consecutivos, respectivamente.

Uma melhora significativa de desempenho é obtida pela incorporação de técnicas de estimação de movimento à análise preditiva. Os vetores de movimento, obtidos por técnicas de casamento de blocos, compensam eventuais movimentos de translação da imagem. Esta informação deve ser transmitida ao decodificador como informação lateral.

Uma técnica mais eficiente é a descrição da imagem através dos coeficientes de uma transformada discreta aplicada sobre os pixels da imagem. A transformada escolhida deve ter características de decorrelação dos coeficientes e concentração da energia sobre alguns dos coeficientes. A transformada ótima, que garante total decorrelação dos coeficientes, é a transformada de *Karhunen-Loève*. Entretanto, esta transformada tem, em geral, grande complexidade, e por isso nos codificadores práticos, faz-se opção por transformadas determinísticas, não dependentes do conteúdo da imagem, que praticamente atingem o desempenho da transformada ótima para a maioria das cenas. Algumas das transformadas determinísticas mais importantes são a transformada discreta de *Fourier* (DFT) transformada discreta do coseno (DCT) e a transformada de *Hadamard* (DHT).

A transformada determinística adotada em praticamente todos os codecs padronizados é a DCT. São tipicamente utilizados blocos de 8x8 pixels. A escolha do tamanho do bloco representa um compromisso entre complexidade e adaptabilidade. Cada coeficiente transformado, carrega informações do conteúdo da imagem em alguma porção do espectro de frequências. Isto permite a priorização dos coeficientes no sentido de obter melhor casamento com a acuidade do sistema visual humano.

A aplicação da transformada proporciona decorrelacionamento espacial entre as amostras, porém persiste forte correlação entre campos e quadros consecutivos. Esta correlação temporal é explorada fazendo-se uma análise híbrida. A ferramenta de análise híbrida compreende a predição com estimação de movimento, seguida da aplicação da transformada sobre o erro de predição. A análise híbrida pode ser também bidirecional no tempo, e neste caso, a predição utilizada é progressiva e/ou regressiva, ou seja, a predição se dá com base em campos anteriores e/ou posteriores. Obviamente, no caso da aplicação de predição regressiva a ordem de codificação das seqüências precisa ser diferente da ordem de exibição. Apesar das vantagens de desempenho em compressão e em facilidade de edição, a análise bidirecional aumenta a complexidade e os requisitos de armazenamento do sistema.

À etapa de análise, segue-se a etapa de quantização onde se dá efetivamente a compressão e onde são introduzidas distorções. Os principais fatores no controle da quantização são: a estatística dos descritores da fonte, as características do sistema visual humano e a

necessidade de controle de taxa. Na codificação de imagens, os intervalos de quantização são normalmente regulares para cada coeficiente (quantização uniforme). Entretanto, os coeficientes mais importantes sofrem quantização mais refinada e os coeficientes de menor importância, uma quantização mais grosseira. A importância dos coeficientes e os intervalos de quantização são derivados empiricamente através de experiências psicovisuais e têm alta correlação com a frequência, tal que coeficientes de baixa frequência são mais importantes do que coeficientes de alta frequência.

O bloco 8x8 de coeficientes quantizados é em seguida arranjado em uma seqüência unidimensional através de uma varredura em 'zig-zag'. Como grande quantidade dos coeficientes quantizados são nulos, realiza-se uma codificação de corrida ("run length coding") desta seqüência, transformando-a em uma seqüência de pares ordenados (comprimento de corrida de zeros, coeficiente quantizado). Cada par ordenado é codificado independentemente utilizando palavras de comprimento variável, de modo que combinações mais frequentes tenham representações mais curtas, obtendo-se assim, uma redução da taxa média de bits por símbolo. Os esquemas de codificação estatística mais comuns são a codificação de *Huffman* e a codificação aritmética.

Na organização dos dados codificados estabelece-se uma sintaxe para o "bitstream". As regras de sintaxe definem tamanho, ordem e convenções relativas a todas as informações que serão multiplexadas para formar um quadro codificado de vídeo. Estas informações podem ser dados codificados provenientes do codificador, informações laterais, informações de sistema ou palavras de sincronismo.

2.4 Padrão MPEG-2: vídeo

As atividades de padronização têm grande importância pois dão incentivo à indústria e garantem a interconectividade entre os sistemas. Os padrões em codificação de vídeo têm estrutura aberta. Assim, embora a sintaxe esteja definida, a solução para a execução das funções dependem do implementador, estimulando a competição entre fabricantes. Nos últimos seis anos as atividades de padronização foram intensas, resultando em diversos padrões para aplicações em sistemas de distribuição e contribuição.

O padrão MPEG-2 [07] é uma marco da evolução das ferramentas de compressão e foi desenvolvido visando diversas aplicações, entre elas, armazenamento em meios digitais e transmissão de televisão. MPEG-2 se presta a uma larga faixa de taxa de bits, resolução e qualidade de serviço, incorporando ferramentas para tratar vídeo entrelaçado e vídeo progressivo. A sintaxe é única para todos os dados codificados, permitindo a troca de informações entre diferentes aplicações, e envolve tratamento a informações escalonáveis, permitindo a estruturação do "bitstream" em duas ou mais camadas e a criação de múltiplos serviços.

A maioria das grandes empresas e laboratórios em telecomunicações participou da elaboração do padrão concluído em 1995. Além da parte de vídeo o padrão trata separadamente

áudio e dados do sistema (transporte). As informações de sistema descrevem as facilidade permitidas em aplicações escalonáveis, cifragem, acesso condicional e múltiplos programas. O padrão especifica funções para dar suporte a redes de pacotes, para operar em canais com taxa de bits variável e garantir robustez frente a erros de bits e perda de células. O MPEG-2/vídeo tem funções particulares para vídeo entrelaçados e não entrelaçados.

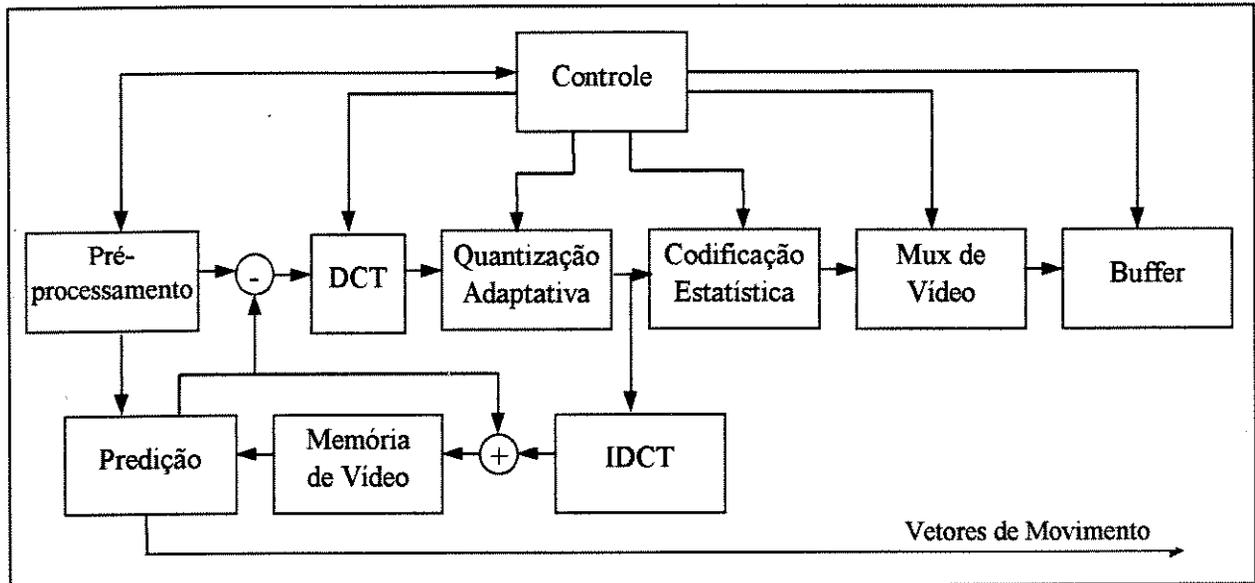


Figura 5: Estrutura de um codec MPEG-2: codificador

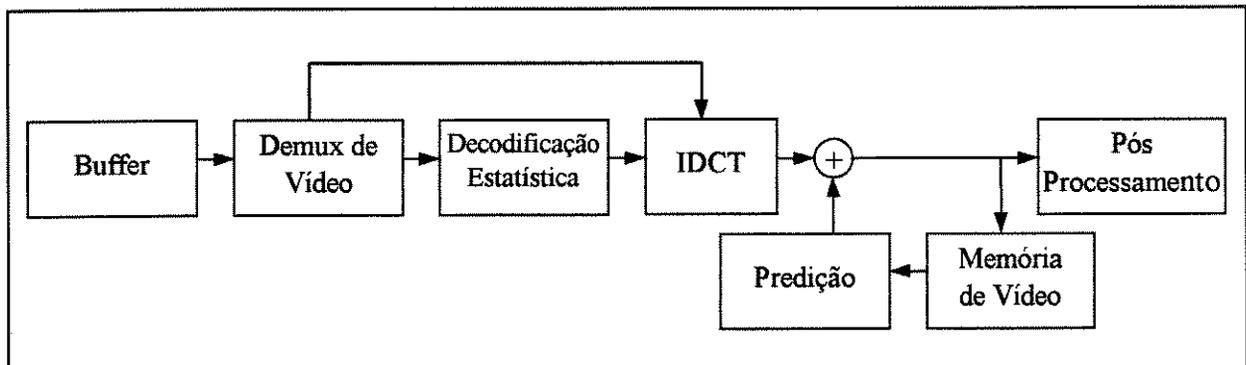


Figura 6: Estrutura de um codec MPEG-2: decodificador

O sinal de entrada deve ser vídeo digital em componentes, com uma componente de luminância Y e duas componentes de croma Cb e Cr. Os formatos de codificação utilizados são YCbCr 4:2:0, YCbCr 4:2:2 ou YCbCr 4:4:4 [03]. A menor estrutura definida é o bloco, um conjunto de 8 linhas de um mesmo campo, ou quadro por 8 amostras de uma mesma componente. A unidade básica de codificação, entretanto, é o macrobloco, que consiste num bloco formado por quatro blocos de luminância e 4 blocos de Cb e 4, de Cr, no caso de codificação YCbCr 4:4:4, 2 blocos de Cb e 2, de Cr, no caso de codificação YCbCr 4:2:2 e 1 bloco de Cb e 1 de Cr no caso de YCbCr 4:2:0. Um conjunto de macroblocos forma uma fatia. Cada fatia tem em seu início, uma palavra de sincronismo os dados codificados. Portanto, na codificação, a imagem é tomada como um conjunto de fatias [35],[41].

Dependendo da forma como são codificadas, as imagens podem ser classificadas em três

tipos. Imagens tipo I, codificadas independentemente de outras imagens, imagens tipo P, codificadas com referência a uma imagem anterior no tempo, e imagens tipo B, codificadas com referência a uma imagem anterior e a uma imagem posterior no tempo. As letras I, P e B são tomadas para codificação Intraquadro, Predição e interpolação Bilateral, respectivamente. As imagens tipo I, constituem pontos de acesso do vídeo codificado. As imagens tipo B nunca são usadas como referência na codificação de outras imagens. Para uso como referência é dada preferência a imagens tipos I e P, por minimizar o efeito de propagação de erros. Como regra geral em caso de imagens campo, os dois campos da imagem são codificados segundo o mesmo tipo, com exceção feita a imagens tipo I, na qual um campo é tipo I e o outro é do tipo P. Um conjunto finito de imagens, que normalmente contém um ponto de acesso, dado por uma imagem tipo I e que pode ser decodificada independentemente de outras imagens, é chamado de grupo de imagens(GOP). As seqüências de vídeo, estruturas de maior hierarquia, são, dependendo do contexto, conjuntos de GOPs formados por campos ou quadros consecutivos.

Na etapa de pré-processamento faz-se a necessária conversão de formato de codificação de croma, que corresponde a uma filtragem e/ou interpolação, seguida de subamostragem e a reordenação dos quadros. Esta última etapa é necessária quando se utilizam imagens do tipo B. Na etapa de pós processamento do decodificador realizam-se os processos de reordenação e conversão inversos.

A codificação de imagens empregada em MPEG-2/vídeo faz uso de técnicas de predição bidirecional, onde a imagem predita é a média entre a imagem anterior e a imagem posterior. A utilização de imagens com codificação bidirecional diminui o ruído proporcionando ganhos de desempenho significativos em baixas taxas, mas tem desvantagens de aumento de complexidade, do atraso de codificação e também dos requisitos de "buffer". O esquema básico de codificação é o algoritmo híbrido combinando DCT e predição com compensação de movimento. É, porém um sistema amplamente adaptativo e que emprega codificação baseada em contexto: há uma variedade de modos de predição e de codificação que são transmitidos ao receptor como informação lateral.

Na etapa de análise preditiva é feita a escolha do macrobloco de referência para o macrobloco atual, a ser codificado, com base no macrobloco que minimiza a potência da diferença entre os blocos. Os candidatos a macrobloco de referência dependem do tipo de imagem a ser codificada. Em imagens tipo I não há predição. Nas imagens P, o único candidato na predição é o macrobloco da imagem anterior mais casado, via estimação de movimento. Já no caso de imagens B, o macrobloco de predição pode pertencer a imagem anterior, ou posterior ou ser a média entre essas imagens.

O algoritmo de estimação de movimento define um vetor de deslocamento do macrobloco de referência mais casado com o macrobloco atual, segundo alguma função erro. Em MPEG-2 utiliza-se DCT baseada em quadro no caso de imagens com pouco movimento e baseada em campos no caso de imagens com muito movimento. Para cada macrobloco é definido um índice de quantização (MQUANT) que depende do controle de taxa e da importância relativa do macrobloco. Existem duas tabelas de codificação básica, uma para macroblocos intra (intraquadro), outra para macroblocos com codificação não intra que são projetadas de acordo

com a importância relativa do coeficiente transformado. Multiplica-se ao valor da tabela, o parâmetro MQANT que pode assumir valores de 2 a 62 em escala linear em situações de taxas baixas ou moderadas, e valores de 1 a 112, não linearmente distribuídos no caso de altas taxas. A etapa seguinte consiste em codificar os coeficientes quantizados por corridas de zeros. Se um macro bloco tiver todos seus coeficientes quantizados iguais a zero, então não é transmitido.

Muitas informações são codificadas por palavras de tamanho variável. Entre elas temos, símbolos produzidos pela codificação de corrida de zeros efetuada nos coeficientes quantizados, vetores de movimento, tipo de macrobloco e mudanças no valor do parâmetro MQANT. As tabelas contendo as palavras de comprimento variável são fixas e não podem ser alteradas pelo usuário.

A função do “buffer” é acomodar a taxa variável de codificação à taxa fixa do canal. O tamanho do “buffer” depende da configuração MPEG-2 adotada. No nível principal (“Main Level”) o tamanho máximo do “buffer” é de 1,8Mbits. Para controle do “buffer” são estabelecidas medidas de complexidade para cada tipo de imagem a partir de imagens anteriores. No início da codificação da próxima imagem; define-se uma meta em número de bits para a codificação da mesma com base na informação de complexidade e do número de bits disponíveis até o final do GOP objetivando manter a qualidade constante em todo o GOP. A estratégia para alocação dos bits consiste em verificar se há uma diferença significativa entre o número de bits já codificados e a quantidade de bits planejada. Se esta diferença é positiva, o valor do parâmetro MQANT é aumentado e, se é negativa, diminuído. Esta estratégia visa manter constante o número de bits por GOP, para que antes da codificação da primeira imagem tipo I do próximo GOP, o “buffer” de transmissão esteja praticamente vazio. Este equilíbrio é obtido se o número médio de bits codificados por intervalo de quadro é igual ao número de bits lidos do “buffer” no mesmo intervalo.

Diferenças na implementação das funções de codificação faz com que as diferenças de desempenho sejam significativas. De fato, o padrão não garante qualidade. As principais funções na determinação da qualidade são, estimação de movimento tanto o algoritmo como a área de busca, o esquema de controle de taxa e as etapas de pré e pós-processamento.

2.5 Caracterização de artefatos associados à compressão

Os sistemas de compressão exploram tanto as propriedades psicovisuais do sistema visual humano como a redundância estatística dos dados da imagem, para conseguir redução da taxa. Entretanto, alguns artefatos podem ser introduzidos no processo de compressão. Estes artefatos são, em alguns casos, perceptíveis e perturbam o usuário, embora um dos objetivos dos sistemas de compressão seja escondê-los.

Os sistemas de vídeo digital produzem degradações que são fundamentalmente diferentes dos artefatos produzidos pelos sistemas analógicos. Estes artefatos podem ser introduzidos através de filtragem, conversão, transformação, quantização e/ou transmissão. Perda de resolução, efeito de bloco, ruído de quantização e erros de bloco são alguns dos indesejados

efeitos do processo da compressão. Essas degradações são influenciadas pelo algoritmo de compressão utilizado, conteúdo da imagem, origem do material-fonte e condições do canal de transmissão. Um aspecto que torna estes artefatos tão diferentes das distorções analógicas é o fato de que estes são influenciados pelo conteúdo da imagem e, portanto, são variantes no tempo fazendo com que algumas cenas sejam mais difíceis de serem comprimidas do que outras.

Vale observar que embora o vídeo comprimido seja geralmente degradado em relação ao original, alguns processos podem reduzir o número de artefatos existentes no material. Isto significa, portanto, que para alguns artefatos a codificação pode significar uma melhora na imagem. É o caso em que o sinal original já está degradado, e, neste caso, pode acontecer, por exemplo, do material comprimido ser menos ruidoso do que o sinal original.

2.5.1 Blocagem

O artefato conhecido como efeito de bloco ou blocagem [09],[69] é uma deterioração, na qual a imagem recebida apresenta padrões retangulares que não estavam presentes na imagem original. As discontinuidades se tornam visíveis quando a quantização é pouco refinada. Este tipo de artefato está presente em cenas complexas ou quando elevadas taxas de compressão são necessárias. Nesta situações o “buffer” do codificador tende a encher e a quantização é ajustada para ser menos refinada e evitar transbordo do “buffer”. Para dada granularidade de quantização, a blocagem é mais perceptível em áreas mais “planas” da figura.

2.5.2 “Blurring”

A perda de resolução é uma degradação caracterizada pela perda, ou quantização pouco refinada, das componentes de alta frequência. As componentes de alta frequência têm, usualmente, menos energia e são menos perceptíveis ao olho humano. Esta informação é descartada para que se atinja a taxa desejada. Na prática, em uma dada situação de taxa, há um compromisso entre alta resolução e baixo ruído. Este ajuste é subjetivo e depende da aplicação.

O efeito conhecido como “blurring” é a redução da definição em bordas e áreas com muitos detalhes espaciais. Em cenas de alta resolução espacial este efeito é consequência do compromisso na alocação de bits para a descrição dos detalhes de alta resolução e na descrição do movimento. No caso geral, é a etapa de pré-filtragem, a maior responsável pela perda de resolução. Se a imagem for convertida de um formato com menor número de linhas para um formato com mais linhas, então, ela aparecerá borrada, antes mesmo da compressão.

Perda de resolução ou erros de crominância em vídeo digital comprimido podem ser mais perceptíveis que os demais artefatos, pois a maioria dos sistemas emprega uma subamostragem na crominância. Na presença de objetos coloridos em movimento rápido, a subamostragem da crominância pode levar a um descasamento entre luminância e crominância. Essa perda de resolução torna os demais artefatos mais visíveis.

2.5.3 “Edge Business”

O artefato conhecido como “*edge business*” é causado, geralmente, pela quantização de um bloco contendo pixels que delimitam áreas planas com diferença significativa entre os níveis médios em cada uma das bordas. O resultado é o aparecimento de uma distorção concentrada nas bordas dos objetos, caracterizada pela variação temporal ou espacial da borda.

2.5.4 Ruído

Há uma grande variedade de degradações associadas ao ruído produzido pelos algoritmos de compressão. Os dois tipos mais importantes são ruído de quantização e ruído mosquito. O ruído de quantização é um dos artefatos predominantes. Na compressão, a quantização é parte fundamental do processo e, em muitos sistemas, a quantização é o único processo que introduz distorções. Este ruído tende a ser descorrelacionado do sinal, mas não é uniformemente distribuído ao longo da imagem. No caso da utilização de quantização adaptativa por blocos, como é o caso de MPEG-2, ele pode estar associado ao conteúdo local da imagem.

O ruído mosquito é um tipo de “*edge business*” caracterizado por artefatos em movimento ou padrões de ruído em blocos superpostos aos objetos. Na recomendação ITU-T P.930 [28] é utilizado o termo ruído correlacionado para designar o efeito resultante da combinação de “*edge business*” e ruído mosquito, já que, em geral a caracterização das diferenças subjetivas entre os dois efeitos é confusa. A distorção ruído correlacionado é, portanto, tipicamente visualizada ao redor de bordas em movimento. Este artefato se caracteriza por um “halo” superposto ou próximo às bordas de objetos em movimento.

2.5.5 “Jerkiness”

“*Jerkiness*” é o artefato no qual o movimento originalmente suave e contínuo, é percebido com uma seqüência de cortes bruscos. No caso de sobrecarga no codificador a imagem pode até mesmo ficar congelada. Tipicamente o “*jerkiness*” pode ser observado em codecs operando a muito baixas taxas. Este efeito é resultado da repetição de quadros para reduzir a quantidade de informação de vídeo. A baixa taxa de quadros resultante é indicativa do nível desta degradação.

2.5.6 Blocos Errados

Devido a erros do canal, pode-se ter no vídeo reconstruído a presença de *blocos errados*. O vídeo digital comprimido é muito susceptível a erros de canal, devido à utilização de ferramentas de codificação preditiva. Dependendo do tamanho do GOP e do tipo de imagem afetada, um bloco errado pode se propagar por diversos quadros. Este espalhamento temporal é mais severo em imagens tipo I e em estratégias de codificação que utilizam GOPs grandes. Também dependendo da estratégia de predição intraquadros e da compensação de movimento, a transmissão de um bloco errado pode contaminar uma grande porção dentro de um mesmo quadro (espalhamento espacial).

3. Avaliação de sistemas de compressão

A compressão de sinais de vídeo é uma solução importante em vista dos requisitos de serviço e das taxas disponíveis. Para atender a demanda crescente diversos sistemas foram desenvolvidos para diferentes aplicações. A taxa de compressão e a qualidade de imagem associada são determinadas pela aplicação e pelas exigências do usuário. Há, portanto, uma necessidade crescente em se avaliar de forma precisa o desempenho destes sistemas. As formas de fazê-lo são alvo de discussões e análises deste capítulo. A avaliação é atividade crítica na monitoração, diagnóstico, bem como na seleção e manutenção de sistemas.

3.1 Insuficiência de testes analógicos

O advento dos sistemas de compressão, armazenamento e transmissão de vídeo expôs as limitações das técnicas e metodologias que tradicionalmente têm sido utilizadas para medir o desempenho de sistemas de vídeo. Tradicionalmente as medidas de desempenho se baseiam na constância de desempenho dos sistemas de vídeo para diferentes cenas de entrada. Assim, poder-se-ia injetar um sinal de teste (ex: "zoneplate" estático) e medir algum atributo do sistema (ex: resposta em frequência) e estar relativamente seguro de que o sistema iria se comportar da mesma forma para outro material de vídeo (ex: vídeo com movimento). Grande esforço foi feito para relacionar as medidas tradicionais de desempenho dos sistemas analógicos (ganho diferencial, distorção de fase, distorção de forma de onda) à qualidade percebida da imagem. Embora o recente advento dos sistemas de transmissão de vídeo, armazenamento e transmissão digitais não tenha invalidado o uso destes parâmetros, tornou sua correlação com os resultados subjetivos mais tênues.

Os sinais de teste analógicos empregam sinais de teste adequados para definir as características da resposta contínua e linear dos componentes analógicos. Muitos dos sinais de teste foram projetados para fornecer uma caracterização baseada em características subjetivas, porém num tempo em que as distorções se limitavam a linearidade dos parâmetros analógicos, com baixa relevância no domínio temporal.

Codecs empregam algoritmos complexos que incorporam processamento espacial e temporal. Em consequência disso, eles apresentam respostas não lineares a mudanças no conteúdo das cenas e das condições de operação. A resposta do sistema a um particular sinal de teste descreve somente seu desempenho para aquele sinal de teste, e, em geral, tem pouca ou quase nenhuma relação com o desempenho para outros sinais de vídeo. Os sinais analógicos são úteis para se testar o processamento analógico de um codec, mas pouco se pode inferir sobre o comportamento do codec em situações de movimento complicado. Testes dinâmicos são essenciais para codecs de alto desempenho, que empregam técnicas de compensação de

movimento.

A natureza estática dos sinais de teste não exercitam completamente os algoritmos de compressão. Mesmo quando os sinais são gerados digitalmente e interfaceados com o sistema em teste, há pouco ruído de quantização para estressar os sistemas. Sinais de alta energia no domínio da frequência como sinais “zoneplate” podem revelar as características de pré-filtragem, ou em caso de sistemas de compressão simples, revelam as limitações de codificação e transbordo do “buffer”. Apesar disso, os sistemas automáticos de teste convencionais indicam, normalmente um desempenho exemplar, mesmo em casos de equipamentos que apresentam problemas em nível subjetivo quando submetidos a programação normal.

Outro aspecto em que sistemas analógicos e digitais diferem, diz respeito à degradação da qualidade da imagem recebida em função da gradual degradação do sinal. Em consequência da utilização de técnicas de compressão de imagens e de códigos corretores de erros, sistemas de transmissão digitais têm excelente qualidade de imagem ou, em condições marginais, nenhuma imagem. Assim, requerem cuidados adicionais na monitoração da transmissão, já que sua degradação pode não se refletir imediatamente na qualidade da imagem.

3.2 Escopo e objetividade dos métodos de avaliação

Na definição da qualidade de uma imagem é útil se fazer uma distinção entre ambientes orientados ao desempenho e ambientes não orientados a desempenho. No primeiro caso, em ambientes orientados ao desempenho, (como, por exemplo, para assistir televisão por diversão), há interesse na utilização desta imagem ou no julgamento da imagem pelo observador. Neste caso, a forma de avaliação encontrada é realmente a opinião de um grupo de avaliadores. Para que se possa opinar sobre os méritos de determinado sistema é necessário considerar a importância relativa de cada fator de qualidade e estabelecer critérios aceitáveis para cada caso.

Uma avaliação global deve incluir o desempenho do sistema de codificação propriamente dito (qualidade básica), o desempenho após sucessivos processamentos e deve levar em conta que seqüências críticas são utilizadas na avaliação dos sistemas. Outro aspecto importante é a forma pela qual erros de transmissão afetam o sistema, ou seja a avaliação de características de falha do sistema. A passagem do sinal com taxa reduzida de bits através de um meio de transporte pode causar erros que podem ocorrer isoladamente ou em blocos. A habilidade do sistema em tolerar tais erros, incluindo erros de canal corrigíveis ou não, e a forma pela qual essas degradações repercutem na qualidade final da imagem deve ser parte da avaliação dos sistemas. Em geral os erros de bit se manifestam como ruído na imagem. Esses erros podem se espalhar espacial e temporalmente pela imagem, dependendo da técnica de codificação utilizada.

A avaliação subjetiva, quando comparada a avaliação objetiva tem grandes desvantagens em todos os aspectos operacionais. Enquanto a avaliação subjetiva é relativamente lenta e de alto custo, de difícil reprodução e fornece apenas medidas qualitativas, as medidas objetivas são quantitativas, reprodutíveis e muitas vezes podem ser efetuadas em tempo real. A grande diferença entre estas duas abordagens está no fato de que objetivamente é possível avaliar apenas

as características físicas da imagem, enquanto a avaliação subjetiva é capaz de medir diretamente as características *percebidas* da imagem.

As características do espaço de percepção humana não são totalmente caracterizadas por medidas objetivas. A compressão de vídeo tira vantagem das características do sistema visual humano descartando componentes da imagem que são menos perceptíveis. Para atingir estes objetivos, técnicas como codificação menos refinada em alta frequência, subamostragem componentes de cromaticidade, codificação de diferença entre campos e quadros consecutivos e codificação adaptativa baseada no conteúdo da imagem são largamente utilizadas pelos sistemas de compressão. A compressão procura tirar proveito do efeito de mascaramento, no qual defeitos da imagem são escondidos devido ao conteúdo da imagem ou devido a cortes de cena, o que é dificilmente modelado por medidas objetivas. Desta forma, a confiabilidade das medidas objetivas na previsão de resultados subjetivos é bastante questionável.

A combinação de testes subjetivos e objetivos é um meio efetivo de se testar um sistema codificador decodificador. Os testes subjetivos formais são úteis para se obter uma medida qualitativa da qualidade da imagem, mas são em geral de alto custo, consomem bastante tempo e não são repetíveis. Testes objetivos, baseados em cálculos por computadores, permitem a avaliação de um grande número de parâmetros e estatísticas, medidas quantitativas precisas e em alguns casos extrapolação dos dados subjetivos.

A vantagem dos testes subjetivos é a capacidade de medir a qualidade perceptível da imagem diretamente, apresentando-se, portanto, como o método bastante efetivo e confiável de se medir a qualidade de vídeo comprimido. Dada a natureza *subjetiva* dos testes, os resultados podem ser influenciados pelo gosto dos observadores, experiência, “background” cultural, idade e condições de saúde. Também há influência das características de projeto do teste, como: apresentação, instruções, material de teste, escala de degradação de imagem, distância de visualização, equipamento, procedimento do “setup” e análise de dados. Pelo fato de consumirem muito tempo, os testes formais são, em geral, precedidos por testes pilotos, cujos resultados são utilizados para o projeto do teste formal de forma mais econômica e útil possível.

4. Avaliação subjetiva: Projeto de experimento

Com o crescente interesse na implementação de sistemas de compressão digital para contribuição, distribuição e armazenamento, atenção deve ser dada à avaliação de seus desempenhos. No estudo de técnicas de processamento de sinais deve-se levar em consideração o usuário final do serviço, a quem compete julgar a qualidade final do sinal. Para que se tenham melhores garantias de sucesso em termos subjetivos, as características dos sistemas de compressão e os artefatos associados devem ser cuidadosamente considerados e avaliados. A caracterização de parâmetros que afetam a qualidade perceptível do sinal pressupõe resultados subjetivos confiáveis e reproduzíveis.

Os procedimentos de teste diferem bastante se o objetivo é testar a qualidade ou a inteligibilidade do sinal digitalizado. A classificação de um codec utilizado em aplicações de distribuição se refere a classificação feita sobre a imagem decodificada após uma passagem pelo codec. Para qualidade de contribuição, entretanto, deve-se analisar a qualidade básica após vários codecs colocados em série (configuração em cascata) de forma a simular uma aplicação de contribuição típica. Em geral, os problemas de especificação de qualidade de imagem têm sido considerados separadamente para sistemas analógicos e digitais. Se de fato existir uma independência psicológica entre as degradações, então elas poderão ser julgadas como degradações causadas pela seção digital, pela seção analógica, ou ainda degradações causadas simultaneamente pelas duas seções. Testes subjetivos são parte importante também na etapa de projeto de codificadores. Mesmo em situações em que medidas objetivas que apresentam boa relação com a qualidade percebida da imagem estão disponíveis, os testes subjetivo são usados para complementar, verificar ou calibrar tais medidas objetivas.

O objetivo destes testes é estabelecer, através de métodos empíricos, bases para os julgamentos. É essencial que o método de medida usado leve a resultados válidos e representativos das opiniões. A Recomendação ITU-R BT. 500-7 [02] é largamente usada para o projeto de avaliação subjetiva de sistemas de compressão de vídeo. Embora a mesma metodologia possa ser usada para os diferentes tipos de codecs, que usam codificação intra-campos ou inter-quadros, usando palavras de comprimento fixo ou variável, as imagens de teste escolhida devem ser diferentes e isso terá grande influência sobre o resultado final do teste. Os métodos existentes não estabelecem níveis absolutos de qualidade. Ao contrário, os resultados obtidos são sempre influenciados pela escolha das imagens de referência.

Além da Recomendação ITU-R BT.500 [02] que se aplica a televisão digital de resolução convencional, existe um conjunto de normas do ITU-R que provêem instruções e métodos disponíveis para o acesso à qualidade da imagem em um ambiente de laboratório controlado. A Recomendação ITU-R BT. 710 [04] trata especificamente de televisão de alta definição e a Recomendação ITU-T BT. 910 [08] faz referência à avaliação subjetiva em aplicações multimídia. Neste capítulo, nosso interesse está voltado para a análise das limitações e

problemas dos métodos padronizados e a sondagem de formas alternativas para a avaliação.

4.1 Fundamentos da metodologia de avaliação

Dentre os vários tipos de métodos psicométricos, três tipos básicos têm sido utilizados em televisão. O primeiro é o método de comparação, no qual a magnitude de um tipo de degradação é variada até provocar o mesmo efeito de uma outra degradação mantida fixa, tomada como referência. O segundo tipo é o método discriminatório que procura, por exemplo, estabelecer a magnitude na qual a degradação se torna visível ou determinar a seqüência de pequenas variações de degradação que produzem as menores diferenças perceptíveis. O terceiro é o método de escalonamento de opiniões, no qual, o caso mais simples, diferentes magnitudes de uma degradação são aplicadas a uma imagem em ordem aleatória. Em seguida observadores classificam cada uma das imagens a partir de uma escala de categorias. Cada um destes métodos têm seu campo de aplicação e muitas variações possíveis no que diz respeito às técnicas experimentais e à forma de análise e apresentação dos resultados [54].

Em geral os dados relativos a testes psicométricos são escassos, pois os testes são inevitavelmente de alto custo e demandam muito tempo. A diversidade de métodos empregados dificulta a comparação de resultados obtidos em diferentes laboratórios, o que é bastante ruim tendo em vista a dificuldade na obtenção de dados. Apesar disto, não há um consenso sobre o melhor método a ser utilizado, pois a determinação de forma quantitativa, das vantagens e desvantagens da utilização de qualquer dos métodos psicométricos, depende da realização de testes em larga escala.

As normas de avaliação subjetiva tratam dos aspectos relativos às condições de observação, seleção e preparação do material, número e tipo dos avaliadores enfatizam, principalmente, a seleção da metodologia adequada ao problema, com relação à seleção da forma de apresentação e ao tipo de escala de notas. A quantidade de dados que precisa ser coletada depende de fatores intercorrelacionados como: intervalo de confiança desejado, desvio padrão dos dados e magnitude relativa do efeito que se deseja medir [01]. Embora os propósitos do teste delimitem a escolha do método, nem sempre é obvio que critério de julgamento deve ser adotado. Em geral, se o projetista do experimento gera diferentes níveis de degradação na imagem, a diferença entre a imagem degradada e a imagem original é o critério relevante e pode-se usar uma escala de degradação. Se por outro lado, não se está forçando uma degradação na imagem, mas testando diferentes algoritmos, a qualidade é o critério relevante.

Este trabalho analisa diversos métodos capazes de fornecer informação sobre a reação do observador sobre larga faixa de magnitudes e diferentes tipos de degradação. Na seleção de uma escala, o principal aspecto é o casamento entre as necessidades do teste, a tarefa dada ao observador e sua capacidade de desempenhá-la. Todos os detalhes devem ser decididos tendo-se em mente, a motivação e satisfação dos observadores. As instruções dadas aos observadores bem como a escala propriamente dita devem ser simples e não causar ambigüidade.

Para o número de pontos na escala, cinco é em geral considerado adequado à capacidade

de observadores não especialistas. A popularidade de escalas de cinco pontos em outras áreas serve de suporte para esta escolha. É recomendado evitar qualquer efeito indesejado pelo uso de números seja 0 a 4 ou 1 a 5. Os adjetivos usados para descrever cada ponto da escala também impõem distorções a escala e devem ser usados como orientação e não como definições rígidas. Levando-se em conta o efeito de condicionamento, é essencial que a faixa de degradação utilizada em cada sessão seja tal que a maioria dos observadores faça uso de toda as notas da escala.

A variação perceptível na qualidade da imagem por determinado tipo de degradação depende fortemente da natureza da imagem, evidenciando a necessidade de escolha da imagem em função do tipo de degradação que se deseja avaliar. Os testes de avaliação subjetiva pressupõem, em geral, condições de teste mais estressantes do que a média.

Na análise dos resultados de um experimento, o objetivo principal é estimar da melhor forma possível a distribuição de opinião para uma determinada condição da imagem. É claro, que uma relação exata não pode ser obtida dada a incerteza que é inerente ao processo de amostragem. Dado que o grupo de observadores escolhido é simplesmente uma amostra da população, os resultados obtidos para um grupo são diferentes dos resultados obtidos para outro grupo. O estudo das funções de distribuição e as técnicas de ajuste de curva são metodologias consideradas eficientes e úteis. A relação entre a distribuição de opinião e a nota média depende de características da classe de observadores utilizada, de tal forma que, o modelo que vale para um grupo de especialistas pode não se aplicar a um grupo de não especialistas. Mesmo dentro de um grupo podem haver diferenças dependendo do nível social e instrução.

4.2 Impacto das condições de observação

Algumas características básicas são comuns a todos os procedimentos subjetivos utilizados para medir qualidade de imagem de televisão, embora existam algumas diferenças entre as condições de visualização para televisão com resolução convencional e para televisão de alta definição.

Todos os parâmetros descritos têm influência sobre a avaliação e devem ser mantidos tanto quanto possível próximos aos valores padronizados para permitir a comparação entre resultados obtidos por diferentes laboratórios.

4.2.1 Condições de visualização padronizadas

Os parâmetros aqui apresentados são um resumo dos valores estabelecidos pela ITU-R BT. 500 [02] e ITU-R BT. 710 [04].

Condições gerais

<ul style="list-style-type: none"> razão entre a distância de visualização e a altura da tela. 	4H-6H (TV Standard) e 3H (HDTV)
---	------------------------------------

Ajustes do monitor

<ul style="list-style-type: none"> luminância de pico 	70 cd/m ²
<ul style="list-style-type: none"> razão entre a luminância da tela do tubo inativo e a luminância de pico 	≤0,02
<ul style="list-style-type: none"> razão entre a luminância da tela, quando mostrando somente o nível preto em uma sala completamente escura, e a correspondente ao nível de pico de branco 	0,01

Condições da sala

<ul style="list-style-type: none"> razão entre a luminância do fundo atrás do monitor e a luminância de pico da imagem 	0,15
<ul style="list-style-type: none"> outra iluminação da sala 	baixa

Condições Especiais

<ul style="list-style-type: none"> número de avaliadores por monitor 	5 (distribuição homogênea a 4H e 6H)
<ul style="list-style-type: none"> monitor 	alta qualidade com 54 cm para TV convencional e 140 cm para HDTV
<ul style="list-style-type: none"> brilho e contraste do "display" 	acertado via sinal PLUGE
<ul style="list-style-type: none"> natureza da sala de visualização 	sala com três paredes na cor 'gelo' e uma parede na cor cinza atrás do monitor

4.2.2 Influência da distância de observação e do tamanho da imagem

Embora todos os parâmetros tenham influência, estes dois aspectos, distância de observação e tamanho da imagem, são analisados mais profundamente com base nos resultados obtidos no estudo de caso (Capítulo 6) e nos resultados dos trabalhos de Westerkink [64].

A sensibilidade do olho humano é influenciada por componentes da imagem como: frequência, cromaticidade, movimento, contraste e brilho. Testes nesta área mostram, por exemplo, que a visibilidade do ruído varia de acordo com o nível de brilho da imagem. Da mesma forma, verifica-se um mascaramento das degradações nas regiões onde há alta atividade espacial e temporal da imagem. Sem dúvida, entretanto, as características mais importantes para a detecção humana são detalhes locais, como bordas e contornos. Além destas características

espaciais e temporais da imagem, diversos fatores relacionados às condições de visualização também influenciam de forma decisiva a qualidade de uma imagem.

A influência da distância se mostrou significativa no experimento de avaliação subjetiva realizado. Calculando-se separadamente as médias das notas dadas pelos avaliadores a uma distância de 4H (H=altura do monitor) do monitor e as notas a 6H, a diferença relativa foi de 2,5%. Embora percentualmente pequena, esta diferença não é desprezível, pois as mesmas pessoas deram notas em uma e outra distância.

Em trabalho análogo, Westerink [64] procura relacionar a qualidade de imagem com a largura, resolução e distância do observador. Fixada a altura da imagem e dada certa largura e resolução, a distância do observador influencia a qualidade de duas formas. Primeiramente, aumento da distância tem um efeito negativo sobre a qualidade da imagem pois o ângulo da imagem diminui. Porém, um segundo efeito está relacionado ao fato de que a resolução angular aumenta com o aumento da distância, e desta forma a qualidade da imagem melhora. Segundo este estudo, a qualidade subjetiva da imagem na distância ótima é independente da largura da imagem e depende linearmente com o logaritmo da resolução (relacionado a banda da imagem).

Uma crítica que pode ser feita ao estudo deste autor é a utilização exclusiva de imagens paradas, pois sua extensão para o contexto de televisão não é direta e deve ser feita criteriosamente. No caso de imagens paradas, a sensibilidade a detalhes locais é um pouco maior. Assim, no que concerne à influência dos detalhes, as imagens paradas podem ser consideradas uma aproximação de pior caso de imagens com movimento.

Resultados relativos à determinação da relação distância resolução ótima têm especial interesse em termos das conseqüências da conversão do sinal de TV para o sinal de "High Definition", HDTV. Em "displays" de alta resolução pode-se melhorar a qualidade da imagem, aumentando-se a largura da imagem. O aumento de resolução por si só não traz melhorias para a imagem, pois há uma saturação da resolução. Dependendo da distância em que o telespectador costuma se posicionar em relação ao televisor, a conversão de TV para HDTV não terá absolutamente nenhum efeito sobre a qualidade da imagem se não for combinada por um aumento na largura da tela.

4.3 Análise de variabilidade entre observadores

Para se garantir um nível de confiabilidade em testes de avaliação subjetiva é preciso envolver um número mínimo de avaliadores, ou seja, uma amostra que seja representativa. Como regra geral [02] devem ser utilizados pelo menos 15 avaliadores. Este valor pode ser usado como um ponto de partida, porém mais pessoas devem ser envolvidas sempre que a análise dos dados revelar que o nível de confiabilidade desejado não foi atingido.

A experiência do avaliador na determinação de qualidade da imagem é um fator de variabilidade nos resultados. Os avaliadores são separados em especialistas e não especialistas de acordo com seu envolvimento com a avaliação de qualidade de imagem como parte do trabalho diário. A experiência do avaliador na classificação de imagens interfere no nível de significância.

O nível de significância, importante parâmetro na caracterização das diferenças entre as imagens, está relacionado à habilidade dos observadores em perceber as diferenças entre as imagens e é obviamente maior no caso de avaliadores não especialistas.

A escolha entre avaliadores especialistas ou não especialistas deve se feita com base nos objetivos dos testes subjetivos. No caso de um experimento para a otimização de parâmetros de um novo sistema de compressão, um pequeno grupo de avaliadores especialistas deve ser selecionado. Por outro lado, um experimento visando o estudo da percepção de artefatos, ou a determinação de qualidade de novos sistemas frente aos atuais, deve envolver um grande número de avaliadores não treinados, ou não especialistas.

Após serem expostos aos artefatos, com o tempo, os usuários podem se acostumar ou se tornarem mais aversos às degradações. Também em função disso, a escolha entre especialistas e não-especialistas em testes subjetivos deve ser cuidadosamente considerada. Comparações entre diferentes grupos de avaliadores é objeto do trabalho de Glasman et al. [31]. Seu resultado mais interessante é a comprovação de que a influência do auto aprendizado durante os experimentos de avaliação subjetiva não é significativo.

A experiência adquirida com a realização dos experimentos de avaliação subjetiva mostrou que os observadores podem não ser capazes de detectar os artefatos, uma vez que não estão familiarizados com os algoritmos de compressão, muito diferentes dos artefatos introduzidos pelo sistemas NTSC e PAL-M convencionais. No experimento 1, descrito em detalhes no Capítulo 6, verificou-se uma diferença entre as notas médias de especialistas e não especialistas de 7,1%, considerada bastante significativa.

4.4 Seleção do material de teste

Um material de teste bem escolhido faz com que o teste subjetivo seja mais confiável o que torna a escolha das cenas de testes uma importante etapa no projeto de avaliação. O primeiro aspecto a ser considerado é o número mínimo de cenas que devem ser envolvidas no teste. Este número fica limitado pela duração máxima do teste, pelo número de níveis de degradação que se deseja testar e pela necessidade de redundância, mas não devendo ser inferior a três. Entretanto, na aplicação de métodos que utilizam comparação de estímulos, envolvendo sistemas com pequenas diferenças de desempenho, o número de cenas envolvidas não deve ser inferior a 6.

Em qualquer sistema de compressão, as informações relativas à percepção temporal e espacial são parâmetros críticos. Estes parâmetros têm um papel importante na determinação do nível de compressão possível do sinal de vídeo e, conseqüentemente, das degradações introduzidas ao se codificar uma seqüência para transmissão em canais de taxa fixa.

Devido à flexibilidade dos codecs, a distorção introduzida por cada amostra codificada depende não só dos valores das amostras em uma mesma localização, mas também, de amostras anteriores que podem se estender por um ou mais quadros. Isto significa, que não faz sentido tentar caracterizar um codec deste tipo provocando distorções locais, não importando se a

codificação é intra- ou inter-quadros. É preciso se ter um conhecimento das características de adaptação do codec quando excitado por cenas naturais para que se possa determinar objetivamente seu desempenho.

Levando em consideração a capacidade do codec em distribuir os bits ao longo da imagem, o conteúdo de cada porção da imagem deve ser analisado cuidadosamente, para que se possa julgar sua “criticalidade”. As medidas objetivas devem estar baseadas no erro de codificação do codec quando submetido a seqüências de testes naturais. Pelo fato de envolver processos adaptativos, alguns artefatos dependem do conteúdo da cena. Estas degradações são funções não lineares da velocidade e dimensão do movimento, complexidade da cena, saturação das cores e variam com o tempo.

Algumas cenas são difíceis de serem codificadas com alguns algoritmos de compressão particulares e com certas taxas de compressão, de forma que, o material de teste deve explorar toda a capacidade do sistema. Em sistemas de compressão que empregam codificação inter-quadro ou inter-campo, cenas com movimentos complexos e com detalhes finos são mais difíceis de serem codificadas do que imagens paradas. Isto não significa, entretanto, que qualquer tipo de imagem parada será exibida sem artefatos. Devido aos mecanismos de “refresh”, necessários para reduzir a propagação de erros, artefatos comuns em compressão, como blocagem, podem se tornar visíveis.

Alguns algoritmos são sensíveis a cortes de cena e, nestes casos: perda de resolução, blocos residuais do campo anterior e efeito de bloco podem ser visíveis na transição entre imagens paradas com detalhes fino e imagens com movimentos complexos ou vice-versa. Geralmente, algoritmos de compressão são sensíveis a ruído no material fonte. O ruído reduz a correlação dos sinais existentes em imagens típicas e confunde os algoritmos de compensação de movimento. Essa informação descorrelacionada, submetida ao codificador, acentua o efeito dos artefatos.

Uma vez que a cromaticidade é subamostrada e quantizada com menor refinamento, a análise cuidadosa do desempenho da cromaticidade é necessária. Para tal devem ser incluídos nos testes seqüências com larga gama de cores, alta resolução e larga faixa dinâmica. Para examinar o descasamento entre luminância e cromaticidade, deve-se usar seqüências de teste que apresentem movimentos rápidos o suficiente para que a posição dos objetos coloridos, na horizontal e na vertical, seja bastante diferente de um campo para outro.

O uso de técnicas de amostragem para a aquisição de material que seja estatisticamente representativo do programa de interesse é uma forma efetiva e apropriada para testes envolvendo novas tecnologias. Material gerado por computador pode ser mais difícil de ser codificado que o material de vídeo natural mas, dependendo da aplicação, este tipo de material também deve ser considerado.

Nos testes padronizados envolvendo codificadores são escolhidas seqüências que sejam críticas no contexto de redução de taxa de bits em televisão [01]. O problema consiste, portanto em determinar quão crítica é uma particular seqüência, ou ainda se uma seqüência é ou não mais crítica do que outra. Algumas seqüências onde aparecem movimentos complexos ou envolvendo texturas são críticas para todos os tipos de codecs. Os movimentos ditos complexos são aqueles

que previsíveis para um observador, não podem ser compensados pelos algoritmos de estimação de movimento, como por exemplo, “zoom” e “pan”.

4.4.1 Critério de ‘criticalidade’ de cenas

Uma vez que cenas ricas em detalhes são difíceis de serem codificadas, estes elementos são utilizados para a determinação da dificuldade de se comprimir um sinal. A recomendação ITU-R BT. 500 [02] fornece diretrizes para a escolha das cenas, baseada no conteúdo de falha da imagem. Esta ferramenta, entretanto, não está bem adaptada para tratar codecs de alto desempenho e alta flexibilidade, o que motivou o estudo de outras ferramentas.

A relação entre taxa de bits e qualidade de imagem é o maior ponto de interesse em serviços de transmissão digital, em especial para radiodifusão. Para avaliar o desempenho da codificação do padrão MPEG-2, é necessário estabelecer comparações entre os diversos atributos e níveis disponíveis no padrão, bem como fazer comparações com o sistema NTSC. A idéia é fazer uma análise estatística da programação com base em um critério objetivo, casado com a aplicação, para selecionar o material para o teste subjetivo.

Esta seleção está baseada na noção de ‘criticalidade’, uma medida quantitativa da dificuldade encontrada na codificação digital. Esta medida permite também, que se faça uma análise da estatística de distribuição de probabilidade das degradações para larga gama de amostras da programação de TV, dado que o material que efetivamente pode ser explorado em um experimento de avaliação subjetiva é limitado. Embora forneça indícios sobre a quantidade de degradação presente nas cenas, esta medida não substitui a avaliação subjetiva. Somente a avaliação subjetiva permite a associação entre a qualidade percebida e as degradações da imagem. A medida de criticalidade fornece um critério para se analisar o restante da programação com base nos resultados subjetivos obtidos.

Os diversos resultados subjetivos na literatura mostram que a qualidade de imagem para uma determinada taxa não é constante para cada seqüência. Isto significa que os resultados de avaliações subjetivas com um pequeno número de seqüências varia dependendo do conteúdo das seqüências utilizadas no teste. A aplicação de codificação digital em serviços de radiodifusão, que inclui material bastante variado, requer o conhecimento da proporção do material que será crítica na codificação e a perda de qualidade esperada em tais casos. Observe que esta medida deve estar associada ao comprimento da seqüência e deve ser capaz de representar, também, seqüências longas.

Uma possível definição para criticalidade [50] é o número de bits por pixel de um codificador que emprega predição inter-campos exclusivamente, com blocos transformados pela DCT, quantização linear fixa e estimação de movimento. Utiliza-se uma estratégia de codificação intra-quadro por macrobloco a cada 0,5 segundos. Esta definição fornece a taxa necessária para obter uma qualidade de imagem praticamente constante para cada seqüência. No experimento realizado no “NHK Science and Technical Research Laboratories”, a análise estatística de material da programação normal de TV, revelou uma distribuição com baixa

probabilidade de ocorrência de seqüências de alta criticalidade.

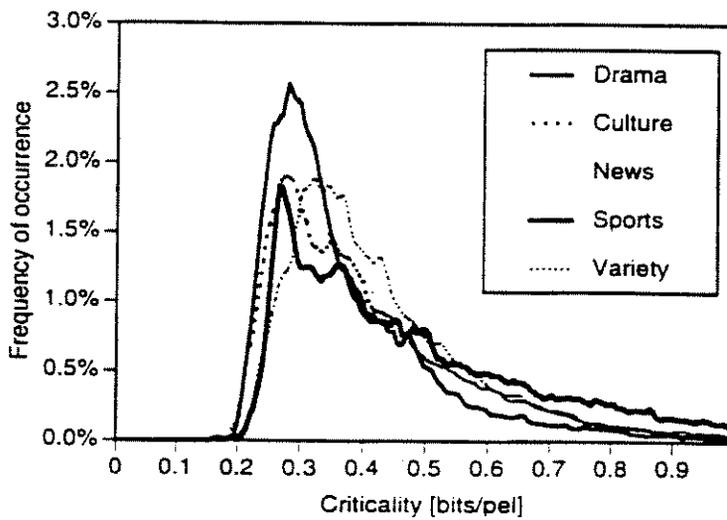


Figura 7: Distribuição de criticalidade em programas de TV (Q=6)

Definida desta forma, a criticalidade reflete as características da imagem propriamente dita, sem depender de particularidades do esquema de codificação como o controle de taxa ou estratégia de predição. Seqüências críticas requerem uma taxa de bits maior do que seqüências não críticas para manter o mesmo nível de qualidade. Numa estratégia em que a taxa de saída é fixa são empregados quantizadores menos refinados em seqüências críticas para baixar a taxa de saída e, conseqüentemente, a degradação em decorrência da quantização geralmente aumenta.

4.4.2 Resultados de simulação de criticalidade

Utilizando esta definição de criticalidade foram feitas simulações para estabelecer a dependência desta medida em relação ao parâmetro Q que determina a granularidade da quantização.

Seqüências	Características	Movimento
Mobile & Calendar	movimento aleatório de objetos	lento
Flower Garden	detalhes coloridos	"pan" lento
Table Tennis	movimentos rápidos múltiplos	"pan", "zoom" e corte de cena
Kielharbour	alta resolução nas dimensões horizontal, vertical e transversal	"pan" lento e "zoom"
Diva with noise	mudanças rápidas de entropia	lento
Popple	cores diversas em movimento	"pan" e rotação

Tabela 1: Cenas utilizadas na simulação de criticalidade - CLASSE 1

As simulações utilizaram dois tipos de cenas. As cinco primeiras cenas selecionadas

avaliação de codecs. Estas cenas estão descritas na Tabela 1.

A segunda classe de cenas selecionadas é composta por trechos da programação normal de televisão. Foram selecionados trechos da programação usual envolvendo esportes, novelas, trechos de documentários e jornalismo.

Categoria	Seqüências	Características
Novela	Cacau (trecho de "Renascer")	cenas envolvendo grandes regiões de textura e com "pan" lento
Novela	Minissérie	expressões faciais e trecho com movimentos rápidos
Esportes	Futebol1 (Corinthians x Palmeiras)	alternância entre cenas da torcida e "zomm" rápido
Esportes	Futebol2 (Botafogo x Vasco)	"pan" rápido e movimentação dos jogadores
Esportes	Moto1 (Campeonato Mundial de Motovelocidade 500cc)	objetos coloridos em movimento, movimentos múltiplos e rápidos
Esportes	Moto2 (Campeonato Mundial de Motovelocidade 250cc)	objetos coloridos em movimento, movimentos múltiplos e rápidos
Documentário	Vistas Aéreas	alta resolução vertical, horizontal e diagonal, "pan" lento
Documentário	Carnaval	detalhes coloridos em movimento, movimento aleatório de objetos
Documentário	Hudson (Reportagem sobre a bacia de Hudson)	características de iluminação variadas e corte de cena
Jornalismo	Jornal	Tons de pele e expressões faciais

Tabela 2: Cenas utilizadas na simulação de criticalidade - CLASSE 2

Para simulação foi utilizado o protótipo MPEG-2 um "software" desenvolvido com base no codec TM5 ("Test Model 5") [14] e adaptado para operar com taxa de saída variável e níveis de quantização fixo. A configuração adotada em cada simulação está detalhada na Tabela 3.

Para cada cena foram selecionados cinco amostras. As amostras são trechos compostos por 12 quadros que corresponde ao tamanho do GOP¹. A seleção de 12 quadros, e não de apenas um quadro aumenta a confiabilidade da medida, sem aumento significativo de sua complexidade computacional. Por exemplo, a seleção aleatória de um único quadro tipo B, geraria resultados de criticalidade inferiores a efetiva dificuldade de codificação da cena.

Característica	Protótipo em Software				

Codificação	intra-quadro, progressiva e bidirecional, com predição bidirecional inter-campos e inter-quadros progressiva e regressiva				
GOP ¹	12	12	12	12	12
M ²	2	2	2	2	2
Formato de codificação de croma	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:0
Quantização	fixa; Q=4	fixa; Q=8	fixa; Q=16	fixa; Q=32	fixa; Q=64
Taxa (média sobre todo o conjunto de cenas utilizadas)	variável 26,2 Mb/s	variável 13,1 Mb/s	variável 6,5 Mb/s	variável 3,3 Mb/s	variável 1,9 Mb/s
Taxa (sobre o conjunto de cenas padronizadas - CLASSE 1)	variável 29,2 Mb/s	variável 14,7 Mb/s	variável 7,0 Mb/s	variável 3,4 Mb/s	variável 1,8 Mb/s

Tabela 3: Especificação dos parâmetros de codificação para medidas de criticalidade

Os cinco níveis de quantização escolhidos estão distribuídos de forma exponencial, dentre os níveis de quantização estipulados no padrão MPEG-2. Observe que as duas últimas linhas da Tabela 3 discriminam a taxa média em Mb/s resultante da compressão com quantização fixa nos níveis escolhidos. As cenas, que compõem o banco de cenas padronizadas e que foram utilizadas para avaliação subjetiva, são em média 9,3% superior à média das demais cenas, também consideradas críticas para cada tipo de programa.

A Figura 8 mostra o resultado das simulações evidenciando as diferenças entre as cenas. Os resultados são apresentados em número médio de bits por pixel da imagem. A coerência dos resultados com relação da granularidade da quantização adotada é acentuada, e as curvas que representam os diversos níveis de quantização não se interceptam. Observe, entretanto, que a adoção de uma quantização muito forçada ($Q > 16$), não é interessante pois as diferenças entre as cenas são muito atenuadas ou desaparecem. Embora a simulação com quantização bastante forçada parecesse, em princípio, interessante para aplicações em baixas taxas, o gráfico evidencia que os resultados obtidos com valores de Q baixos (quantização bem refinada) se aplicam a qualquer taxa, sem perda de generalidade e com maior refinamento das medidas.

¹ GOP ou 'Group of Pictures' indica o número de quadros entre dois quadros consecutivos com codificação intra-quadro.

² O parâmetro M indica a distância entre quadros com codificação preditiva.

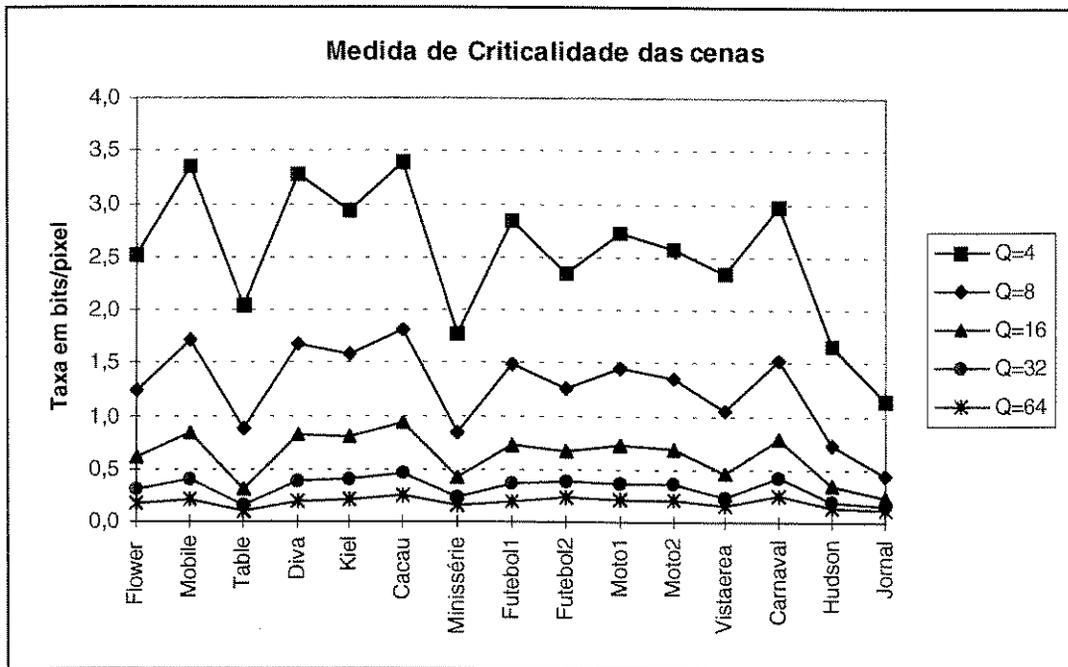


Figura 8: Medida de criticalidade - Variação do parâmetro de quantização

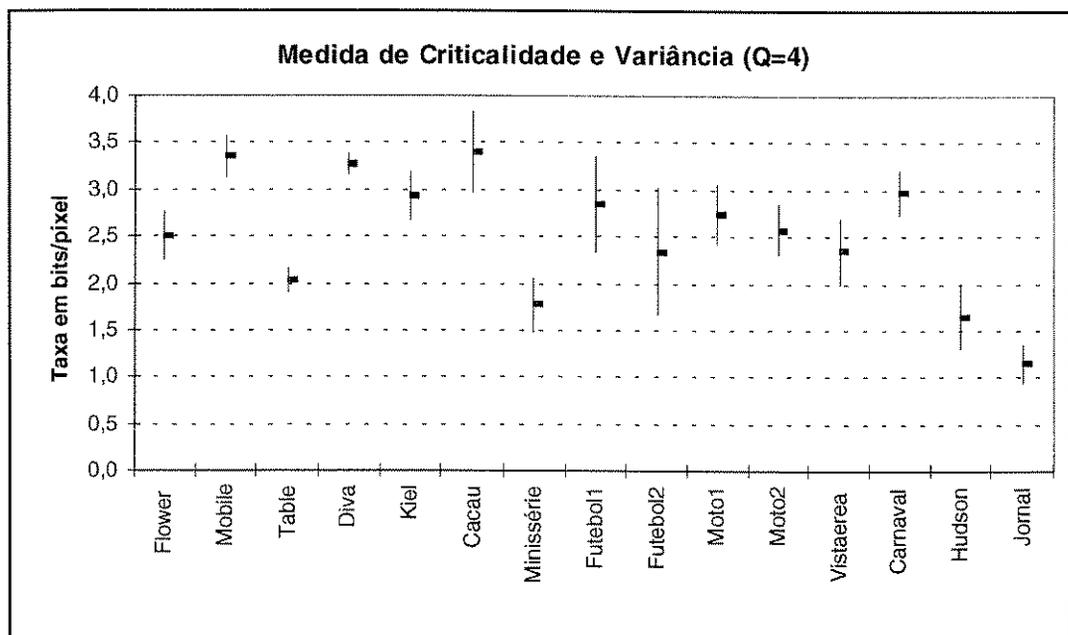


Figura 9: Medidas de criticalidade - Variância da Programação

Embora os resultados mostrados na Figura 9 sejam uma tentativa de analisar a variância das cenas, os resultados devem ser analisados cautelosamente dado que a amostra utilizada é pequena. Uma análise significativa da distribuição de criticalidade das sequências só é possível através da criação de um banco de dados que seja percentualmente representativo da programação, ou do tipo de serviço que se deseja caracterizar.

4.5 Forma de apresentação das seqüências

Em relação a forma de apresentação das cenas, dois aspectos são fundamentais: a utilização ou não de imagens de referência e a repetição de estímulos. A apresentação simultânea de duas cenas, seja uma delas referência ou não, é a forma mais efetiva de avaliar um sistema, promovendo uma comparação mais eficiente entre estímulos. Esta técnica implica, entretanto, na disponibilidade de um sistema de gerência que garanta a correta sincronização entre as cenas e sem o qual a apresentação simultânea tende a deteriorar os resultados da avaliação.

4.5.1 Utilização de cenas de referência

A utilização de uma imagem de referência tem fortes implicações em todo o projeto do experimento pois aumenta o tempo necessário para exibição das unidades de apresentação e, conseqüentemente, diminui o número de cenas envolvidas e o número de situações de degradação que podem ser exploradas. Entretanto, quando disponível, a utilização de imagens de referência pode ser preferível devido aos efeitos de adaptação do observador. A adaptação consiste na variação da magnitude das notas dos avaliadores em função do contexto das demais cenas em teste.

Nos testes de opinião utilizados para avaliar o efeito subjetivo de vários tipos de ruídos e distorções que podem degradar a transmissão de sinais de televisão, a opinião do observador com relação a primeira imagem é baseada em experiências passadas. A medida que o teste progride, entretanto, seu padrão de julgamento é modificado por uma adaptação para a qualidade das imagens que ele já viu. Desta forma, os observadores se tornam mais críticos quando a qualidade da imagem é elevada e vice-versa no caso de imagens de baixa qualidade.

A realização de experimentos, utilizando imagens com grande e com pequena variação de qualidade, revelou grande variação nas notas dadas a mesma imagem em um e em outro teste. Resultados de experimentos apresentaram um deslocamento do teste com estreita faixa de variação de qualidade para o experimento com ampla variação de qualidade quase tão grande quanto o range de notas utilizado no experimento envolvendo pequenas diferenças de qualidade. Este resultado evidencia uma aparente falta de significado dos rótulos verbais utilizados, pois uma imagem, julgada boa passou a ser considerada pobre variando-se o contexto do teste.

Um outro aspecto importante identificado no estudo de Narita [51] é que as notas dadas às imagens de referência são afetada pelo nível de qualidade médio das imagens codificadas. As diferenças nas notas dadas às imagens de referência são particularmente significativas na comparação entre altas e baixas taxas.

Torna-se necessário balancear o projeto do experimento de forma a minimizar a adaptação de curto prazo e de longo prazo. O efeito de adaptação de curto prazo afeta os resultados, por exemplo, em situações em que uma imagem particular segue várias imagens com baixos níveis de degradação. Uma situação que pode, eventualmente, ocorrer em testes onde a ordem de apresentação das cenas é totalmente aleatória. Mesmo compensando estes efeitos de

curto prazo, a adaptação de longo prazo ainda pode ocorrer, pois dependem das características gerais das imagens apresentadas.

Fica claro, portanto, que a característica que relaciona a nota a magnitude objetiva de determinada degradação não é única, mas depende da faixa e do nível médio de degradação apresentados no teste [50]. Algumas tentativas foram feitas no sentido de padronizar estas características. Uma condição padrão deve estar baseada em medidas subjetivas para ser efetiva, mas, é somente após a realização do teste que estas medidas subjetivas estão disponíveis. Por isso, estes resultados de padronização têm aplicação limitada.

4.5.1.1 Descrição e quantificação do efeito de adaptação

Uma vez que a não utilização de referência é preferível por diminuir o tempo de duração do teste, algumas técnicas são propostas para se compensar a adaptação dos observadores. Segundo Corbett [27], a magnitude do efeito de adaptação para um tipo particular de degradação é dependente de um coeficiente de degradação. O estudo de Corbett procura determinar a tolerância e propor uma padronização dos resultados.

O que define o condicionamento do teste é a medida subjetiva no nível de degradação médio. A distribuição das opiniões para um dado valor médio de degradação não é, relativamente, muito afetada pela situação de condicionamento. Assim, as medidas que melhor descrevem o condicionamento em termos subjetivos são a média geral do teste, ou seja, a média das notas médias para cada situação de degradação e o desvio padrão das médias em relação a média geral. No que diz respeito à adaptação, portanto, a obtenção de um valor médio geral de 0,5, em um intervalo de 0 a 1, é sempre desejada.

Embora as distorções causadas pela adaptação, em testes bem projetados, sejam relativamente pequenas, é desejado o desenvolvimento de metodologias para diminuir a susceptibilidade dos avaliadores a este efeito. Uma estratégia utilizada é o uso de estímulos-âncoras, condições de teste padrão com qualidade acima e abaixo do range em teste e apresentadas em pontos pré-determinados durante o teste. O uso deste tipo de estímulo, entretanto, tem a grande desvantagem de envolver condições muito diferentes das imagens efetivamente testadas. É possível, por outro lado, treinar os observadores para que se tornem mais imunes à adaptação. O uso de observadores treinados, entretanto, leva a resultados atípicos e difíceis de serem interpretados.

Ainda que métodos com resultados satisfatórios no que diz respeito à adaptação sejam utilizados, informações a respeito do efeito de adaptação são necessárias em aplicações rigorosas dos resultados dos testes às situações de condicionamento que podem surgir na observação da programação em casa. Observadores que estão perto dos transmissores e, em geral, recebem imagens de boa qualidade, têm padrões de qualidade mais elevados do que aqueles que estão mais distantes, acostumados com imagens pobres.

4.5.2 Técnicas de comparação

Este método é aplicado a ordenação de seqüências de acordo com sua qualidade subjetiva. A utilização de comparação tem a vantagem de que a forma dos dados obtidos facilita a análise de transitividade nos julgamentos e análise da concordância de critérios usados por diferentes observadores. Não é sempre óbvio que as condições de transitividade são atingidas, principalmente em situações em que degradações complexas da imagem estão envolvidas.

Em contrapartida à possibilidade de discriminação e ordenação de efeitos subjetivos muito semelhantes, tem-se a impossibilidade de cálculo preciso das distâncias entre os estímulos. Em termos subjetivos há sempre um compromisso no que diz respeito à obtenção de resultados relativos à ordem e às distâncias entre os estímulos.

A confiabilidade do método é extremamente depende do número de observadores e da diversidade de seqüências utilizadas. A Recomendação ITU-R BT. 500 [02] estabelece como número mínimo seis cenas. A necessidade de exploração exaustiva de comparações entre os estímulos determina um crescimento exponencial do tamanho do teste com o aumento do número de situações. As alternativas encontradas neste caso são dividir as observações entre os avaliadores ou fazer uso de uma amostra dos possíveis pares. A seleção de uma amostra apropriada para o teste em questão requer conhecimento e sensibilidade dos efeitos que se deseja testar no nível de percepção e, muitas vezes, pressupõe a realização de testes pilotos.

4.5.2.1 Ferramenta de análise de transitividade: método Zeta

Um ciclo de preferências é um caso tipo de inconsistência presente em testes por comparação de estímulos. Os ciclos aparecem devido à falta de habilidade do observador ou às diminutas diferenças entre as imagens avaliadas. Entretanto, ciclos também podem acontecer se as diversas imagens utilizadas induzem o avaliador a prestar atenção a diferentes artefatos.

Suponha que em determinado teste foram utilizadas n cenas, N avaliadores e que, em ordem aleatória, foram apresentados todos os possíveis pares. O número total de pares é:

$$z = \frac{n(n-1)}{2} \quad (4.1)$$

Os resultados de cada indivíduo devem ser combinados em uma matriz de dominância formada por “zeros” e “uns”. O número “1” na coluna t e linha s indica que na opinião do avaliador em questão, a imagem da seqüência t é melhor que a imagem s . Um “0” na mesma posição da matriz significa o contrário. Uma vez organizados todos os dados na forma matricial, deve-se checar a “transitividade” dos resultados.

No caso da avaliação de três sistemas, a tríade é dita “intransitiva” se o avaliador decide que a imagem A é melhor do que B, a B é melhor que a C, porém a C é melhor do que a A. Esta observação deixa claro que só é possível fazer uma ordenação dos sistemas e situações em teste, se o julgamento dos observadores for feito de forma sistematicamente transitiva.

O número de tríades intransitivas na matriz de dominância é dado por:

$$d = \left(\frac{n(n-1)(2n-1)}{12} - \frac{1}{2} \right) \sum_i D_i^2 \quad (4.2)$$

onde D_i é a soma da i -ésima coluna. O número máximo de tríades intransitivas vale:

$$d_{\max} = \begin{cases} \frac{n(n^2-4)}{24}, & n \text{ par} \\ \frac{n(n^2-1)}{24}, & n \text{ ímpar} \end{cases} \quad (4.3)$$

A relação entre o número máximo possível de tríades intransitivas e o número efetivamente calculado fornece uma medida de transitividade.

$$\zeta = 1 - \frac{d}{d_{\max}} \quad (4.4)$$

Se ζ igual a um, os resultados são absolutamente transitivos e, contrariamente, se ζ vale zero, os resultados são absolutamente intransitivos. Em casos práticos, a transitividade é usualmente aleatória. Nestes casos, o “método Zeta” define um critério para que se estabeleça se a transitividade é ou não sistemática. Para tal calcula-se a função:

$$x(d) = \frac{8}{n-4} \left(\frac{1}{4} \binom{n}{d} - d + \frac{1}{2} \right) + \frac{n(n-1)(n-2)}{(n-4)^2} \quad (4.5)$$

A segunda parcela corresponde aos graus de liberdade do experimento. Para um número de cenas superior a seis (e daí, o número mínimo de cenas sugerido na Recomendação) a função $x(d)$ é uma qui-quadrada (χ^2). A distribuição χ^2 é bem conhecida e seus valores tabelados [01-pag 29] de acordo com os graus de liberdade do sistema e com o percentual aceitável de intransitividade, que está diretamente relacionado à confiabilidade dos resultados subjetivos. Assume-se que a transitividade é sistemática se o valor de x calculado for superior aos valores tabelados.

Resta saber ainda, se há concordância entre avaliadores, ou seja, concordância nos critérios de avaliação utilizados. Para tal, os resultados são combinados em uma matriz agregada, onde as linhas representam cada um dos pares testados no experimento, em ordem aleatória, e as colunas representam os avaliadores. Os elementos x_{ij} da matriz valem 1(0) se na opinião do avaliador j a primeira imagem é melhor(pior) do que a segunda.

Novamente, a concordância entre os observadores pode ser sistemática ou aleatória. Para tal, calcula-se:

$$Q = \frac{\binom{n}{2} \left[\binom{n}{2} - 1 \right] \sum_i (L_i - \bar{L})^2}{\binom{n}{2} \sum_j C_j - \sum C_j^2} \quad (4.6)$$

onde L_i é a soma da i -ésima linha da matriz agregada e C_j a soma da j -ésima coluna. Para esta análise o número de graus de liberdade é dado por $\binom{n}{2} - 1$. Analogamente à análise de transitividade, fixado um percentual de confiabilidade e calculado número de graus de liberdade, a decisão se baseia nos valores contidos na tabela da referência [01]. Se o valor calculado para Q

for maior do que o valor da tabelado a concordância entre avaliadores é sistemática.

4.5.3 Influência do número de apresentações

Especialmente nas situações em que a imagem de referência não está presente, pode-se apresentar as cenas a serem avaliadas mais de uma vez. O objetivo da repetição das apresentações é facilitar a atividade dos avaliadores, melhorando a capacidade de percepção de detalhes. Estas metodologias de repetição de estímulos envolvem, geralmente, a repetição do estímulo por duas vezes. Mais apresentações fornecem pouca ou nenhuma informação adicional segundo [45].

Este recurso é utilizado quando há necessidade de discriminação de cenas muito semelhantes, seja porque o range de qualidade é estreito ou porque as cenas a serem comparadas são muito parecidas. Embora nestas situações, a aplicação desta técnica produza resultados significativos, a replicação de estímulos deve respeitar a limitação de tempo das sessões de avaliação, ou seja, não deve ser utilizada nos casos em que a tempo de exposição dos avaliadores é crítico.

4.5.4 Duração dos trechos de avaliação

No caso de seqüências de vídeo dinâmicas, estabeleceu-se a utilização de trechos vídeo de 10 a 15 segundos de duração. Para seqüências de vídeo estáticas o tempo de exibição pode ser mais curto, 3 a 4 segundos. Neste esquema de apresentação sugerido pela Recomendação ITU-R apenas uma nota é dada para cada estímulo, mas, especialmente no caso de trechos de vídeo com movimento, faz-se necessário alguma forma de ponderar esta impressão de qualidade ao longo do tempo.

A limitação imposta por seqüências de curta duração se torna um problema na avaliação de novos sistemas. As seqüências, em geral, envolvem grande variação de qualidade e não são distribuídas ao longo do tempo. Não é raro que algumas degradações severas somente ocorram uma vez a cada 10 ou 20 minutos e no restante do tempo a qualidade seja boa. Por outro lado, um trecho pré-selecionado, onde a qualidade da imagem é particularmente ruim, também não representa a qualidade do sistema de forma geral.

A limitação do tempo de exibição a seqüências de teste relativamente curtas se torna um problema quando se está interessado em avaliar sistemas digitais avançados como MPEG-1 ou MPEG-2. Estes sistemas apresentam variações significativas que não estão distribuídas ao longo do tempo. Seqüências longas são necessárias para apreciação da influência destas variações, que podem ocorrer apenas uma vez em períodos de alguns minutos. Particularmente na codificação a taxa constante, e baixa o suficiente para causar distorção, a qualidade em trechos de 10 segundos sucessivos pode ser muito diferente. Nesta situação, em cenas com movimentos acentuados, pode ocorrer o transbordo do "buffer" e a introdução de diversos níveis de

distorção no vídeo codificado durante o tempo.

Reconhecidamente, as metodologias de avaliação subjetiva padronizadas pelo ITU-R não são capazes de lidar facilmente com variações na qualidade da imagem. Este problema motivou o desenvolvimento de novas metodologias direcionadas para fazer aquisição de dados no tempo. Estas metodologias estão em discussão no grupo de estudo 12 do ITU, em vias de ser incorporado às novas versões da recomendação ITU-R BT. 500 e serão discutidas em maior profundidade na secção 4.8.3.

A avaliação de seqüências de longa duração trás algumas dificuldades para o projetista. O primeiro problema é que para seqüências muito longas, métodos de estímulo duplo são dificilmente aplicáveis, pois o tempo entre a exibição seqüencial de duas versões de uma seqüência se torna muito longo. O segundo problema está relacionado ao efeito de esquecimento nos julgamentos subjetivos, ou seja, as partes mais recentes da seqüência têm relativamente, maior contribuição para a impressão de qualidade.

4.5.4.1 Capacidade de avaliação contínua

Os primeiros estudos neste campo foram feitos por Hamberg et al.[34], onde comprovou-se a possibilidade de avaliação contínua da qualidade de uma seqüência de vídeo de forma consistente e confiável. Para tal, estes pesquisadores fizeram um estudo comparativo utilizando seqüências de longa duração e fragmentos de imagem retirados desta longa seqüência. O trabalho destes dois pesquisadores se concentrou no método de medida. Eles utilizaram imagens paradas com níveis de “blurring” variável. Dois tipos de materiais são utilizados: no primeiro, o nível de “blurring” é mantido constante em cada cena para se quantificar seu efeito subjetivo e no segundo, o nível do parâmetro é variado ao longo do tempo com o objetivo de verificar as características temporais do escalonamento contínuo.

Há interesse, portanto, em verificar especificamente, como o comportamento temporal dos avaliadores se reflete nos dados. É intuitivo que haja um atraso na percepção e um atraso relacionado a atividade motora para registro das variações. Uma solução possível para compensar estes dois efeitos é a utilização de um filtro linear causal, que minimize a diferença entre os dados subjetivos e as medidas objetivas da seqüência. O filtro resultante não é uma função simples e bem conhecida, dada às diferenças de natureza dos efeitos de atraso que se procura explicar. Além disso, seu formato é influenciado pela frequência de amostragem das notas subjetivas. É esperado, entretanto, que um filtro real tenha componente nula para $\Delta t=0$ e para valores maiores, os coeficientes estejam distribuídos ao longo do tempo com máximo próximo a $\Delta t=0,5$ segundos, tempo previsto para o atraso motor. Este resultado mostra que a qualidade de um imagem pode ser medida instantaneamente, com pequena mediação temporal. Para a descrição do sistema, entretanto, é preciso fazer um ajuste dos dados e projetar um filtro adequado.

4.5.4.2 Caracterização do efeito de esquecimento

Observou-se que os avaliadores, quando solicitados a fornecer uma única nota após assistirem a uma sessão de vídeo degradado são fortemente influenciados pela última porção da seqüência. Se esta sessão tem boa qualidade, ela tende a elevar a nota média, mas se ao contrário tem baixa qualidade a tendência é causar um decréscimo da média. Tudo se passa como se o avaliador fizesse uma média temporal tal que o máximo peso é dado às seqüências exibidas mais recentemente.

Este efeito na formação de opinião dos avaliadores recebe o nome de efeito de esquecimento e está fortemente relacionado ao tamanho da “*memória de trabalho humana*”. Aldridge et al. [17] realizaram estudos para descobrir como os avaliadores formam suas opiniões relativamente à qualidade da imagem e como são influenciados pela distribuição de qualidade ao longo do tempo. Estes estudos envolveram seqüências de 30 segundos de duração nas quais é possível observar variações consideráveis de qualidade e onde vídeo de boa qualidade é entremeado por picos de degradação visível. As seqüências foram selecionadas de forma a apresentarem trechos de 10 segundos durante os quais as degradações são bastante visíveis. Quando o trecho degradado (10s) se localiza no final do trecho codificado (30s), as notas obtidas se assemelham a apresentação apenas do trecho degradado. A diferença obtida nos julgamentos pelo posicionamento do trecho de maior degradação no começo ou no final da seqüência, neste estudo [17], foram de aproximadamente 10%.

Verifica-se que, os observadores tratam distorções com variações temporais e variações espaciais de forma diferente. No caso de distorções que variam espacialmente, o avaliador tende a fazer uma avaliação pela região de pior qualidade [45] enquanto que, variações temporais têm efeito diferente dependendo da região da seqüência em que ocorrem. De fato, quando um trecho de vídeo de boa qualidade se sucede a um trecho bastante degradado, os observadores tendem a esquecer o trecho ruim e as notas médias são maiores e têm baixa correlação com as notas dadas ao trecho degradado isoladamente. Da mesma forma, se o último trecho de vídeo apresentado tem baixa qualidade o observador tende a julgar toda a seqüência com base nesta última porção, aparentemente dando baixa relevância ao material apresentado antes.

4.6 Escalas

Principalmente em função, dos elevados custos das avaliações subjetivas, grande empenho é dedicado ao desenvolvimento de metodologias que permitam a obtenção de resultados de máxima confiabilidade e mínimo esforço. Um aspecto importante na otimização desta relação é o casamento entre o problema, os efeitos de degradação da imagem e o tipo de escala utilizada. A escala não deve causar dúvida ao avaliador e deve ter sensibilidade suficiente para se ajustar a todos os critérios utilizados pelo avaliador.

4.6.1 Caracterização das escalas

4.6.1.1 Escala de categorias

As escalas de categorias são o tipo de escala mais comumente utilizada em avaliação subjetiva e o primeiro tipo a ser estudado. O termo escala de categorias é usado para definir escalas onde a graduação dos níveis é definida por termos da língua portuguesa.

São de dois tipos as categorias usadas em avaliação subjetiva de imagens, as categorias relacionadas à qualidade da imagem e as categorias que buscam definir o nível de degradação da imagem. A Recomendação ITU-R BT. 500 [02] estabelece dois tipos de escala com cinco níveis principais, conforme ilustrado na Tabela 4.

Nível	Categorias de Qualidade	Categorias de Degradação
5	Excelente	Imperceptível
4	Bom	Perceptível, mas não incomoda
3	Regular	Perceptível e incomoda pouco
2	Pobre	Incomoda
1	Ruim	Incomoda muito

Tabela 4: Categorias para descrição da imagem para o português falado no Brasil

4.6.1.2 Escala numérica linear

As escalas numéricas exploram a habilidade do observador de emitir julgamentos em uma escala linear. Uma vez que os diversos níveis da escala não estão limitadas pelos valores dos adjetivos, a escala poder ser usada em diferentes faixas de degradação. É uma escala mais flexível, que permite a variação do número de pontos da escala.

Quando largo “range” de qualidade perceptível é empregado, os resultados obtidos com a utilização de escalas numéricas lineares é consistente com os resultados obtidos por escalas não métricas. Embora o estudo de Hidder e Majoor [56] se restrinja a cenas paradas e envolva poucos observadores, os resultados apresentados são indicativos de que estas escalas podem ser utilizadas com sucesso na avaliação de efeitos usualmente encontrados em imagens digitalmente codificadas com poucos artefatos perceptíveis.

Apesar de amplamente utilizadas, a eficiência da avaliação por categoria nos casos em que a visibilidade das degradações é pequena, é questionável. Todas as opiniões para sistemas de alta qualidade estão na categoria excelente, assim, embora existam diferenças de qualidade, não será possível distinguir entre os resultados. Uma alternativa é aumentar a flexibilidade da escala por categorias pela utilização de números ao invés de adjetivos. A escala numérica pode, em princípio, se adaptar a qualquer faixa de qualidade, sem os problemas referentes a adequação dos adjetivos.

A desvantagem em se utilizar uma escala de qualidade contínua sem os indicadores de

qualidade, é que este não é capaz de associar o significado com a posição na escala. Enquanto um observador deseja indicar que uma imagem está *ruim* o outro deseja indicar que se trata de uma imagem apenas *regular*. Por esta razão, a utilização de números na descrição de cenas em contexto de baixa qualidade é crítica.

4.6.1.3 Escala racional

A idéia da utilização de escalas baseadas em proporções na avaliação de imagens foi aventada após um estudo dos processos sensoriais dirigido à medição de insatisfação política. Este tipo de escala fornece informações que outras não são capazes, como, por exemplo, a distância entre as variáveis. Tem também a vantagem de ser ilimitada, pois é característica da avaliação humana que os avaliadores evitam os extremos da escala, reduzindo a informação que pode ser obtida. Seu estudo é recente, e sua utilização ainda não está totalmente padronizada.

Da mesma forma que as escalas numéricas, a motivação de aplicação deste tipo de escala na avaliação de imagens é a emergência de sinais de televisão de alta qualidade onde o range de qualidade das imagens é estreito, e o efeito de adaptação em função do contexto, maior. A construção da escala racional permite que o avaliador do teste classifique os elementos a serem analisados e também descreva a magnitude relativa entre os descritores. Ou seja, pode-se dizer não apenas que uma dada imagem A é melhor do que a imagem B, mas também quão melhor é esta imagem. A utilização da escala por categoria permite a ordenação, mas não a determinação dos intervalos, quando na realidade, a estimação de magnitude é o principal objetivo do teste.

O observador é apresentado a uma série de imagens em ordem aleatória, e é convidado a fornecer um valor numérico para a qualidade de cada uma das seqüências. A escala construída tem, portanto, a vantagem, de prover uma “razão” entre intervalos indicativos de quão melhor é uma imagem em relação a outra. O organizador da avaliação deve garantir que as variações de qualidade envolvidas e o número de estímulos sejam grandes o suficiente para permitir que o observador utilize todos os critérios de julgamento disponíveis.

Justamente pela forma como é construída, a escala racional dá origem a intervalos igualmente espaçados, isto aumenta a precisão e permite uma descrição mais completa dos dados. Uma vez que os observadores geram suas próprias escalas, contorna-se uma importante falha das escalas fixas, as quais impõem rótulos e/ou números potencialmente não apropriados ou pouco confortáveis. A escala racional, ao contrário, é construída pelo observador ao longo do experimento.

De forma a estabelecer uma referência e permitir a comparação entre resultados realizados em diferentes laboratórios e com diferentes sistemas de televisão, os observadores são convidados, ao final de cada sessão a fornecer um valor numérico que seja apropriado a sua concepção de qualidade de imagem “ideal”. O “ideal” deve estar associado a melhor qualidade de imagem imaginável. Na análise dos dados o julgamento do “ideal” deverá ser normalizado para fornecer o número 100, e este valor irá fornecer um padrão uniforme. É uma limitação do método, entretanto, que a estabilidade somente é atingida quando mais de quinze observadores são envolvidos. Melhores resultados são obtidos se a primeira cena apresentada tiver um nível de

qualidade no meio do range e não em um dos extremos.

Para a análise dos dados, a medida mais comum e mais indicada da tendência central da estimação de magnitude é a média geométrica dos dados, pois leva em consideração a distribuição das respostas. A média geométrica é uma estimativa não polarizada do valor esperado do logaritmo da magnitude. Para o cálculo da média não é necessário normalizar os dados basta que as médias sejam multiplicadas por $100/R_1$ onde R_1 é o valor numérico associado a resposta “ideal”.

4.6.1.4 Escala gráfica

As escalas gráficas não identificam por si só uma classe de métodos de avaliação subjetiva, mas caracterizam a forma de apresentação e utilização das escalas. Na adaptação lingüística das escalas de categorias, as escalas gráficas passaram a ser importante ferramenta de estudos sobre as diferenças semânticas entre as diversas línguas dos rótulos escolhidos para representar cada categoria.

A motivação deste estudo são as diferenças entre os resultados obtidos por escalas contínuas e escalas de categorias. As escalas gráficas são usadas para determinar o intervalo percebido entre os descritores. Adjetivos e advérbios são usados para determinar sua força relativa como modificadores.

Os resultados dos testes de escalonamento gráfico são inerentemente válidos [39], pois os avaliadores emitem seu julgamento na sua língua nativa, na ausência de restrições, e sem a necessidade de interpretação numérica.

4.6.1.5 Escala multidimensional

A comparação de estímulos em imagens de televisão deu origem a métodos multidimensionais. Neste caso deseja-se explorar a similaridade entre imagens com diferentes níveis e tipos de degradação. A similaridade entre os julgamentos são uma medida da “distância” entre condições no espaço de percepção de dimensão n . Escalas multidimensionais podem ser usadas para estabelecer as dimensões em que os fatores de projeto e de transmissão variam. As coordenadas utilizadas para medir o nível de cada fator neste espaço estabelecem relações entre as medidas objetivas, os valores percebidos e a relação entre ambos na descrição do nível de satisfação do observador.

Para explorar toda a potencialidade deste método, podem ser utilizadas múltiplas variáveis. Os experimentos nesta área buscam identificar e interpretar grupos de variáveis e fatores subjetivos que afetam significativamente a qualidade de uma imagem. As variáveis mais relevantes segundo documento do ITU [11] são: bordas, luminosidade, detalhes, movimento (especialmente horizontal), cores, ruído, contraste, cintilação, conteúdo global e local, clareza, expressão facial, razão entre a clareza do rosto e o fundo, continuidade das seqüências e posicionamento dos objetos.

4.6.2 Comparações e análise de desempenho

Na seção 4.6.1 foi feita a apresentação de diversos tipos de escala utilizados em avaliação subjetiva. Nesta seção são feitas comparações indicativas das aplicações e limitações de cada tipo de escala.

4.6.2.1 Escala de categorias: qualidade vs degradação

A escolha entre termos associados à qualidade ou ao nível de degradação da imagem pode ter impacto significativo sobre os resultados da avaliação. A diferença está associada também ao efeito de adaptação já descrito, uma vez que os métodos de categorias de degradação, em oposição aos métodos de qualidade padronizados, utilizam imagens de referência. Esta constatação é o resultado de um estudo comparativo entre métodos que utilizam as escalas de qualidade e de degradação, com e sem a presença de uma imagem de referência realizado por Sallio et al. [58].

O primeiro resultado interessante da comparação entre os dois tipos de referência diz respeito à nota associada à imagem de referência. No experimento utilizando o método de degradação, a imagem de referência é sistematicamente apresentada antes de cada uma das imagens avaliadas (Método de Estímulo Duplo [02]) e recebe notas próximas a 5, correspondentes a degradação imperceptível, em todos os casos. Esta imagem de referência funciona como âncora para a escala subjetiva utilizada. Ao contrário, nos casos de avaliação de qualidade foi utilizado método de estímulo único conforme ITUR BT. 500 [02], e nesta situação, a imagem de referência não é sistematicamente associada a imagem de qualidade excelente.

Com boa aproximação, a diferença entre as notas médias dadas às diversas imagens de referência é mantida para os estímulos das cenas correspondentes cujas notas que estão entre 5 e 3 na escala de degradação. Na maioria dos casos, este resultado também vale para altos valores de degradação, notas abaixo de três com boa aproximação. Este resultado fornece um método empírico para a conversão de medidas de degradação em medidas de qualidade. Para tal basta fazer um deslocamento das notas de um valor igual à média associada à degradação residual, ou seja a degradação presente na imagem de referência. Assim subtraindo-se uma constante do valor medido na escala de degradação obtém-se o correspondente valor na escala de qualidade. Este valor pode, entretanto, variar de acordo com o tipo da degradação presente no experimento e com a distância ao monitor.

Este deslocamento nas escalas sugere um eixo de percepção subjacente para o qual as imagens de referência funcionam como âncoras, determinando o range de variação das notas. Na ausência da imagem de referência, as notas flutuam em bloco. Este deslocamento entre notas dadas para qualidade e para degradação são estáveis dentro de um mesmo experimento, devido à adaptação que ocorre durante a sessão, mas são instáveis entre experimentos. As notas para a degradação são fortemente ancoradas na parte superior da escala. As notas de qualidade, na ausência de referência, flutuam em blocos em relação às notas de degradação de um experimento

para outro. Na região de boa qualidade (notas de 4 a 5), o desvio padrão associado às medidas de degradação são significativamente menores do que os desvios referentes a medidas de qualidade. Na área correspondente ao meio da escala, entretanto, os resultados são variáveis, podendo ocorrer desvios para as medidas de degradação maiores ou menores do que os valores medidos na escala de qualidade, de tal forma que, na média, as notas em qualidade e em degradação têm variações equivalentes.

Estes resultados indicam a existência de um contínuo entre a percepção em qualidade e em degradação. As respostas dadas pelos observadores, isto é, a adaptação deste contínuo em termos absolutos, depende do contexto das imagens que são solicitados a julgar. Os principais efeitos perceptíveis são o efeito de adaptação ao range de degradação, influência do contexto e influência de imagens que funcionem como âncoras. Outro resultado que justifica a hipótese de existência deste contínuo é a similaridade da influência do contexto sobre os resultados.

Uma observação significativa na comparação entre os dois métodos é a constatação de que o contexto tem especial relevância na determinação dos resultados na região correspondente a baixa qualidade. Este fenômeno é bastante dependente da natureza das degradações sobre investigação. Uma vez escolhida a degradação, os dois métodos reagem exatamente da mesma forma a mudanças de contexto. Da mesma forma, no que diz respeito a variações na faixa das degradações, os dois métodos são igualmente sensíveis para permitir a seleção do ponto de operação ótimo.

4.6.2.2 Comparação entre escalas racionais e ordinais com relação ao efeito de adaptação

Principalmente no que diz respeito à utilização da tecnologia de alta definição em substituição ao sistema convencional, a informação relevante não é simplesmente saber qual o melhor sistema, mas sim quão melhor. A forma encontrada para obter este tipo de informação foi a utilização de escalas racionais geradas por estimação de magnitude.

Os testes realizados com o método de qualidade contínua mostram que há grande deslocamento das notas quando são comparados experimentos realizados com um range de qualidade estreito em relação a um contexto em que mais níveis de qualidade são testados. Os testes realizados com escala racional apresentam o mesmo tipo de comportamento e de deslocamento, porém em proporções muito menores. No experimento de Jones et al. [40], o deslocamento máximo chega a 34% no caso de escalas racionais, contra 63% para escala ordinal de qualidade. Um aspecto relevante nesse caso é que quando a faixa dos estímulos se estreita a faixa de valores utilizados pelos avaliadores também se estreita, evidenciando sua adequação para este tipo de teste pela fácil adaptabilidade. Os observadores têm possibilidade de reagir pela expansão e compressão das escalas em resposta a variações na faixa de estímulos utilizada, pois não há necessidade de interpretação lingüística.

Outros resultados de Jones et al. [39] detectaram problemas relacionados ao acúmulo de termos na região inferior da escala, chamado “efeito do fim de escala”. Este efeito é outro forte motivo para a adoção de uma escala sem limites ou de uma escala racional.

O mesmo teste mostra que a avaliação numa escala de cinco pontos reforça a influência

do contexto, evidenciando que a interpretação dos rótulos relacionados aos valores médios é problemática. A estimação de magnitude, ao contrário, apresenta menor efeito de contexto e tem a virtude de prover intervalos significativos. Quando o range de estímulos utilizado é o mesmo, os dois métodos são estáveis o suficiente para permitir a comparação entre resultados provenientes de diferentes laboratórios. Entretanto, no caso das faixas utilizados serem diferentes somente o método da razão de escala é estável o suficiente para prover resultados relativos confiáveis entre experimentos.

4.6.2.3 Análise de intervalos nas escalas de categorias por meio de escalas gráficas

A ordenação de imagens pode, em geral, ser feita facilmente, resultando em uma lista ordenada. Muitas vezes, entretanto, essa ordenação não contém nenhuma informação sobre a distância entre as posições relativas. Na construção de uma escala relativa a qualidade da imagem, entretanto, as distâncias podem eventualmente ser a informação mais importante. Em algumas situações pode ser óbvio do que uma imagem é melhor que a outra, entretanto, decidir quão melhor pode ser o fator determinante, por exemplo, para se fazer uma especificação ou tomar decisões.

A natureza e as propriedades dos dados precisam ser conhecidas para que se possa fazer a análise estatística adequada. As medidas de tendência central apropriadas, em se tratando de escalas ordinárias, são: moda e mediana, já que as distribuições não são conhecidas. Média e desvio padrão são apropriadas para tratar intervalos e razões onde a distribuição é conhecida. As escalas de categoria são por natureza ordinárias, porém, podem se tornar uma escala de intervalos se os intervalos forem iguais [20].

A escala por categorias padronizada assume que os descritores são percebidos subjetivamente em intervalos igualmente espaçados. Esta hipótese entretanto, não é verdadeira. Estudos das escalas com ênfase em examinar o significado e o espaçamento entre os descritores, em diversas línguas, evidenciam grandes diferenças nas distâncias percebidas nos intervalos entre os termos nas diferentes línguas. Por razões associadas as diferenças de significado associadas a estes descritores, a correlação de resultados entre laboratórios não é considerada boa o suficiente para que sistemas de alta qualidade sejam avaliados em diferentes laboratórios e o resultado absoluto comparado. De fato, para que a tradução dos termos referentes às categorias de qualidade seja significativa, devem ser escolhidos termos que apresentem baixo desvio padrão na análise da correlação entre o termo e seu valor numérico numa escala familiar aos avaliadores. Uma das implicações de se ter elevado desvio padrão associado às categorias da escala é a baixa confiabilidade da escala de intervalos associada, já que as distâncias subjetivas entre os termos diferem grandemente de um observador para outro.

Em geral, os termos indicativos de qualidade servem como guia e são capazes de discriminar se uma imagem é boa ou ruim. Por sua vez, a concordância em termos subjetivos das escalas de degradação é maior se comparada às escalas de qualidade, levando a um resultado mais estável. Adicionalmente, nos testes realizados pelo NASA Lewis Research Center em Cleveland os termos utilizados na escala de degradação pelo ITU-R, em inglês, fornecem

intervalos que são regulares e iguais. Também para japonês, francês e alemão, a escala de degradação é praticamente uniforme [11].

4.6.3 Análise de intervalos da escala de categorias de qualidade padronizada para o português falado no Brasil

4.6.3.1 Descrição do teste

Uma vez que a escala de categorias utilizada em métodos subjetivos é empregada como escala de intervalos nos procedimentos estatísticos de análise de resultados, há interesse em determinar o intervalo percebido entre termos sucessivos utilizados para descrever a qualidade subjetiva de imagens e sons. Para verificar a validade da suposição de categorias equidistantes no que se refere ao português falado no Brasil, os rótulos padronizados, traduzidos do inglês para o português, foram ordenados através de escalas gráficas.

Preparou-se um formulário no qual os avaliadores são apresentados a doze termos usualmente utilizados na descrição de qualidade de imagem. Todo o conjunto de termos em estudo é conhecido do avaliador, de forma a permitir a correlação entre eles. A tarefa do avaliador consiste em assinalar numa linha vertical não numerada, a posição de cada um dos doze termos selecionados. Os extremos da escala estão assinalados e identificados, respectivamente, como a melhor e a pior qualidade imaginável.

Participaram do teste, 45 pessoas. Para a conversão dos dados da escala gráfica em números, posteriormente a marcação dos termos, a linha vertical foi dividida em 40 intervalos iguais. Isto corresponde a uma quantização dos resultados em intervalos de 0,1 numa escala de 5 (melhor imaginável) a 1 (pior imaginável).

Na análise dos resultados as diferenças em termos de idade e nível de escolaridade são consideradas. O teste não leva em consideração variações nos resultados provenientes de regionalismos. Embora existam representantes de todas as regiões do país, há grande predominância de avaliadores da região sudeste.

4.6.3.2 Apresentação dos resultados

Nos gráficos a seguir são apresentados os valores médios obtidos e os respectivos desvios para cada um dos doze termos em estudo. No Anexo 9.1 são apresentados os valores médios obtidos bem como os resultados médios discriminados por idade.

O desvio padrão obtido fornece indicações da aplicabilidade de cada um dos termos em testes subjetivos. Observa-se que os termos associados à elevada qualidade apresentam menor variabilidade do que os termos associados à baixa qualidade. Selecionando os quatro termos que obtiveram as mais baixas médias e os quatro termos de mais altas médias, o desvio padrão médio associado é 1,6 vezes maior para os termos associados a baixa qualidade. O adjetivo 'pobre' é o termo de maior desvio padrão, um valor que é 3,6 vezes maior do que o menor

desvio, associado ao termo excelente, o que deixa claro sua inadequação na caracterização de qualidade. Apesar disto, é um dos cinco termos selecionados pelo ITU-R. Na Figura 11 são apresentados apenas os termos efetivamente selecionados pela Recomendação ITU-R [02].

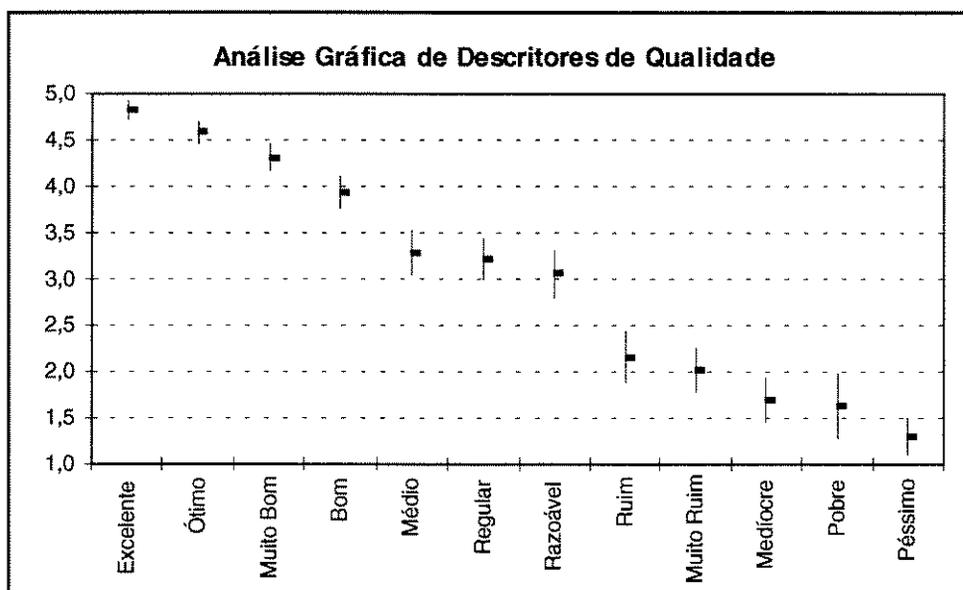


Figura 10: Resultados de Análise Gráfica para Descritores em Português

Da mesma forma, as diferenças observadas com relação a idade do avaliador são significativas apenas para os termos relativos à baixa qualidade. Considerando os quatro termos de menor média, a diferença percentual entre avaliadores acima e abaixo de 25 anos é de 22%. Os avaliadores mais jovens tendem a dar notas mais baixas para o adjetivo 'pobre' e notas mais altas para o 'mediocre'. O termo onde a influência da idade é mais significativa é o adjetivo 'péssimo' para o qual a diferença entre as médias por idade é de 28%, sendo os avaliadores adultos os mais criteriosos.

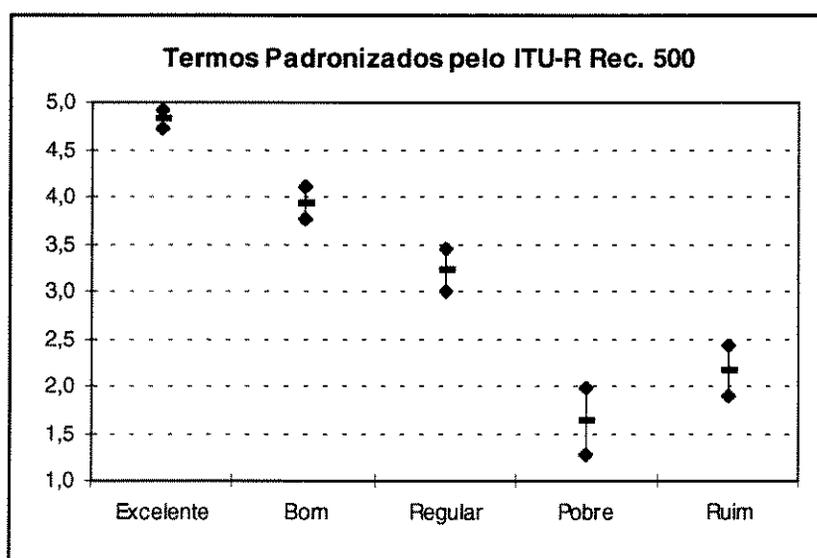


Figura 11: Análise Gráfica dos Termos Padronizados

O resultado mais significativo em relação aos termos padronizados é a inversão na ordem entre os termos *pobre* e *ruim*. Conforme padronizado, com base nos resultados da escala em inglês, o termo “*poor*” se aplica a cenas com qualidade superior às cenas classificadas como “*bad*”. A Figura 11 deixa claro que a tradução literal destes termos para aplicação em testes subjetivos no Brasil produz uma escala que é inadequada por apresentar uma inversão no significado comumente associado aos termos ‘*pobre*’ e ‘*ruim*’. Esta inversão entre estes dois termos está presente em 73,8% das notas.

4.6.3.3 Comparação dos intervalos de qualidade em diversas línguas

A literatura [01] relata experimentos semelhantes utilizando escalas gráficas aplicados em diversos países e línguas para determinar as diferenças de significado e o tamanho dos intervalos entre eles. Na Figura 12 são analisados resultados de experimentos feitos em japonês, alemão, inglês, italiano e francês. Nos testes em alemão foi objeto de estudo, também, analisar se os termos utilizados no escalonamento após a realização do teste de qualidade da imagem, os tornariam mais relevantes para as imagens vistas, mas isso não ocorre.

Observa-se que também em outras línguas, especialmente para francês e italiano, alguns termos tendem a ser agrupados, indicando similaridade ou correspondência de significado no caso extremo. Para o inglês, a escala de qualidade supostamente com cinco pontos e quatro intervalos é na realidade uma escala com quatro pontos e três intervalos com espaçamentos desiguais. Os termos “*bad*” e “*poor*” são muito próximos enquanto que a distância ao termo “*fair*” é relativamente grande.

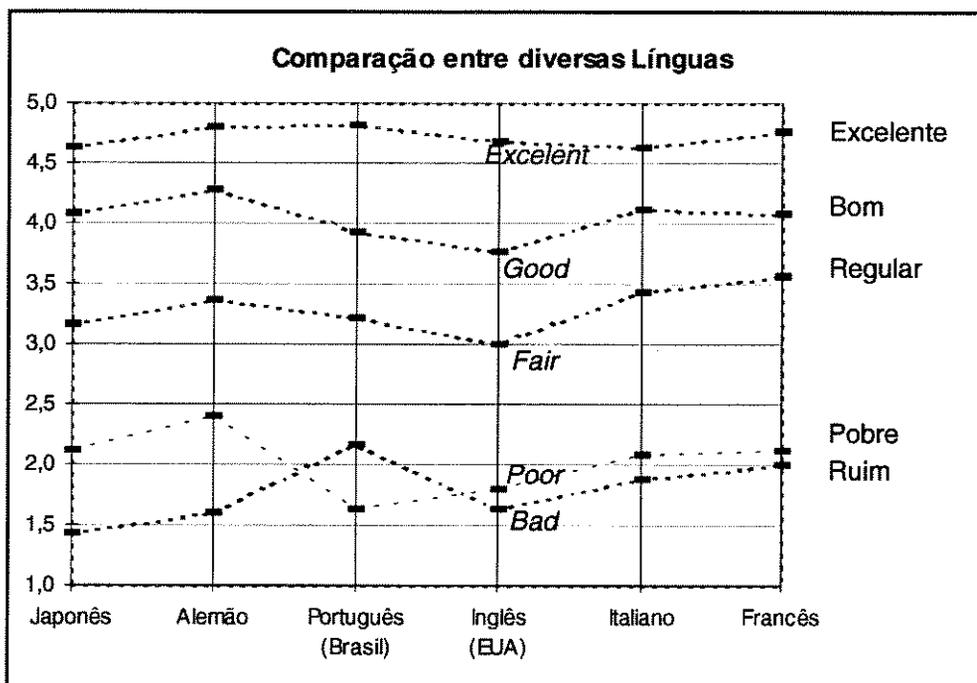


Figura 12: Comparação dos intervalos adotados para os descritores de Qualidade

As escalas de categorias são utilizadas em todo o mundo e os resultados reportados na forma de média e desvio padrão. Verifica-se, entretanto, que a escala de categorias não possui propriedades das escalas de intervalos. Desta forma, testes estatísticos paramétricos; ou seja, aqueles que requerem uma distribuição normal, não são apropriados e a extrapolação para pontos intermediários da escala não é permitida.

4.7 Ferramentas de análise de resultados

A variabilidade é inerente à avaliação subjetiva. A não homogeneidade no julgamento dos avaliadores está relacionada a diversos aspectos que não podem ser controlados, como educação e nível social. Esta característica dos testes subjetivos leva à adoção de procedimentos estatísticos específicos, cujas condições de contorno são delimitadas pelo conhecimento prévio do comportamento do sistema. O formato de apresentação dos resultados subjetivos foi padronizado, na versão 7 da Recomendação ITU-R BT. 500 de 1995.

4.7.1 Procedimentos estatísticos de descarte de notas e avaliadores

A primeira etapa na compilação das notas obtidas em um experimento de avaliação consiste em calcular para cada cena o valor médio e o desvio padrão associado considerando as notas de todos os avaliadores.

A Recomendação ITU-R BT.500 descreve procedimentos de análise da coerência dos avaliadores e define critérios para que todas as notas de um avaliador sejam desconsideradas, em casos de grande discrepância. Da mesma forma fornece diretrizes para o descarte de notas discrepantes.

O descarte de notas se baseia no cálculo da variância e do momento de quarta ordem, em geral calculado separadamente para avaliadores especialistas e não especialistas. A relação entre estes dois valores permite que se estabeleça um critério sobre o tipo de distribuição das notas obtidas e que se determine uma faixa de valores fora da qual as notas são descartadas. Este critério está explicitado a seguir.

$$\beta_2 = \frac{\text{Momento de quarta ordem}}{\text{Variância}} \quad (4.7)$$

β_2	Tipo de Distribuição	Largura da Faixa de Aceitação
$2 < \beta_2 < 4$	normal	2*desvio padrão
$\beta_2 < 2$ ou $\beta_2 > 4$	não normal	$\sqrt{20}$ *desvio padrão

Tabela 5: Definição do critério de descarte de notas

4.7.2 Estimativa do intervalo de confiança associado ao teste

Testes subjetivos são realizados para estimar a opinião média dos usuários quando a imagem é corrompida. É questão fundamental, portanto, a estimativa da opinião média e do intervalo de confiança associado. O intervalo de confiança está intimamente relacionado ao tamanho da amostra considerada e é parâmetro importante para a correta avaliação da influência da distorção na qualidade da imagem.

Seja x_{ij} a nota dada pelo j -ésimo observador à qualidade da i -ésima imagem. O conjunto de todas as notas constitui a população, cuja esperança é chamada opinião média.

$$Q = E[x_{ij}] \quad (4.8)$$

Define-se $\bar{x}^{(j)} = E_i[x_{ij}]$ como a esperança das notas dadas pelo j -ésimo observador. Da mesma forma, define-se $\bar{x}_i = E_j[x_{ij}]$, a esperança das notas subjetivas dadas à i -ésima imagem.

Considerando-se as definições usuais para a variância, temos:

$$\sigma^2 = E[(x_{ij} - Q)^2], \text{ variância da população;}$$

$$\sigma_o^2 = E_j[(\bar{x}^{(j)} - Q)^2]; \text{ variância da distribuição da opinião média dos observadores;}$$

$$\sigma_p^2 = E_i[(\bar{x}_i - Q)^2]; \text{ variância da distribuição das opiniões médias relativas às imagens;}$$

A nota subjetiva x_{ij} é a determinação de uma variável aleatória, significando que o j -ésimo observador pode, em diferentes observações, emitir diferentes opiniões sobre a i -ésima imagem. Entretanto, para assegurar a reprodutibilidade dos testes subjetivos, a estatística da população deve ser invariável. Na discussão que segue é feita esta hipótese.

Para a avaliação da opinião média da população e de seu intervalo de confiança, é prática comum o uso de amostragem duplamente estratificada. Uma amostra deste tipo é tal que um grupo de n observadores e m imagens são escolhidas aleatoriamente e cada observador emite sua opinião sobre cada imagem. Conseqüentemente para uma amostra de $n.m$ notas subjetivas temos x_{ij} tal que $i=1,2,\dots,m$ e $j=1,2,\dots,n$.

No caso de uma amostra duplamente estratificada, como descrito acima, a estimativa da opinião média é da forma:

$$\bar{x} = \sum_{i=1}^m \sum_{j=1}^n \frac{x_{ij}}{n.m}; \text{ onde } \bar{x} \text{ é uma estimativa não polarizada de } Q.$$

Para as variâncias temos:

$$s^2 = \sum_{i=1}^m \sum_{j=1}^n \frac{(x_{ij} - \bar{x})^2}{n.m} \quad (4.9)$$

$$s_o^2 = \frac{1}{n} \sum_{j=1}^n \left(\sum_{i=1}^m \frac{x_{ij}}{m} - \bar{x} \right)^2 \quad (4.10)$$

$$s_p^2 = \frac{1}{m} \sum_{i=1}^m \left(\sum_{j=1}^n \frac{x_{ij}}{n} - \bar{x} \right)^2 \quad (4.11)$$

A variância da população, da distribuição das imagens e dos observadores se relacionam às estimativas acima da forma:

$$\sigma^2 = E \left[\frac{((m-1)(n-1)-1)s^2 + ns_p^2 + ms_o^2}{(m-1)(n-1)} \right] \quad (4.12)$$

$$\sigma_o^2 = E \left[\frac{mns_o^2 - n(s^2 - s_p^2)}{(m-1)(n-1)} \right] \quad (4.13)$$

$$\sigma_p^2 = E \left[\frac{mns_p^2 - m(s^2 - s_o^2)}{(m-1)(n-1)} \right] \quad (4.14)$$

Assumindo que a distribuição de \bar{x} é normal, pelo teorema do limite central temos, que os limites de confiança para a opinião média com confiabilidade de 95% é dado por:

$$\bar{x} - \frac{\delta}{2} \leq Q \leq \bar{x} + \frac{\delta}{2}; \text{ com } \delta = 4\sigma_{\bar{x}} \quad (4.15)$$

onde $\sigma_{\bar{x}}^2$ é a variância da distribuição de \bar{x} e δ é o intervalo de confiança de 95%.

A variância $\sigma_{\bar{x}}^2$ vale:

$$\sigma_{\bar{x}}^2 = E[(\bar{x} - Q)^2] = \frac{\sigma^2 + (n-1)\sigma_p^2 + (m-1)\sigma_o^2}{n \cdot m} \quad (4.16)$$

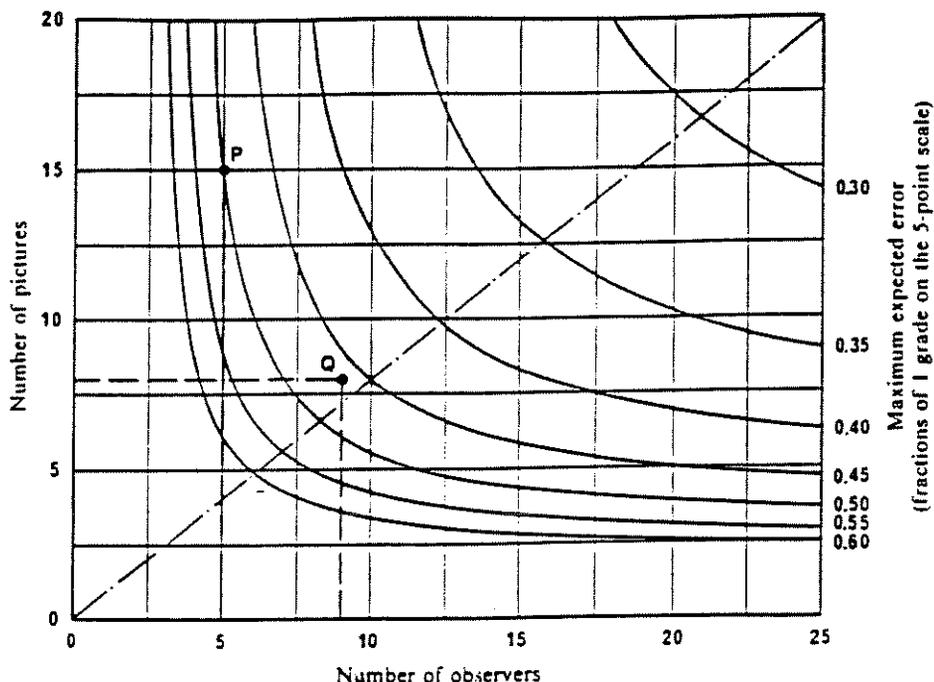


Figura 13: Curvas com o erro máximo esperado em testes subjetivos

Assumindo-se o modelo de Miceli e Orlando [49] para uma amostra duplamente estratificada, o intervalo de confiança é uma função da opinião média Q , do número de observadores n e do número de imagens m . O interesse recai então na possibilidade de minimizar

a amostra, escolhendo n e m de tal forma que o intervalo de confiança esteja abaixo de um valor previamente estipulado. Os gráficos do ITU-R fornecem curvas para a minimização do tamanho da amostra, sendo que os valores de intervalo de confiança assinalados são obtidos com Q próximos a 3,0, na escala de 1 a 5 [01].

A Figura 13 apresenta o erro máximo esperado em um experimento de avaliação subjetiva em função do número de avaliadores e do número de cenas empregados, para um intervalo de confiança de 95%. A Figura 14 traz um fator de correção a ser aplicado ao erro máximo nos casos em que a média geral do teste não é 3,0.

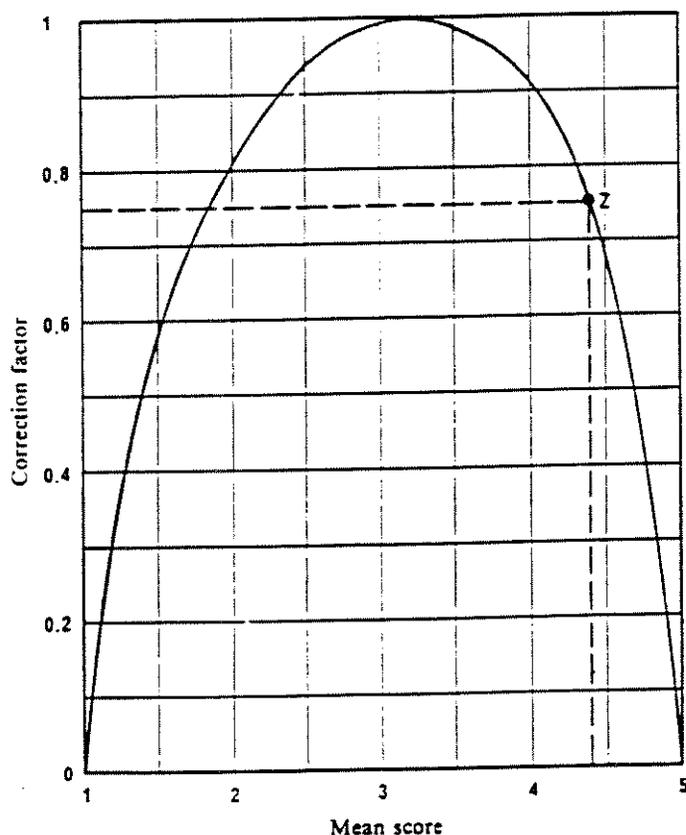


Figura 14: Fator de correção do erro máximo com base na média geral do teste

4.8 Novos problemas: novas metodologias

Neste item são abordadas metodologias não padronizadas, mas que oferecem boas perspectivas de aplicação e resultados.

4.8.1 Escalonamento multidimensional de imagens

A abordagem de projeto de experimento em múltiplas dimensões trata de métodos para se analisar o efeito subjetivo da combinação de degradações que escapam ao caso aditivo

[48] para descobrir dimensões de percepção ortogonais subjacentes à avaliação de qualidade e em seguida determinar como estas dimensões se relacionam aos parâmetros físicos.

O escalonamento multidimensional é uma ferramenta bastante útil para a determinação de “clusters” relacionados ao mesmo atributo de percepção. É uma ferramenta que permite a identificação do número de dimensões numa comparação entre diversas imagens de televisão. Este método vem ao encontro da necessidade de avaliação das recentes técnicas de codificação digital, pois permite a monitoração quantitativa e pode ser usada para avaliar diferenças entre sistemas de forma mais informativa. Com os métodos usuais de avaliação, surgem problemas quando há significativo avanço da tecnologia. Nestes casos, estímulos que receberam notas correspondentes a alta qualidade passam a receber notas mais baixas, principalmente em função do efeito de contexto. Pela abordagem de escalonamento multidimensional, vê-se que os valores associados ao estímulo permanecem constantes embora a tecnologia evolua. A exposição a novas tecnologias altera somente a forma como estes fatores são mapeados na qualidade final.

As vantagens em flexibilidade são contrabalançadas pela maior complexidade e maior tempo gasto na coleta dos dados. Em especial, a parte mais demorada consiste justamente na identificação das dimensões de percepção, pois nesta etapa devem ser realizados testes envolvendo todos os estímulos dois-a-dois ou, pelo menos, uma seleção cuidadosa dos possíveis estímulos. Uma vez tendo obtido um número adequado de dimensões físicas na área de interesse, teoricamente, novas dimensões poderiam ser integradas a estrutura através de medidas exploratórias mais simples.

O ganho na utilização de múltiplas dimensões é a possibilidade de descrição formal de toda a cadeia de eventos desde os parâmetros físicos até a qualidade percebida da imagem e de utilização na otimização de serviços. A descrição de como as dimensões físicas são mapeadas em dimensões psicológicas auxiliam na predição do efeito combinado das degradações. A seguir, a descrição dos pesos mostrando o impacto relativo das dimensões psicológicas na qualidade final é útil para identificar melhorias efetivas dos serviços.

4.8.1.1 Princípios da análise em múltiplas dimensões

A tarefa de medir impressões subjetivas tem uma longa história. A questão original consiste em descobrir uma função psicofísica, ou seja, uma função capaz de relacionar a magnitude física das degradações às sensações psicológicas. As técnicas de escalonamento são ferramentas para o desenvolvimento de modelos que mapeiem as impressões subjetivas em espaços psicológicos de múltiplas dimensões, ou seja, que determinem como variações em dimensões específicas da imagem se relacionam à avaliação subjetiva da qualidade de imagens de televisão.

A premissa do escalonamento em múltiplas dimensões é que a qualidade percebida pelo avaliador é determinada não pelas dimensões físicas como relação sinal ruído por exemplo, mas sim, por dimensões de percepção. Obviamente, os parâmetros físicos afetam os parâmetros percebidos na imagem, esta alteração pode, porém, não ser um-para-um. Diversos parâmetros físicos pode afetar uma dimensão percebida ou, ainda, um parâmetros físico pode afetar diversas

dimensões perceptíveis. De qualquer forma, o julgamento quanto à qualidade de uma imagem se forma a partir de uma combinação, possivelmente aditiva, dos pesos que um estímulo gera em cada uma das dimensões.

É necessário, portanto, identificar as dimensões perceptíveis apropriadas, determinar as relações entre as dimensões físicas e perceptuais e os pesos que cada uma das dimensões, em percepção, têm na determinação da qualidade da imagem.

A análise multidimensional começa com a criação de um estímulo que apresenta variações em diversas dimensões físicas. Para Goodman e Pearson [32] é importante que esta variação englobe o maior número de dimensões possíveis e que seja sistemática em cada uma delas. Os avaliadores julgam os estímulos em pares, apontando semelhanças (ou diferenças) para cada par. As notas dadas refletem as semelhanças (ou diferenças) percebidas e são tomadas como uma medida da distância entre os estímulos num espaço psicológico multidimensional. As distâncias obtidas são usadas para reconstruir o espaço psicológico de dimensão-n.

4.8.1.2 Consideração de projeto em Múltiplas Dimensões

Tendo construído um modelo para o espaço psicológico, uma estratégia para determinar as relações entre as dimensões em percepção e as dimensões no espaço físico é a utilização de regressão não linear. O princípio é simples e intuitivo. As dimensões físicas que provocam variações significativas em uma dimensão psicológica contribuem para esta dimensão. Este critério define dentre o conjunto de parâmetros físicos aqueles que possibilitam uma predição da qualidade de imagem e os pesos associados a cada um deles.

O objetivo desta análise é a obtenção de uma função que relacione a qualidade aparente de um estímulo a uma soma de valores em cada uma das dimensões identificadas, fornecendo bases para a predição de geral de qualidade. Um modelo construído desta forma, permite também analisar as variações nos julgamentos entre avaliadores, pelas diferenças de pesos em cada uma das dimensões da percepção, fornecendo desta forma, um índice de confiabilidade nas predições de qualidade. Estas informações relativas ao impacto subjetivo (peso) de cada dimensão de qualidade são extremamente importantes para a especificação e melhoria dos serviços.

Segundo esta abordagem, o modelo segundo o qual as dimensões da percepção se combinam de forma aditiva para determinar a qualidade percebida da imagem, conforme proposto por Lewis e Allnatt [43], é consistente, mas não se aplica a todos os casos. Dado que as dimensões percebidas da imagem podem se combinar de forma aditiva; as dimensões físicas mapeadas em diferentes dimensões de percepção, isto é, com efeitos diferentes também se combinam de forma aditiva. Entretanto, não há nenhuma garantia de que dimensões físicas mapeadas na mesma dimensão de percepção, isto é, com efeitos semelhantes se comportem da mesma forma.

4.8.1.3 Dimensões de qualidade: Resultados de estudos preliminares

Experimentos realizados com o objetivo de descrever os diversos atributos na formação da qualidade global da imagem identificaram: *definição*, *ruído* e *contraste*, como as três dimensões de qualidade mais importantes. Este resultado é surpreendente por não incluir a cor. No experimento de Teunissen e Westerink [61], aqui analisado, cenas com diferentes parâmetros físicos como luminância de pico, contraste de luminância, resolução espacial e reprodução de cores, são comparadas com respeito à percepção da qualidade.

No experimento analisado, foram selecionados quinze atributos para a avaliação: resolução, brilho das cores, “flicker” em grande áreas, brilho, definição na transição de cores, “flicker” em linha, naturalidade do contorno da imagem, contraste, escuridão das áreas escuras, contornos, naturalidade das cores, ruído, distorção geométrica da imagem, detalhes e distorções da imagem. O estudo destes pesquisadores procura identificar neste conjunto de atributos aqueles que são julgados de forma similar pelos avaliadores. Entretanto, estas semelhanças podem ser resultado de semelhança na interpretação dos observadores ou simplesmente consequência do restrito conjunto de imagens explorado.

Os resultados médios são analisados com base em um gráfico onde as distâncias entre os pontos é dada pela equação (4.17).

$$d = \frac{(1-r)}{2}, \text{ onde } r \text{ é a correlação entre os atributos} \quad (4.17)$$

Atributos avaliados de forma similar têm alta correlação e por isso são plotados próximos. Os 16 atributos, incluindo qualidade geral da imagem, pode sempre ser plotados em um espaço 15-dimensional, de tal forma que as distâncias derivadas neste espaço correspondam exatamente aos valores medidos. Entretanto, embora uma dimensão elevada se ajuste melhor aos dados, isto não significa que, de fato, uma dimensão tão alta seja necessária. No caso do experimento descrito, três dimensões são suficientes para explicar cerca de 93% da variância observada [20]. A introdução de novas dimensões, não representa significativa melhora para explicar as variações observadas. Em três dimensões é possível fazer uma representação gráfica dos atributos, neste caso, a formação de “clusters” evidencia a correlação entre os atributos.

Os atributos de maior relevância na definição da qualidade de uma imagem são: definição e naturalidade. De fato, apenas estes dois atributos fornecem uma boa predição para a qualidade percebida de imagem. O interessante no resultado obtido é a correlação entre os diversos atributos e a qualidade geral da imagem. Há uma alta correlação entre o grupo de atributos relacionados à definição e às notas de qualidade, o que sugere que as notas de qualidade têm uma relação linear com a impressão de resolução. Já a correlação com os atributos de naturalidade é pequena, sugerindo assim uma relação não linear entre naturalidade e qualidade da imagem.

Quando as cenas observadas são idênticas com exceção do esquema de cores os observadores dão notas diferentes não apenas para naturalidade, mas também para a impressão de definição. As notas relativas a naturalidade por sua vez não são muito afetadas por mudanças na resolução. Isto foi mostrado pelo decrescimento do peso dado à resolução após a subdivisão das cenas em teste em duas classes de cores, uma onde predominavam as tonalidades vermelhas

e outra na qual os tons de azul e verde predominam. Este dado é uma comprovação de que o desbalanceamento das componentes RGB afeta mais a qualidade percebida da imagem do que as diferenças de resolução. No que se refere ao balanceamento das componentes RGB, os avaliadores têm significativa preferência pelas imagens avermelhadas em comparação com os tons de azul e verde. Esta preferência se reflete em todas as notas. Além disso, verificou-se que mudança no contraste tem um importante papel na percepção de definição e, portanto, na qualidade percebida da imagem.

Os avaliadores utilizam uma estrutura hierárquica na avaliação da qualidade percebida da imagem. Primeiramente, a maior diferença perceptual é avaliada e só então as diferenças de percepção menores são levadas em consideração. Em alguns casos as diferenças observadas não são consideradas relevantes aos observadores e somente são levadas em conta quando adicionadas diferenças substanciais.

4.8.2 Refinamento da metodologia de “escala de degradação a estímulo duplo”

O refinamento da metodologia de “escala de degradação a estímulo duplo” a que se refere este trabalho trata de propor variações ao método que permitam melhorar a capacidade de discriminação de pequenas diferenças de qualidade. Neste contexto, a confiabilidade do testes será tomada como a habilidade em discriminar pequenas diferenças de qualidade. O ponto de partida para este estudo é uma comparação entre o método de “escala de qualidade contínua a estímulo duplo” e o método de “escala de degradação a estímulo duplo”.

No método de “escala de qualidade contínua a estímulo duplo”, descrito na Recomendação ITU-R BT. 500 [02], assim como no método de “escala de degradação a estímulo duplo” as cenas são avaliadas em pares. Uma das imagens do par é a imagem de referência, enquanto a outra pode, ou não estar degradada. Ao contrário do método “escala de degradação” entretanto, o avaliador não conhece a posição da imagem de referência no par, escolhida aleatoriamente. Adicionalmente, o observador assiste a cada um dos pares do experimento quantas vezes achar necessário, contra apenas uma apresentação na metodologia definida como “escala de degradação”.

A proposta é uma modificação no método “escala de degradação” no que diz respeito à ordem de apresentação das cenas de referência e do número de apresentações para torná-lo, nestes aspectos, semelhante ao método de “escala de qualidade”. A estratégia proposta, para o caso de cenas dinâmicas, é apresentar as seqüências por duas vezes, sendo a votação feita durante a segunda exibição. No caso de cenas estáticas, o tempo de exibição das cenas pode ser menor, 3 ou 4 segundos, porém o número de repetições deve ser maior, tipicamente, cinco vezes. Preserva-se também a aleatoriedade no posicionamento da cena de referência dentre as cenas do par.

A motivação para estas alterações no método padronizado é que no método de “escala de qualidade”, as notas são assinaladas em linhas contínuas divididas nas cinco categorias de qualidade já citadas. Como mostrado no experimento com escalas gráficas para as categorias em

português realizado, a confiabilidade destas medidas é pequena, pois os intervalos entre as categorias de qualidade não são uniformemente espaçados. Para as categorias de degradação, entretanto, os intervalos se distribuem de forma aproximadamente linear. A idéia então é obter os dados através da escala de degradação e interpretá-los em termos descritivos.

Em [51] Narita compara os resultados para uma escala de degradação com resultados onde as notas expressam diretamente a diferença de qualidade em relação à imagem de referência, numa escala de qualidade. As conclusões deste experimento são indicativas de que o método de “escala de degradação” modificado é mais confiável do que o método de “escala de qualidade” quando pequenas diferenças de qualidade estão envolvidas. O pesquisador obteve resultados equivalentes para os dois métodos em condições em que na “escala de qualidade” as cenas são apresentadas duas vezes e no método de “escala de degradação” apenas uma vez.

5. Estimação de qualidade subjetiva com base em parâmetros objetivos

A necessidade de medir a qualidade de vídeo é consequência natural do desenvolvimento destes equipamentos. As duas categorias principais para avaliação de vídeo são avaliação subjetiva que utiliza observadores humanos e a avaliação objetiva que se utiliza de medidas físicas da imagem.

As atividades em avaliação objetiva de qualidade de vídeo comprimido se restringiram, inicialmente, a aplicação de sinais de teste artificiais, para medidas de ganho, deslocamento, resposta em frequência 2D e área ativa. A natureza custosa e demorada dos testes subjetivos, aliada ao fato de estarem baseados somente nos artefatos visíveis, motiva a elaboração de métodos de avaliação de desempenho de codecs baseada em medidas objetivas.

Assim, embora acredite-se que a melhor avaliação da qualidade de vídeo é feita por observadores humanos, é necessário o desenvolvimento de medidas objetivas de qualidade de vídeo comprimido. Medidas são necessárias para as organizações de padronização, usuários finais e provedores de serviços. Medidas objetivas imparciais, confiáveis, repetíveis e de baixo custo, tanto de vídeo como de desempenho dos sistemas de transmissão de imagem, aumentam a competição entre os provedores de serviços e equipamentos e, também, permite melhor especificação e avaliação de novos sistemas. A possibilidade de realização de medidas objetivas em tempo real abre a perspectiva de monitoração contínua da qualidade do vídeo. Além disso, espera-se que inúmeros novos algoritmos de codificação de imagem sejam desenvolvidos, para os quais a hipótese de se realizar testes subjetivos para todos os novos sistemas não é factível.

O objetivo é obter medidas com boa correlação com os resultados subjetivos, de forma que se possa prever a aceitação dos usuários frente aos artefatos de compressão. Deseja-se que estas medidas forneçam valores absolutos dos erros de codificação, e que se desenvolvam regras para sua concatenação, possibilitando assegurar o nível de qualidade desejado ao longo da cadeia de produção.

O desenvolvimento de um modelo de qualidade de vídeo geral e útil para sistemas de operação e transmissão de vídeo é, entretanto, um processo bastante complicado. Um exemplo simples ilustra a complexidade envolvida. Considere duas aplicações diferentes, transmissão de gráficos de alta resolução e transmissão de linguagem de sinais. Estas duas aplicações requerem diferentes características de desempenho em diversas dimensões da qualidade de vídeo; por exemplo, a reprodução de cores, resolução espacial e temporal. A aplicação de transmissão de gráficos requer alta resolução espacial com baixas taxas de quadro, enquanto que a transmissão de linguagem de sinais requer alta taxa de quadros porém com baixa resolução espacial. Um modelo objetivo que produz estimativas gerais de qualidade de um conjunto de parâmetros específico, deve ser capaz de lidar com este tipo de efeito dependente da aplicação.

5.1 Considerações gerais: O que medir?

Os trabalhos de predição de qualidade subjetiva com base em medidas objetivas têm dois objetivos teóricos básicos: encontrar o melhor conjunto de medidas objetivas para predizer o resultado subjetivo e determinar quão perto do ótimo estão os preditores.

Como desenvolver os parâmetros de desempenho de vídeo e como eles devem ser relacionados a percepção humana pode dar margem a considerações filosóficas. Considere um tipo de degradação que não seja detectado por nenhum observador. Neste caso, as medidas objetivas devem ou não ter sensibilidade suficiente para detectar este tipo de deficiência do sinal? Se as medidas objetivas forem tomadas como substitutivas das medidas subjetivas, então os parâmetros objetivos devem se comportar de forma similar e ser insensível a este tipo de degradação. Parece claro, entretanto, que ter conhecimento de que a distorção existe é desejado, tanto para o conhecimento do usuário quanto para o isolamento e manutenção do sistema defeituoso.

Estas considerações mostram a necessidade do desenvolvimento de uma hierarquia de medidas objetivas. Esta hierarquia deve levar em conta o usuário final bem como o provedor do serviço e garantir flexibilidade na especificação dos parâmetros de desempenho de um grande número de sistemas e aplicações. A Figura 15 ilustra este raciocínio e evidencia a necessidade de experimentos para a determinação das dimensões perceptíveis de qualidade e a necessidade de desenvolvimento de parâmetros para quantificar cada uma destas dimensões. Na Figura 15, assume-se que as dimensões de qualidade de vídeo podem ser independentemente quantificadas por parâmetros objetivos. Sempre existe, no entanto, certo nível de dependência entre os parâmetros objetivos.

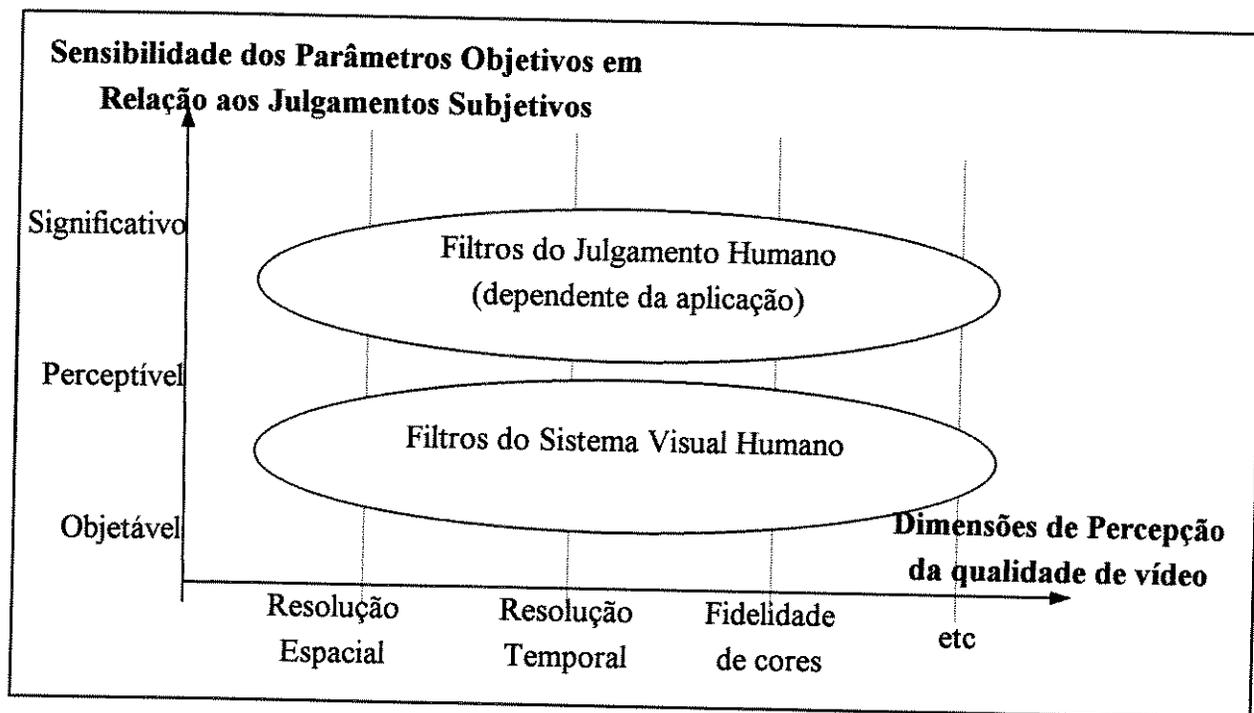


Figura 15: Hierarquia de desenvolvimento de parâmetros objetivos

Esta abordagem, contrariamente às tentativas de desenvolvimento de um único parâmetro, garante máxima flexibilidade na especificação dos atributos do sistema e no casamento destes atributos à qualidade. Imagens médicas por exemplo, requerem alta resolução espacial, enquanto que imagens para aplicações em teleconferência necessitam apenas de uma resolução espacial moderada para serem julgadas “excelentes”.

Observe que no eixo vertical estão representados os graus de sensibilidade dos parâmetros objetivos em relação aos julgamentos subjetivos. No nível mais baixo, as medidas objetivas detectam distorções que não são visualmente perceptíveis. O terceiro nível leva em conta as medidas que afetam o julgamento humano; é o nível mais difícil de ser quantificado. Os filtros de julgamento humano devem transformar as distorções percebidas em medidas de distorção significativas, levando em conta a aplicação e a população envolvidas.

5.2 Norma ANSI de parâmetros objetivos

Diversos grupos vêm trabalhando no desenvolvimento de novas medidas objetivas de qualidade de imagem, entre eles o centro de pesquisa do RAI; o laboratório David Sarnoff, National Telecommunications and Information Administration, Institute for Telecommunication Sciences (NTIA/ITS) e seus associados, Bellcore e KPN Research. Em 1996 a ANSI adotou ANSI T1.801.03 [06]. Este documento traz novas métricas para avaliação objetiva de qualidade de vídeo para quantificar o efeito da compressão e transmissão digital. A norma não se limita a sinais de teste artificiais, mas utiliza cenas naturais e especifica parâmetros sensíveis a distorções no codificador, no decodificador e no canal de transmissão. A norma é bastante abrangente, mas destina-se, preferencialmente, à comparação de uma condição operacional atual de determinado sistema com o desempenho verificado em uma configuração anterior do mesmo sistema.

O primeiro aspecto considerado é a calibração dos equipamentos. Desconsiderar a calibração pode produzir grandes erros de medidas nos valores dos parâmetros. Esta norma ANSI especifica métodos robustos para medida de ganho, nível de “offset” e deslocamento de área de vídeo ativa, ou seja, método para o registro espacial dos quadros de entrada e de saída. Entretanto, os métodos propostos requerem a utilização de cenas paradas e são necessárias adaptações da metodologia para aplicações em cenas naturais com movimento.

Da mesma forma, cuidado deve ser tomado para garantir o alinhamento temporal no registro dos quadros de vídeo. Em geral há desalinhamento entre entrada e saída, como resultados dos atrasos acumulados nos circuitos de compressão e transmissão. Existem dois métodos para o alinhamento do vídeo. O primeiro método, alinhamento constante, resulta em uma única medida de alinhamento para todo o “stream” de vídeo. Um segundo método fornece medidas de atraso relativas a cada quadro. Este dois métodos produzem resultados bastante diferentes nos casos em há descarte de quadros na codificação. Em sistemas de alta qualidade, como é o caso de MPEG-2, os métodos de alinhamento constante têm se mostrado uma técnica simples e que produz bons resultados. É esperado que em geral, as medidas obtidas para alinhamento temporal constante tenham maior correlação com os resultados subjetivos do que as

medidas obtidas a partir de alinhamento variável, uma vez que este último remove as penalidades do descarte de quadros.

A primeira dentre as medidas ANSI é o conjunto das estatísticas espaciais, indicativas de quantidade e tipo de informações espaciais ou bordas na cena. Em segundo temos estatísticas temporais indicativas de quantidade e tipo de informação temporal ou movimento na cena. Além disso, os gradientes espaciais e temporais são úteis pois fornecem indicações da quantidade de informação perceptível ou de mudança na imagem. O uso de grandezas matriciais empregadas para representar as degradações tem a grande desvantagem de requerer enorme capacidade extra de armazenamento e largura de banda para sua transmissão.

Características como detalhes espaciais e temporais são transformadas em quantidades ou informações associadas a cada um dos quadros. Os parâmetros são calculados utilizando funções de comparação que operam em na seqüência de entrada e de saída. Alguns parâmetros escalares, ou seja, baseado em medidas escalares da imagem e não em medidas vetoriais ou matriciais, têm se mostrado útil, com boa correlação com resultados subjetivos, indicando que a quantidade de informação requerida do vídeo de entrada para se obter uma medida de qualidade significativa é muito menor do que todo o quadro de vídeo.

As medidas na norma ANSI T1.801.03 foram selecionadas após extensivos testes de avaliação de qualidade de imagem realizados em diversos laboratórios. Os estudos realizados envolveram sistemas com taxas variando entre 64kbits/s a 45 Mbit/s e cenas com diferentes níveis de dificuldade de codificação a nível espacial e temporal. Os testes objetivos dos equipamentos foram acompanhados de julgamentos subjetivos. Embora 25 sistema tenham sido testados, os testes não incluíram sistemas MPEG e não cobriram taxas entre 1.6 e 10Mbits/s. As medidas propostas a seguir, neste capítulo, e o experimento descrito no Capítulo 6 cobrem justamente esta lacuna.

5.3 Linhas de pesquisa

Embora a norma ANSI permita descrição de diversas características dos codecs de vídeo, no que diz respeito à estimação de qualidade subjetiva, não apresenta resultados conclusivos. Muitos trabalhos estão em desenvolvimento nesta área em todo o mundo trazendo diferentes abordagens para o problema. As diversas maneiras de se obter uma caracterização objetiva do sistema bem correlacionada ao seu desempenho subjetivo têm limitações e aplicações diferentes, mas, desafios e objetivos comuns.

5.3.1 Modelo do sistema visual humano

A primeira tentativa de se procurar obter uma resposta equivalente a do sistema visual humano se deu através do desenvolvimento de modelos visuais. A grande dificuldade encontrada nesta tarefa é que um mesmo modelo não é completamente aplicável em situações de alta e baixa

degradação.

No desenvolvimento de um modelo visual, dois níveis de conhecimento devem ser incorporados. O primeiro deve caracterizar quão visível é uma determinada degradação dada sua localização na imagem, e o segundo consiste em determinar como a visibilidade de todas as degradações devem ser combinadas para se obter um determinado nível de qualidade.

Os modelos desenvolvidos até o momento tratam da resposta visual humana no limite mínimo de visibilidade, o que é adequado para aplicações em HDTV, por exemplo. Os modelos relativos a respostas para grandes distorções ainda são pouco explorados, em especial o desenvolvimento de modelos para a determinação objetiva de medidas de qualidade na presença de degradações que surgem como consequência do processo de codificação digital e dos processos adaptativos não lineares presentes nos codificadores.

Os primeiros estudos de correlação entre medidas objetivas e os resultados da percepção humana para qualidade de uma imagem se concentraram no desenvolvimento de modelos de baixo nível, ou seja aqueles que procuram descrever o funcionamento da visão. Estes modelos são adequados para imagens paradas. Através deles foi possível identificar, que no processamento de imagens estáticas monocromáticas, os principais aspectos da observação humana são a adaptabilidade à iluminação da cena, decomposição em escalas, seleção de objetos com base na orientação e mascaramento por padrões específicos [68].

Entretanto, dada à complexidade de movimento no vídeo digital e aos cortes de cena em contraposição à limitada capacidade de processamento do sistema visual e da memória de trabalho humano, a extrapolação dos resultados para imagens paradas na avaliação da qualidade de imagens em movimento não é eficiente. É necessário incorporar a forma como os observadores reagem às degradações visíveis e variantes no tempo resultantes da codificação digital, levando também em consideração os processos de memória humana. É fundamental, portanto, a inclusão de aspectos psicológicos no modelo de avaliação de qualidade [30]. Em função disso, os modelos originados do olho humano deram lugar ou foram associados a simuladores do sistema cognitivo humano, correspondentes a modelos de alto nível.

Em particular, a visibilidade de uma degradação depende dos quadros adjacentes. Claramente este aspecto aumenta a complexidade do problema, fazendo com que o julgamento de degradações variantes no tempo se desvie do julgamento emitido para um único quadro. Observa-se que se o intervalo entre dois estímulos é menor do que a duração da memória de trabalho humano, a resposta aos dois estímulos é combinada. Desta forma, quando trechos não degradados de curta duração são alternados por trechos com degradação visível, todos os quadro parecem distorcidos. Este efeito, entretanto, se limita ao mascaramento de trechos não distorcidos alternados por trechos com degradação visível e não o contrário.

Os esforços para o desenvolvimento de uma métrica capaz de prever de forma rápida e precisa a magnitude das diferenças perceptíveis entre imagens, resultou no desenvolvimento do modelo "Sarnoff" [47] para diferenças perceptíveis. O modelo do laboratório Sarnoff estabelece um mapeamento entre duas imagens de entrada em uma imagem de saída, cujos valores dos pixels procuram estimar unidades de percepção. A imagem de saída define a probabilidade e a região em que as diferenças são observadas, estabelecendo uma relação aproximadamente linear

entre a magnitude dos pixels e a percepção das diferenças.

O modelo aplicado faz uma decomposição piramidal da imagem em bandas de frequência para facilitar um tratamento diferenciado em cada região da percepção, seguida por uma etapa de normalização. A normalização corresponde a um ajuste do ganho de acordo com um valor médio temporal da componente de luminância. Esta mediação temporal procura representar os fenômenos relativos à insensibilidade relativa ao nível de luminosidade e à perda de sensibilidade na transição entre uma cena clara para uma cena escura. Após esta etapa são feitas medidas de contraste, três medidas diferentes são calculadas. A ponderação das medidas de contraste é feita de acordo com o nível da decomposição piramidal, procurando modelar os efeitos de decréscimo de sensibilidade em regiões de grande atividade.

Este modelo, bastante completo, tem sido aplicado, com bom nível de correlação subjetiva, na predição de discriminação e fidelidade.

5.3.2 Medidas no “bitstream”

Medidas objetivas para a avaliação de imagem podem também ser feitas diretamente no “stream” de vídeo codificado. Este “bitstream” carrega diversos parâmetros de codificação que devem ser transmitidos ao decodificador para que o sinal seja corretamente decodificado. Para cada macro-bloco codificado estão disponíveis nos bits codificados, a escala de quantização, os vetores de movimento e os modos de codificação.

O passo de quantização é um fator importante na qualidade da imagem e está relacionado ao ruído de quantização introduzido pelo codificador. Vale lembrar que, a matriz de quantização leva em conta a sensibilidade do olho a diferentes frequências espaciais. Não são consideradas, entretanto, variações da sensibilidade do olho a estímulos visuais em cenas com diferentes atividades, ou seja, o fato de existirem regiões da imagem nas quais a percepção de artefatos é limitada. Na compressão de imagens, a granularidade da quantização é função da “criticalidade” do bloco e preponderantemente do estado do “buffer”, de forma que, ao analisar o valor dos parâmetros de escalonamento da quantização, não é possível separar os dois efeitos.

Nos experimentos do RAI [16], o máximo valor do parâmetro **quantizer-scale**, encontrado em pelo menos seis quadros tipo B, foi correlacionado com as medidas objetivas. Essa restrição quanto ao número mínimo de quadros, se deve ao fato de que uma degradação deve durar algum tempo - alguns quadros - para que seja perceptível para os avaliadores. Para estes pesquisadores italianos a informação espacial não está diretamente, mas indiretamente relacionada à qualidade. A conclusão está baseada na observação da informação espacial, quadro-a-quadro. Em uma das seqüências testadas é possível observar importantes degradações, como distorções e efeito de bloco, por alguns segundos no final do trecho avaliado. O parâmetro de informação espacial decresce neste trecho, embora o valor médio do **quantizer-scale** permaneça constante durante toda a seqüência. Relacionando estes dois efeitos, o parâmetro de avaliação selecionado é o produto entre o **quantizer-scale** e o inverso da informação espacial. Entretanto, a correlação entre os resultados da avaliação subjetiva das cenas e parâmetros

objetivos calculados a partir do “bitstream” de vídeo, não é satisfatória para todas as cenas.

5.3.3 Medidas em cenas naturais

Nesta linha de pesquisa, a multiplicidade de opções é ainda maior. Uma das opções é realizar um conjunto de medidas da imagem original e transmiti-la como informação lateral para o decodificador. As mesmas medidas são realizadas sobre a imagem decodificada e os resultados comparados. Mais correlacionado com a atividade de avaliação subjetiva, entretanto, são os trabalhos que procuram extrair parâmetros da imagem processada, com o sem o conhecimento da imagem original, e relacioná-los a uma base de dados subjetiva. A ideia geral para este tipo de abordagem é testar os equipamentos de compressão e obter dois tipo de dados: um conjunto de medidas objetivas e um conjuntos de notas obtidas por avaliação subjetiva.

As características de ruído e redundância inerentes a uma base de dados subjetiva dificultam o problema de encontrar o melhor conjunto de descritores objetivos de qualidade. O ruído na base de dados faz com que diferentes conjuntos de preditores se ajustem igualmente bem aos dados. Porém, este casamento nunca é de 100%. Como o conjunto de dados tem redundância, é possível fazer estimativas de qualidade através de subconjuntos distintos de variáveis, obtendo essencialmente o mesmo resultado. Verifica-se porém, que se muitas variáveis redundantes forem utilizadas como preditores, os resultados podem ser instáveis, especialmente na presença de ruído.

Dadas estas limitações, os trabalhos de análise de dados se concentram em obter um conjunto significativo de preditores através do qual seja possível fazer generalizações, sem buscar um casamento de 100%, sabidamente impossível. Por isso, a confiabilidade também é um item importante, estabelecendo os limites de casamento mesmo para medidas objetivas perfeitas. A definição teórica de confiabilidade é dada pela razão entre a variância da imagem, se fosse perfeitamente medida, e esta variância acrescida de erro. E, pela própria definição, não é um valor computável.

Uma possível abordagem estatística, para a estimativa desta confiabilidade real e para análise dos “clusters” de parâmetros objetivos obtidos, se baseia em estipular um número máximo de variáveis a ser utilizado com base em medidas de redundância do subconjunto de variáveis escolhido. As medidas candidatas a preditores de qualidade subjetiva podem substituir um ou mais preditores já existentes se for uma generalização dos anteriores. Sempre que possível são escolhidas variáveis que sabidamente têm relação de causalidade com as variáveis dependentes. Este processo chama-se análise em componentes principais. A medida da redundância determinada por esta análise é o número de variáveis não correlatadas necessárias para representar a maior parte da variância do conjunto de medidas. Este tipo de análise é dita análise em componentes principais.

Em geral, o número de componentes principais é tomado como o número de autovalores maiores do que um. Na prática, a análise é considerada bem sucedida se é capaz de justificar cerca de 70% a 80% da variância da base de dados com um número de variáveis igual a um

terço ou um quarto do número original de variáveis.

5.3.3.1 Método de estimação empregado: Casamento do banco de dados subjetivo

Está claro que a avaliação objetiva de imagens dá origem a uma outra linha de pesquisa e que não é o foco principal deste trabalho. Entretanto, uma vez obtida uma base de dados subjetivos, a utilização destes dados na estimação de qualidade através de parâmetros objetivos é uma atividade empírica relevante e com bons resultados. Com objetivos menos pretensiosos, a estimação empírica de qualidade permite esmiuçar e estender a base de dados.

O estudo desenvolvido neste trabalho aborda a estimação de qualidade subjetiva por comparação entre a imagem de entrada e a imagem decodificada incorporando segmentação de imagens como ferramenta de avaliação. O esquema utilizado está descrito na Figura 16. O correto registro das imagens de entrada e de saída é fundamental para o sucesso das medidas e esforço considerável é empenhado nos algoritmos de calibração da imagem.

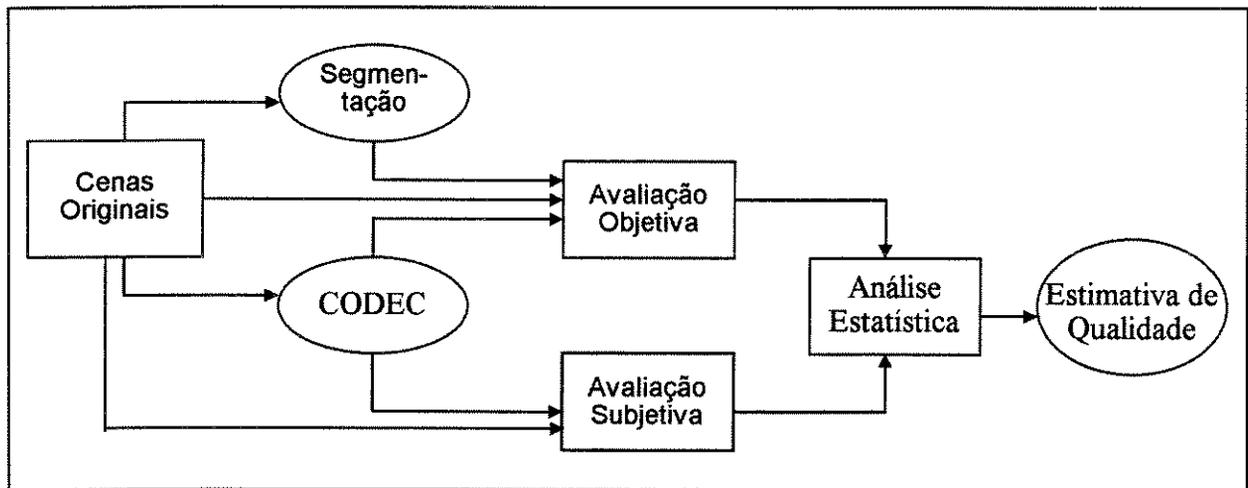


Figura 16: Algoritmo de estimação de qualidade subjetiva com base em cenas naturais

Esta metodologia é aplicada ao estudo de caso do próximo capítulo.

5.4 Segmentação como ferramenta de avaliação

Os equipamentos de redução de taxa de bits procuram casar a introdução de degradações com a região da imagem em que estas são menos perceptíveis. A comprovação da existência de diferenças de sensibilidade quanto à presença de degradações na imagem, motivou o estudo das imagens por regiões [53].

Entende-se por região, uma porção não necessariamente conectada de um campo ou quadro de vídeo com determinadas características espaciais e temporais que a diferenciam das demais regiões. A atividade de segmentação consiste na identificação destas, podendo ser tomada como uma ferramenta para análise de desempenho baseada em características particulares do sinal-fonte. O objetivo desta divisão em regiões com base no contexto visa

na imagem e a percepção da degradação. Esta divisão possibilita a utilização de um conjunto menor de estimadores, porém mais casados com a degradação presente na região.

O método de segmentação escolhido divide a imagem em regiões planas, regiões de borda e regiões de textura. Cada pixel da imagem é associado a uma das três regiões. Um exemplo para a cena “Mobile & Calendar” definida na Recomendação ITU-R BT. 802 [05] é apresentado na Figura 17. O nível de preto foi associado a regiões de textura, o nível cinza a regiões de borda e as regiões mais claras são identificadas como regiões planas da imagem.

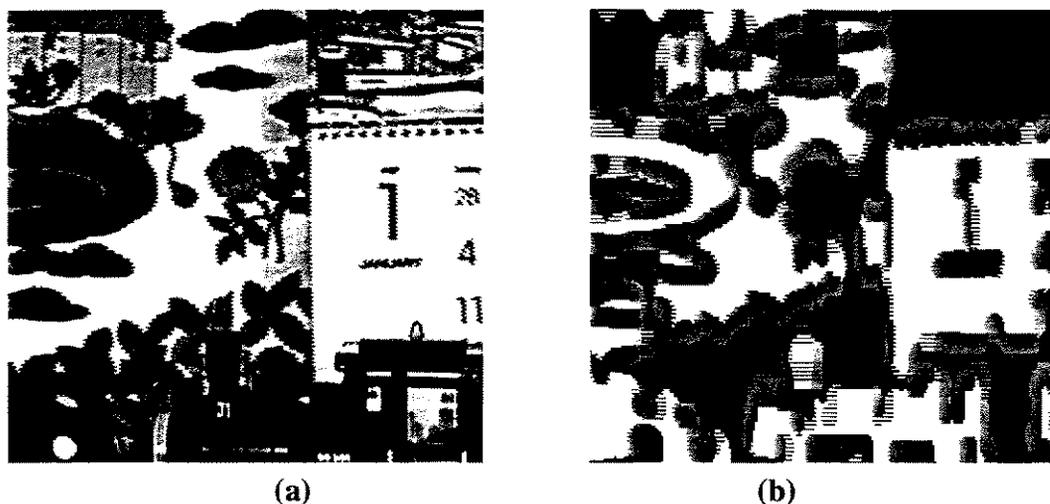


Figura 17: Ilustração do resultado de segmentação.

Cena “Mobile & Calendar”: (a) região da cena original (b) resultado da segmentação em três regiões.

Diversos são os algoritmos aplicáveis à segmentação. Uma possibilidade para a identificação de regiões na imagem é encontrar regiões planas através de medidas baseadas nas variações do gradiente. Computa-se a variância de um bloco centrado no pixel que se deseja classificar e se o valor obtido for inferior a um determinado limiar, o pixel é classificado como região plana. Aplica-se então uma filtragem mediana à imagem resultante. Regiões planas com área inferior a um determinado limiar mudam de classificação para que sejam eliminados conglomerados espúreos.

Emprega-se, em seguida, um algoritmo de detecção de borda. As bordas obtidas são dilatadas e se fizerem fronteira com áreas planas são classificadas como bordas. As demais regiões da imagem são classificadas como textura.

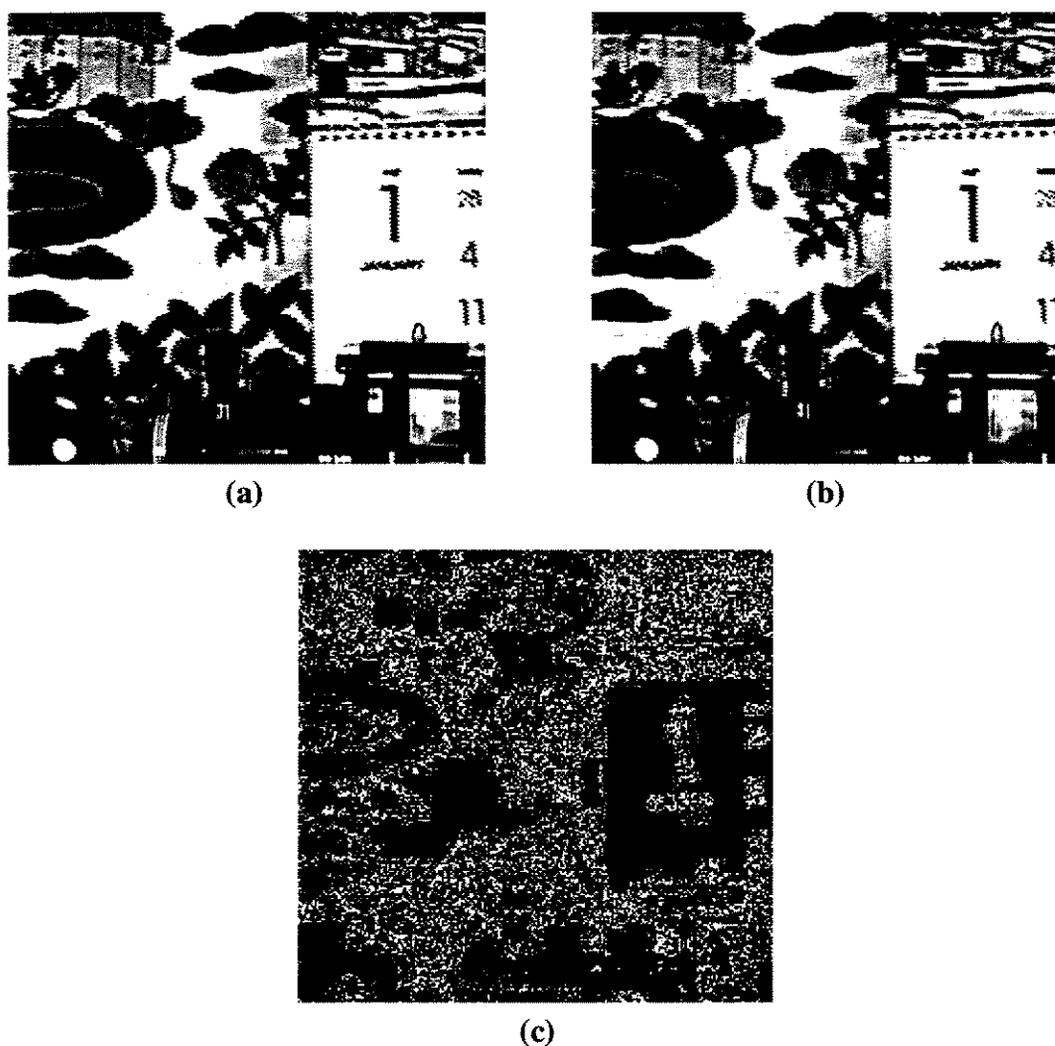
5.5 Seleção de estimadores objetivos

A total quantificação das características de desempenho dos sistemas de vídeo pressupõe a existência de parâmetros de desempenhos tais que cada parâmetro é sensível a uma única dimensão de qualidade do vídeo ou a uma dimensão de degradação. Esta propriedade discriminatória é útil na otimização de certos atributos em relação a outros e na obtenção de resultados que indiquem não somente que o sistema está falhando, mas também onde e como. Da mesma forma há interesse em saber como os avaliadores ponderam os diferentes atributos da

imagem, resolução, reprodução de cores, resolução temporal, acurácia, ao fornecerem sua nota a respeito da imagem.

5.5.1 Seleção de medidas objetivas

Grande esforço computacional é despendido na segmentação da imagem, realizada sobre as cenas originais apenas uma vez. Já no cômputo dos parâmetros são aplicadas operações de baixa complexidade, cujos cálculos são repetidos para os diversos sistemas avaliados. A escolha das operações mais apropriadas é atividade que exige conhecimento do problema a ser avaliado. A norma ANSI TI.801.03 dá grande ênfase a medidas da informação temporal da imagem. Entretanto, em codecs de alto desempenho, em que não há descarte de quadros, estas medidas têm baixa correlação com os resultados subjetivos.



**Figura 18: Caracterização do medidor “erro quadrático médio” (MSE)
Região da cena “Mobile & Calendar” (a) original (b) codificado a 2Mbit/s (c)MSE**

Com base nos resultados obtidos por Wolf [67] em trabalho de contribuição para o ITU foram selecionadas duas operações básicas: o erro quadrático médio e a filtragem Sobel, ambas

relacionadas ao conteúdo espacial da imagem. Estas medidas são extremamente dependentes do registro espacial e temporal das imagens de entrada e saída. O padrão ANSI, por outro lado, estipula um conjunto de parâmetros não dependentes de registro espacial, o que o torna mais abrangente dentro de seus propósitos.

5.5.1.1 Erro quadrático médio

O erro quadrático médio (MSE) é a medida mais comumente utilizada e em muitos casos, a única. Este estimador é o valor quadrático médio da diferença pixel-a-pixel entre o sinal de entrada e o sinal de saída. Sua ilustração está mostrada na Figura 18.

5.5.1.2 Filtragem Sobel

O valor Sobel é o resultado de uma filtragem passa-altas não linear. Esta filtragem é direcional, e indica a atividade local nas direções horizontal e vertical. Os filtros Sobel nas direções horizontal e vertical [44] estão descritos na Figura 19.

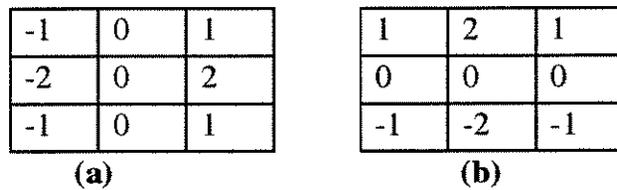


Figura 19: Descrição do filtro Sobel.
Filtro para realce de diferenças horizontais (a) e verticais (b).

A filtragem Sobel é empregada para ressaltar as diferenças horizontais e verticais entre os quadros que estão sendo comparados e, portanto, o interesse não está no valor calculado, em si, mas na diferença entre o valor Sobel do sinal de entrada e do sinal de saída. Os estimadores escolhidos são a diferença Sobel positiva e a negativa.

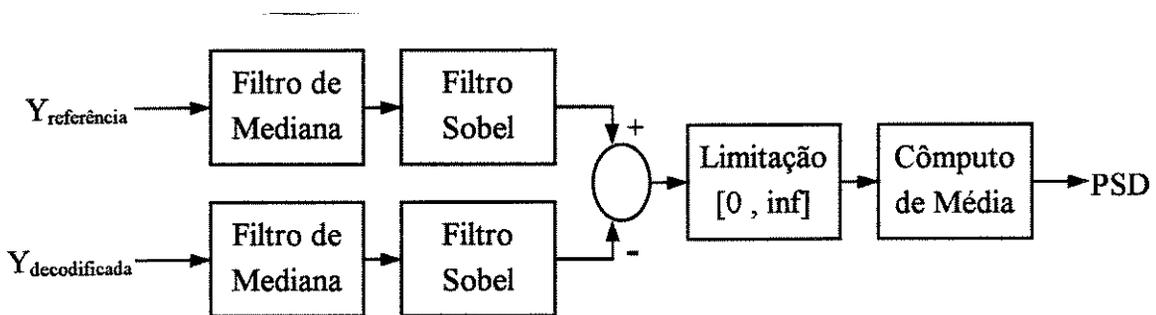


Figura 20: Determinação da Diferença Sobel Positiva - PSD.

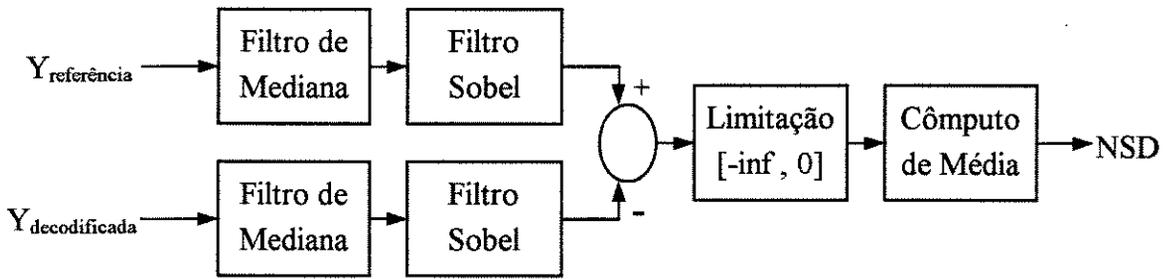


Figura 21: Determinação da Diferença Sobel Negativa - NSD.

A diferença Sobel positiva é a média da parte positiva da diferença pixel-a-pixel entre o valor Sobel de entrada e o valor Sobel do sinal de saída. Esta operação permite detectar o surgimento de falsas bordas, resultantes de blocagem na codificação. Analogamente, a diferença Sobel negativa é média da parte negativa. Esta medida detecta a suavização das bordas da imagem original, à perda de resolução e “blurring”. Estas operações estão apresentadas esquematicamente nas Figuras 20 e 21.

A Figura 22 caracteriza a filtragem Sobel tomando como exemplo o mesmo quadro da sequência “Mobile & Calendar” utilizado na exemplificação do erro quadrático médio.

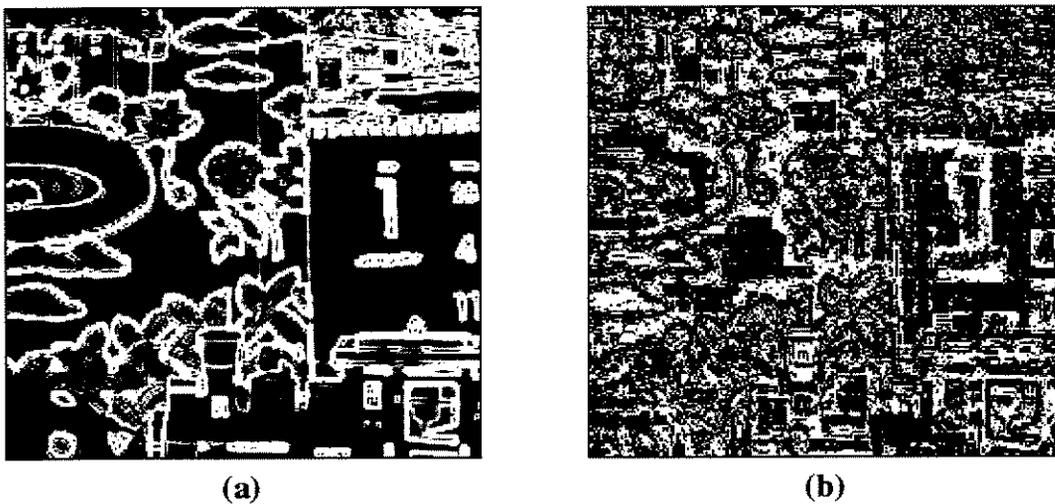


Figura 22: Caracterização da filtragem Sobel.

(a) Filtragem na cena original (b) Diferença entre a filtragem da cena original e codificada a 2Mbps/s

5.5.2 Correlação entre estimadores e degradações

O estimador define um conjunto de operações, possivelmente envolvendo filtragens, aplicadas sobre os sinais de entrada e de saída do sistema sob avaliação, em um determinada região. Os resultados obtidos pelos estimadores refletem o nível de degradação introduzido pelo sistema em teste sobre o material de vídeo de entrada.

Cada estimador está mais relacionado a um tipo de degradação como detalhado na Tabela 6.

Estimador	Região	Componente	Degradação Associada
MSE	plana	Luminância	ruído granular, blocagem
MSE	borda	Luminância	ruído tipo mosquito, perda de definição, efeito de Gibbs (“ringing”)
MSE	textura	Luminância	ruído, perda de definição, “ringing”
MSE	plana	Crominância	“cross color”, ruído, “blurring”
MSE	borda	Crominância	“cross color”, ruído, “blurring”
MSE	textura	Crominância	“cross color”, ruído, “blurring”
PSD	plana	Luminância	blocagem
PSD	borda	Luminância	“blurring”
PSD	textura	Luminância	“blurring”
NSD	plana	Luminância	suavização
NSD	borda	Luminância	“cross luminance”, “ringing”
NSD	textura	Luminância	“cross luminance”, “ringing”

Tabela 6: Associação entre medidor e degradação.

5.6 Introdução de funções logísticas como modelo para descrição de degradações

A recomendação ITU-R BT. 500 [02] define funções sigmoidais não simétricas para a descrição da relação entre medidas objetivas e subjetivas feitas sobre uma imagem. Estas funções procuram modelar os efeitos de saturação e assimetria da percepção próprios do sistema visual humano.

5.6.1 Efeito de saturação e assimetria da percepção

As medidas de distorção devem levar em conta, a faixa de degradações da imagem e a fidelidade das observações humanas. É comum assumir que a qualidade de imagem percebida pelos observadores é proporcional aquela estimada por medidas objetivas. Esta relação, entretanto, não é verdadeira, especialmente se a qualidade da imagem se encontra nos limites da percepção, imagens de alta fidelidade ou imagens muito degradadas. Nestas regiões há uma limitação na capacidade humana de observar alterações na qualidade da imagem. A forma encontrada para descrever este efeito é através de um função com comportamento linear na região central, mas com características de saturação nas bordas.

Outra característica relevante é a assimetria da percepção. Esta assimetria significa que tomando-se uma imagem e a ela associando um ganho e uma perda na qualidade equivalentes do ponto de vista objetivo, a perda de qualidade é, em termos subjetivos, mais significativa. Esta

qualidade da imagem, resulta em uma habilidade assimétrica por parte dos observadores em acompanhar variações de qualidade em testes subjetivos, e faz com que o comportamento na região central se afaste, ainda mais, de uma reta.

5.6.2 Modelo para descrição subjetiva de uma única degradação

Inicialmente deseja-se descrever como a variação de determinada dimensão de degradação se relaciona à qualidade subjetiva de imagens em televisão, admitindo que a degradação seja produzida pela variação de um único parâmetro físico. A questão principal desta discussão é estabelecer uma relação funcional entre o nível dos parâmetros e as notas subjetivas.

Prosser, Allnatt e Lewis [54] foram os primeiros a sugerir esta formulação através de uma função sigmoide com dois parâmetros conforme descrito pela equação (5.1). De fato, uma vez obtidos os resultados médios para cada situação em teste no experimento, a relação entre a magnitude da degradação e o efeito subjetivo resultante é, em geral, bem descrita por uma função logística.

$$d_{\text{est}} = \frac{1}{1 + \left(\frac{D}{D_M}\right)^G} \quad (5.1)$$

onde d_{est} é uma estimativa da nota subjetiva média no intervalo (5,1). A nota cinco se aplica a imagem perfeita ou sem degradações e nota um ao outro extremo. D é a magnitude da degradação;

Esta função é otimizada, tal que os parâmetros D_M e G são específicos para a dimensão de degradação investigada e devem ser estimados empiricamente. Os valores podem variar bastante de estudo para estudo e são fortemente dependentes do contexto, ou seja, da utilização de imagens de alta ou de baixa qualidade.

O parâmetro D_M corresponde ao valor de D para o qual d vale 3. Pelo fato de ser proporcional a declividade no ponto médio da sigmóide, este valor pode ser considerado como a taxa na qual os observadores mudam de opinião em resposta a mudanças de magnitude da degradação. O parâmetro G é simplesmente um parâmetro de crescimento e pode variar de degradação para degradação bem como variar de acordo com o range de estímulos. Considerações de ordem psicofísica a respeito da dependência de G com o tipo de degradação, entretanto, são puramente especulações.

5.6.3 Efeito subjetivo de múltiplas fontes de degradação

Assumindo que, o modelo de descrição do efeito subjetivo de uma degradação através de uma sigmóide é adequado, sua extensão para cenários em que mais de uma degradação está

presente, não é trivial. O efeito subjetivo da coexistência de diversas degradações é pouco conhecido. Os primeiros estudos, por Lewis et al. [43], na década de sessenta, sugeriam a existência de um tipo de aditividade entre as degradações. De fato, a validade deste modelo é comprovada para a combinação de degradações como eco e ruído e em outros estudos [60]. Estes resultados, entretanto, não atestam a validade da lei de aditividade para todas as degradações.

Lupker [48] mostrou que a aditividade é encontrada nos casos em que a degradação não tem efeito subjetivo similar, mas não vale quando dimensões com percepção semelhantes são combinadas. Assim embora esta proposta de aditividade tenha bom valor heurístico, sua capacidade de predição é limitada, pois o resultado subjetivo da combinação de efeitos semelhantes não pode ser predito assumindo-se a aditividade das degradações.

Algumas degradações apresentam um efeito de “superaditividade” que tem sérias implicações em telecomunicações, pois quando se assume a aditividade das degradações está se fazendo uma análise de melhor caso, e não há formas satisfatórias de se estabelecer quão ruim irá se tornar o serviço no caso em que houver uma combinação de degradações. Contrariamente, podem ser observados efeitos de “subaditividade”, menos problemáticos. Nos casos em que o efeito produzido pela combinação de degradações é inferior a soma destes efeitos separadamente, a análise feita com base na aditividade dos efeitos, assume um caráter de estimativa de pior caso.

O problema é não poder estabelecer a priori se duas degradações irão ou não se combinar de forma aditiva. Em princípio, a única maneira de se determinar se a aditividade vale ou não é realizar um teste completo, examinando todas as possibilidades. Não há formas seguras de se prever a validade do modelo de aditividade proposto e nem de que forma o modelo falha. Nos casos de interferência, exemplificados por Michelli, a aditividade falha nos casos em que, a dimensão de degradação combinada, são parâmetros que contribuem para a mesma classe de interferência e se verifica, nas situações em que as dimensões em estudo se relacionam a classes de interferência diferentes.

6. Estudo de caso

Este experimento teve como objetivo a caracterização do desempenho de sistemas de compressão de imagem para aplicações em televisão. Os experimentos realizados permitiram a obtenção de um banco de dados com resultados subjetivos representativos das degradações introduzidas em cenas naturais por codecs de vídeo e sua comparação à qualidade de estúdio alcançada por sistemas NTSC e PAL-M.

Nesta avaliação subjetiva foi utilizado o “Método da Escala de Degradação a Estímulo Duplo”, definido na Recomendação ITU-R BT.500[02]. Todos os aspectos referentes às condições de visualização foram cuidadosamente analisados e controlados. A Figura 23 apresentada a seguir ilustra o ambiente de avaliação utilizado.

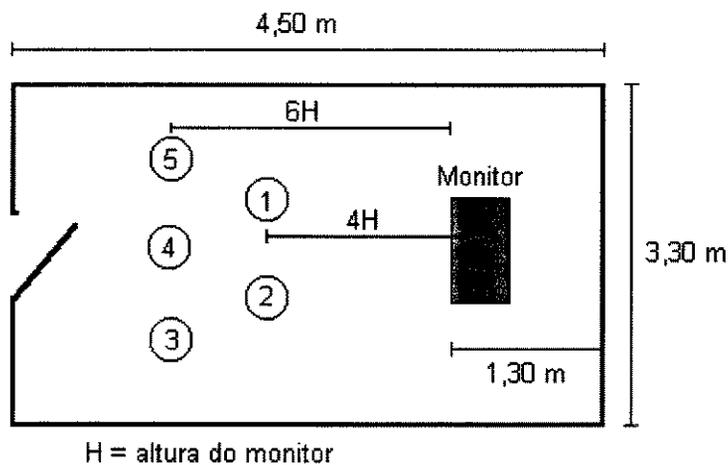


Figura 23: Sala de avaliação subjetiva

6.1 Material de teste

Esta etapa da avaliação subjetiva compreendeu a reprodução e organização do material de forma criteriosa e sistemática.

6.1.1 Cenas escolhidas

As cenas utilizadas no experimento são as mesmas cenas CLASSE 1 que foram utilizadas nas simulações de criticalidade de cena. Estas cenas fazem parte do conjunto de seqüências padronizadas pelo ITU-R BT.802 [05] para avaliações subjetivas de codecs e estão descritas em maiores detalhes na Tabela 1 (pag 39).

As cenas foram armazenadas em uma máquina D1 [15] no formato YCbCr4:2:2 padronizado pela Recomendação ITU-R BT.601 [03]. Esse material é composto por trechos de 300 quadros, ou seja, com dez segundos de duração.

6.1.2 Processamento

A Tabela 7 apresentada a seguir é uma descrição dos sistemas que foram analisados no experimento.

Descrição
Seqüência original não degradada
Subamostragem Espacial e Temporal
Codecs MPEG-2
Codificação MPEG-2 e degradação adicional
Sistema de Codificação com qualidade constante e taxa variável
Codecs para Sinal Composto

Tabela 7: Descrição das condições de degradação avaliadas.

6.1.2.1 Cenas originais

Processamento: Nenhum

As cenas originais, ou seja, os que não sofreram qualquer tipo de processamento, são tomadas como referência nas avaliações e foram incluídas entre as seqüências a serem avaliadas para verificar a coerência dos avaliadores e medir a confiabilidade do teste. A ausência de defeitos nestas cenas é fundamental para a obtenção de resultados estáveis, o que, entretanto, nem sempre foi possível obter em função de “flicker” ou da resolução espaço-temporal do sinal fonte.

6.1.2.2 Subamostragem temporal e espacial

Processamento:

Este item engloba dois tipos de subamostragem. No processo de subamostragem temporal é feito o descarte de um em cada dois quadros e a replicação do quadro anterior em substituição ao quadro descartado. No caso do processo de subamostragem espacial é feita uma redução da resolução por subamostragem de ordem dois e descarte de um dos campos. A seguir faz-se interpolação para retornar a escala original e também a interpolação dos campos vizinhos de paridade oposta para a reconstrução do campo descartado.

Este tipo de processamento introduz degradações bastante perceptíveis, principalmente “jerkines”, relacionado ao descarte de quadros ou de campos, e de “blurring”, relacionado à

redução de resolução. Estes processamentos são empregados em codecs MPEG-1 operando a baixas taxas (abaixo de 2 Mb/s) em aplicações como vídeo conferência, por exemplo.

6.1.2.3 Codificação MPEG-2

Processamento:

No caso da codificação MPEG-2 [07] foram testadas cenas naturais processadas por quatro codecs comerciais distintos, cedidos pela TV Globo, e pelo protótipo MPEG-2 em *software* desenvolvido no CPqD/TELEBRÁS com base no codec TM5 (“Test Model 5”) [14], modelo utilizado na padronização do sistema MPEG-2. Para efeito de comparação, o protótipo em “*software*” foi configurado com os mesmos parâmetros utilizados nos codecs comerciais.

Situações em teste:

As imagens são codificadas a taxas de bits constantes e com mecanismos de quantização adaptativa. Algumas características dos sistemas testados são dadas na Tabela 8. O parâmetro M, corresponde a distância entre “*frames*” cuja codificação é intra-campo ou codificação preditiva com base em campos anteriores. O GOP, ou “*Group of Pictures*”, fornece a distância entre dois quadros com codificação exclusivamente intra-campo.

Característica	Codec comercial 1	Codec comercial 2	Codec comercial 3	Codec comercial 4	Protótipo software CBR
Denominação	C1	C2	C3	C4	C5
Codificação	intra-quadro, progressiva e bidirecional com predição inter-campos e inter-quadros, progressiva e regressiva				
GOP ¹	12	12	12	2	12
M	1 e 2	1 e 2	2	1	1 e 2
Formato de codificação de croma	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:2	YCbCr4:2:0
Quantização	adaptativa	adaptativa	adaptativa	adaptativa	adaptativa
Taxa	constante 5 a 15 Mb/s	constante 5 a 15 Mb/s	constante 3,5 a 8,5 Mb/s	constante 18 Mb/s	constante 2,5 a 15 Mb/s

Tabela 8: Especificação dos parâmetros de codificação em teste.

As taxas utilizadas para o codec C3 são diferentes das demais devido a limitações de operação dos equipamentos, na configuração montada.

6.1.2.4 Codificação MPEG-2 e degradação adicional

Processamento:

O processamento consiste em introduzir degradações via *software* em um material fonte já codificado a 2,5 Mb/s. A simulação dos artefatos tem como referência os algoritmos

¹ GOP ou ‘*Group of Pictures*’ indica a distância entre quadros com codificação intra-quadro.

fornecidos pela recomendação ITU-T P. 930 [09].

Esta condição de degradação é utilizada para que se tenha referências de baixa qualidade, de forma a tornar o teste mais equilibrado e a garantir que todos os níveis da escala sejam utilizados. Dado o contexto do teste, os resultados obtidos para estas seqüências têm alta variabilidade e pouco interesse prático.

Situações em Teste

As degradações introduzidas estão descritas na Tabela 9.

Processamento	Efeito
filtragem passa-baixas	"blurring"
10% de blocos degradados	blocagem
1000 pixels degradados no quadro	chuveisco

Tabela 9: Descrição das degradações adicionadas por simulação.

6.1.2.5 Sistema de codificação com qualidade constante e taxa variável.

Processamento:

Trata-se de um processo de compressão, no qual o nível de quantização é mantido fixo e não há controle de taxa. Utiliza-se o protótipo MPEG-2 em *software* operando com codificação intra-quadro (como um codec JPEG [10]), taxa de bits variável e parâmetro de controle de quantização Q fixo. O parâmetro Q corresponde ao *quantiser-scale* do MPEG-2 [07].

O nível de degradação introduzido no material de vídeo codificado está, portanto, diretamente relacionado a granularidade da quantização. O parâmetro Q define esta quantização, tal que, quanto menor é o seu valor, mais refinada é a quantização.

Situações em Teste

A Tabela 10 apresenta as características gerais deste sistema, para as cinco configurações testadas.

Característica	Protótipo software VBR				
Denominação	S1	S2	S3	S4	S5
Codificação	intra-quadro	intra-quadro	intra-quadro	intra-quadro	intra-quadro
GOP	1	1	1	1	1
Formato de codificação de croma	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:0	YCbCr4:2:0
Quantização	fixa, $Q=4$	fixa, $Q=8$	fixa, $Q=16$	fixa, $Q=32$	fixa, $Q=62$
Taxa (média sobre o conjunto de cenas utilizadas)	33,2 Mbits/s	21,0 Mbits/s	13,0 Mbits/s	7,9 Mbits/s	5,0 Mbits/s

Tabela 10: Configuração do sistema de codificação com qualidade constante.

6.1.2.6 Codecs para sinal composto

Processamento:

A título de comparação foram incluídas no teste cenas que sofreram conversão para sinal composto. O sinal original, no formato 4:2:2 é convertido para NTSC ou PAL-M, e em seguida decodificado.

Situações em teste:

Foram utilizados os protótipos em software de codificação/decodificação para NTSC/PAL-M, desenvolvidos pelo CPqD, operando no modo 2-D não-adaptativo e um equipamento de conversão disponível, descrito na Tabela 11 a seguir.

Equipamento	Configuração
Encoder	2D-Padrão
Decoder	2D-Adaptativo

Tabela 11: Especificações do encoder e decoder.

6.1.3 Preparação do material

6.1.3.1 Arranjo das Seqüências

A material preparado atendeu às condições de máxima aleatoriedade explorando todas as condições de teste e observando a limitação de trinta minutos, necessária para que se consiga manter a atenção dos avaliadores. É dada maior importância, relativamente ao número de apresentações, ao material processado pelos codecs comerciais e ao codec protótipo em software do que aos demais sistemas.

Os tipos de processamento apresentados totalizam 41 condições a serem exploradas. Entretanto, cada sessão de avaliação continha um total de 48 unidades de apresentação quantidade que se mostrou suficiente para medir o ruído associado e a variabilidade das decisões subjetivas.

6.1.3.2 Ordenação

Foram preparadas cinco apresentações que diferem entre si na escolha das seqüências que serão utilizadas em cada uma das situações de teste descritas e na ordem em que estes itens são apresentados. O conjunto das cinco apresentações é tal que cobre todas as 205 possibilidades (41 situações em teste e 5 cenas). A ordenação das cenas é diferente para cada apresentação a fim de compensar efeitos de adaptação ou de cansaço que possam vir a interferir no julgamento dos avaliadores.

As seqüências são arranjadas de forma pseudo-aleatória e cuidados foram tomados na ordenação para a evitar a exibição seqüencial de imagens perceptualmente '*semelhantes*'. A semelhança entre as unidades de apresentação é medida em termos da cena e do tipo de processamento. Assim, tomou-se o cuidado para que a mesma cena não fosse apresentada duas vezes seguidas ainda que sob condições de degradação diferentes. Da mesma forma, duas

restrição busca compensar, ao longo do teste, os efeitos de adaptação de curto prazo.

Na definição de semelhança entre as diversas categorias de processamento foram definidos grupos onde há semelhança em termos de qualidade objetiva nas seqüências processadas. Desta forma, os itens pertencentes ao mesmo grupo não são exibidos duas vezes em seguida. Os grupos de processamento estão definidos na Tabela 12 a seguir.

Grupo	Processamentos “Semelhantes”
I	Original, S1 e codecs MPEG-2 de 12,5 a 18 Mb/s
II	S2, S3 e codecs MPEG-2 de 7,5 a 10 Mb/s
III	codecs MPEG-2 a 5 Mb/s
IV	codecs MPEG-2 a 3,5 Mb/s
V	codecs MPEG-2 a 2,5 Mb/s
VI	codecs MPEG-2 a 2,5 Mb/s e degradação adicional
VII	S4 e S5
VIII	NTSC e PAL
IX	Resolução espacial e temporal

Tabela 12: Descrição dos grupos de proximidade no experimento.

6.1.4 Forma de apresentação

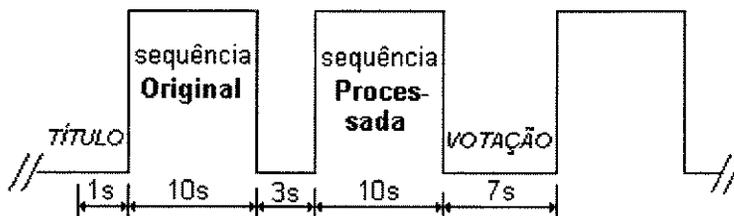


Figura 24: Forma de apresentação das seqüências.

6.2 Detalhes do teste

6.2.1 Avaliadores

Os testes envolveram um total de 56 pessoas dentre funcionários do CPqD e da Rede Globo, professores e alunos da UNICAMP. Os avaliadores foram divididos em dois grupos, especialistas e não especialistas, resultando num total de 14 especialistas e de 42 não especialistas.

O critério para a separação destes grupos foi a experiência das pessoas na avaliação de

cenar de televisão, participação em atividades de análise da qualidade de imagens como parte do trabalho diário e conhecimento do tipo de degradação introduzida por codecs de vídeo.

Avaliadores	CPqD	Rede Globo	Unicamp
especialistas	2	11	1
não especialistas	32	-	10
total	34	11	11

Tabela 13: Descrição dos avaliadores do teste subjetivo

Cada avaliador participou de duas sessões de avaliação nas quais são modificadas a ordem de apresentação das seqüências, as situações de teste e a distância ao monitor. Em uma das apresentações o avaliador fica a uma distância de 4H do monitor, e na outra a 6H, onde H é a altura do monitor (conforme Figura 23).

As sessões de avaliação produziram um total de aproximadamente 5300 notas (96 notas para cada avaliador) para serem compiladas.

6.2.2 Votação

A duração de uma sessão de avaliação é um compromisso entre a necessidade de se ter redundância explorando diferentes parâmetros e o tempo em que é possível manter concentrada a atenção do avaliador. No teste realizado, cada sessão de avaliação foi dividida em duas partes com duração média de 14 minutos cada e com um rápido intervalo entre elas para descansar o avaliador.

Ao início de cada sessão, os avaliadores foram apresentados ao método de avaliação e às degradações que aparecem no teste, através de seqüências de treinamento. Entretanto, as notas dadas às seqüências de treinamento não foram consideradas nos resultados finais. As instruções incluíram informações sobre o tipo de avaliação, escala utilizada, temporização das seqüências e duração do intervalo para votação.

O julgamento das seqüências foi colhido através de uma escala de degradação por categorias com nove pontos, formada pelos cinco níveis de degradação estabelecidos pelo ITU-R [02] e pelos quatro níveis intermediários de degradação. Notas de 1 a 5, igualmente espaçadas, foram relacionados à escala adotada como mostrado na Figura 25.

Categoria	Posição na Escala	Nota
imperceptível	<input type="checkbox"/>	5,0
	<input type="checkbox"/>	4,5
perceptível, mas não incomoda	<input type="checkbox"/>	4,0
	<input type="checkbox"/>	3,5
incomoda pouco	<input type="checkbox"/>	3,0
	<input type="checkbox"/>	2,5
incomoda	<input type="checkbox"/>	2,0
	<input type="checkbox"/>	1,5
incomoda muito	<input type="checkbox"/>	1,0

Figura 25: Relação entre a escala de degradação e o valor numérico das notas.

6.3 Resultados e conclusões

6.3.1 Médias gerais

6.3.1.1 Descarte de observadores e de notas

A primeira etapa de compilação das notas obtidas no teste consistiu em calcular para cada cena o valor médio e o desvio padrão associado, considerando as notas de todos os avaliadores.

Inicialmente foi feita uma análise para avaliação da coerência dos avaliadores, segundo o qual nenhum avaliador foi desconsiderado. A seguir foram aplicados os critérios de descarte de notas conforme descrito no Capítulo 4. A aplicação deste procedimento implicou no descarte de um total de 121 notas, sendo percentualmente mais significativo no caso de avaliadores especialistas, conforme mostrado na Tabela 14:

Avaliadores	Número de Notas Descartadas	Percentual de Descarte (%)
especialistas	35	2,52
não especialistas	86	2,11

Tabela 14: Discriminação do percentual de descarte de notas.

6.3.1.2 Ajuste de coerência por tipo de processamento

Devido às restrições relativas ao tempo de duração máximo das sessões e ao número mínimo de cenas, não foi possível garantir que todas as situações sob teste fossem julgadas por todos os avaliadores. Portanto, que existe uma forte dependência do resultado de um par (cena, sistema sob teste) com relação ao contexto da avaliação (apresentação, conjunto de avaliadores).

A polarização dos julgamentos em função do contexto é fonte de discrepâncias no quadro geral de notas. Esta constatação levou à adoção de procedimentos de suavização das notas, pela imposição de restrições de desempenho em cada um dos subsistemas. Os critérios foram determinados pelo conhecimento prévio dos sistemas, baseado nos resultados de avaliação objetiva e em características construtivas dos equipamentos, no caso dos codificadores para sinal composto.

critério

Minimização do erro quadrático médio entre o MOS (*Mean Opinion Score*) calculado inicialmente e o MOS obtidos após a suavização das notas, para cada uma das cenas.

restrições

Sistema	Restrição Imposta a Mesma Cena Natural
Sistema de codificação com qualidade constante	<ul style="list-style-type: none"> A nota média de um sistema com determinado nível de quantização não poderá ser superior a nota média obtida por um sistema com um nível de quantização inferior, ou seja; $MOS_{Q=4} \geq MOS_{Q=8} \geq MOS_{Q=16} \geq \dots$
Codecs MPEG-2	<ul style="list-style-type: none"> A <u>mínima</u> nota média dada para os codecs operando a uma determinada taxa não poderá ser inferior a <u>máxima</u> nota média dada em uma taxa inferior O desempenho de um codec operando com o parâmetro $M=2$ não poderá ser inferior ao desempenho do mesmo codec, a mesma taxa e com $M=1$
Codecs para sinal composto	<ul style="list-style-type: none"> O desempenho da codificação PAL não pode ser superior ao desempenho da codificação NTSC. Esta comparação deve ser feita entre equipamentos e entre o protótipo em “<i>software</i>” separadamente.

MÉDIAS	MAE médio	MAE/MOS (%)
Gerais	0,05	1,26
Especialistas	0,12	3,09
Não Especialistas	0,06	1,56
Distância=4H	0,08	2,12
Distância=6H	0,06	1,58

Tabela 15: Erro absoluto médio resultante do ajuste de notas.

Este procedimento foi aplicado às médias gerais, bem como a cada um dos subgrupos entre os quais observou-se significativas variações entre as notas. A aplicação destes critérios de otimização implicou em variação média no MOS de 1,92%. A Tabela 15 detalha, para cada

MOS médio, após ajustes.

As médias finais obtidas são apresentadas no Anexo 9.2.

6.3.2 Diferenças perceptuais entre cenas e entre grupos

A etapa seguinte na análise dos resultados consistiu em identificar os principais fatores de variabilidade nas notas para uma mesma situação de processamento e analisar seu impacto no resultado geral.

As cenas escolhidas para o teste têm características espectrais e conteúdos diferentes de forma que a percepção das degradações também é diferente. As cenas em que a percepção da degradação é mais difícil são 'Flower Garden'(FL) e 'Diva with Noise'(DI). Em média, as notas destas cenas foram 11% superiores às notas da seqüência 'Mobile & Calendar'(MO), cena na qual os avaliadores foram mais críticos. Os valores médios, considerando o conjunto de sistemas sob teste, são mostrados na Figura 26.

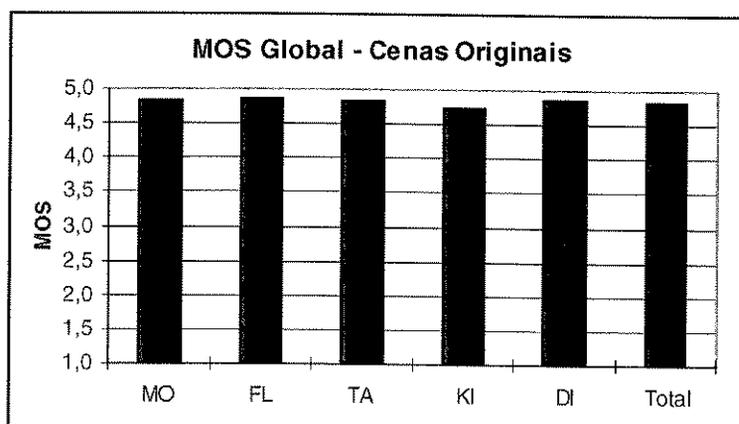


Figura 26: Notas médias obtidas por cada seqüência.

Explorando estas diferenças entre as cenas, as notas médias para cada situação em teste foi normalizada de acordo com a expressão abaixo. Esta normalização vem ao encontro da percepção de que avaliadores humanos deram as notas com base nas diferenças entre as cenas e permite que as notas sejam interpretadas como um percentual de degradação em relação à cena original.

$$d_n = \frac{MOS_{original} - MOS_n}{MOS_{original} - 1} \times 100\% \quad (6.1)$$

Além das diferenças observadas entre cenas há também diferenças relacionadas à distância do observador ao monitor e a sua experiência em avaliação de imagens, ou seja, se ele é um especialista ou não.

Avaliadores	Média	Avaliadores	Média
especialistas	3,80	distância=4H	3,98
não especialistas	4,09	distância=6H	4,08
<i>diferença percentual média</i>	7,1%	<i>diferença perceptual média</i>	2,5%

Tabela 16: Diferenças percentuais entre notas por grupos.

A Tabela 16 permite inferir, que, as diferenças com relação à distância, para os casos considerados, são menos significativas do que as observadas entre os grupos de especialistas e não especialistas.

As notas dos avaliadores especialistas, mais críticos e mais treinados, é em geral, inferior a nota dada pelo não especialistas. Com relação à distância, a percepção das degradações é mais aguçada quando se está mais perto do monitor. Observou-se também, que a variação das notas entre os especialistas com relação à cena é ligeiramente superior à variação presente no grupo de não especialistas.

Para análise de notas, somente entre os diversos grupos de avaliadores, é feita uma normalização das notas que leva em conta, não a média geral das cenas originais, mas sim a média do grupo considerado.

6.3.3 Análise de resultados dos subsistemas em teste

6.3.3.1 Codec C4 (YCbCr422 18 Mbits/s)

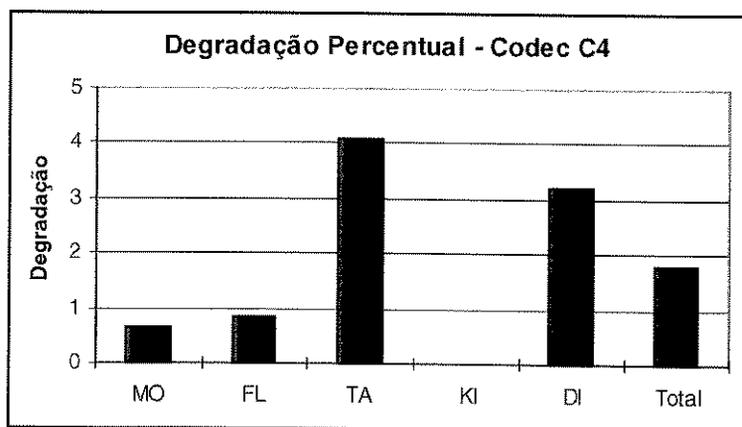


Figura 27: Resultado subjetivo do codec C4 a 18 Mbits/s - Média geral.

Este codec tem ótimo desempenho subjetivo, conforme mostrado na Figura 27. O desempenho obtido é comparável ao obtido pelas seqüências originais, com relativa homogeneidade de notas entre cenas. A diferença entre a melhor e a pior nota é inferior a 4,1%. A última coluna do gráfico mostra que a diferença percentual entre as seqüências originais e as cenas codificadas a 18Mbit/s é de apenas 1,77%. Estes resultados mostram que, internamente ao codificador, a utilização de codificação YCbCr4:2:2 ao invés de YCbCr4:2:0, tem relevante efeito subjetivo positivo.

6.3.3.2 Subamostragem espacial e temporal

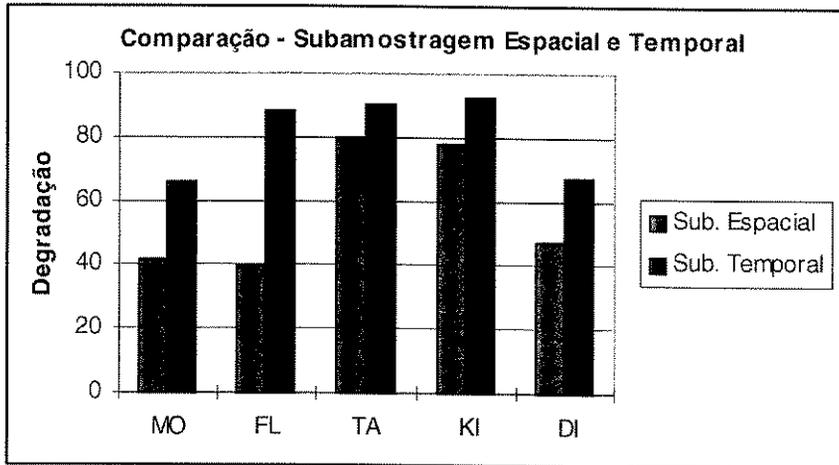


Figura 28: Resultado subjetivo da subamostragem espacial e temporal - Médias por cena.

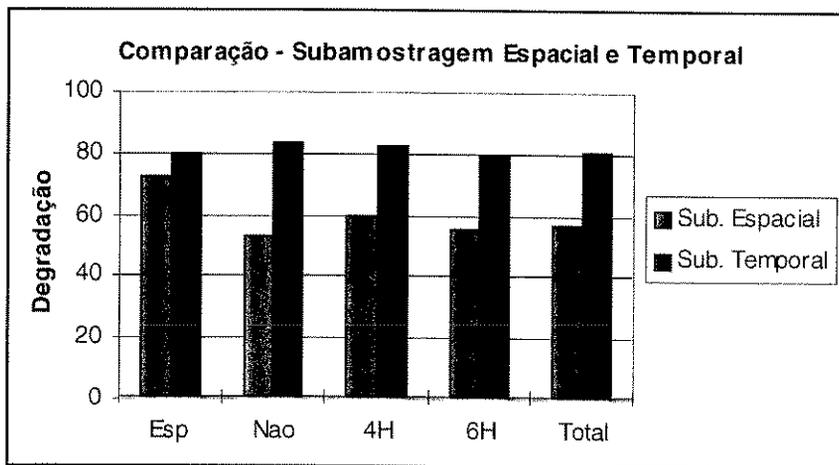


Figura 29: Resultado subjetivo da subamostragem espacial e temporal - Médias por grupo.

Os resultados apresentados evidenciam que estes dois tipos de processamento têm efeito subjetivo bastante desagradável. Ao contrário do que foi observado na seção 6.3.1, o efeito subjetivo é extremamente dependente da cena, e, em especial, da quantidade de movimento presente na seqüência. A variação em relação ao valor médio chega a 60% para subamostragem temporal e 54% no caso de subamostragem espacial.

6.3.3.3 Sistema de codificação com qualidade constante

6.3.3.3.1 Relação MOS x Nível de quantização

A diferença de desempenho entre os níveis consecutivos considerados, em termos de relação sinal-ruído, é de 4dB. Entretanto, estes níveis permitiram estabelecer, em média, apenas quatro níveis distintos na escala de degradação subjetiva. Considerando todas as notas, a diferença percebida entre os níveis 4 e 8 é de apenas 0,18, o que representa uma diferença de 4,5% do intervalo de notas.

Considerando os resultados médios apresentados por cena, (Figura 30), verifica-se que

há significativas diferenças entre as cenas. Na cena 'Mobile' a diferença de notas entre os níveis de quantização 4 e 8 é significativa (cerca de 10,1%) enquanto que para as sequências 'Kiel' e 'Diva' não há diferença. De fato, as características da cena explicam estas diferenças na sensibilidade do avaliador em identificar as degradações. Na 'Diva', o ruído presente em volta da cena mascara o efeito da quantização, no caso de 'Kiel' o mascaramento está relacionado à imperfeições presentes na imagem original e que dificultam a identificação das degradações com relação a imagem de referência.

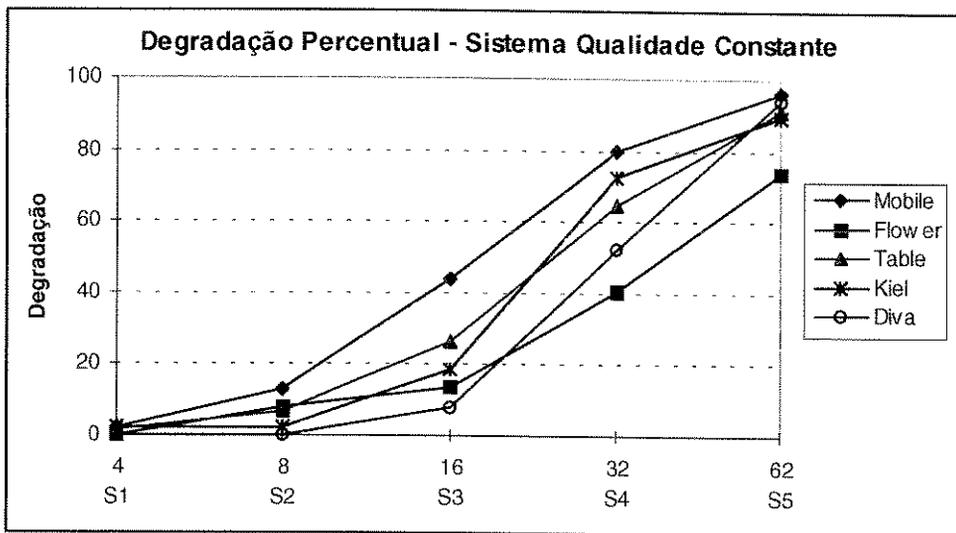


Figura 30: Desempenho subjetivo do sistema de qualidade constante - Média por cena

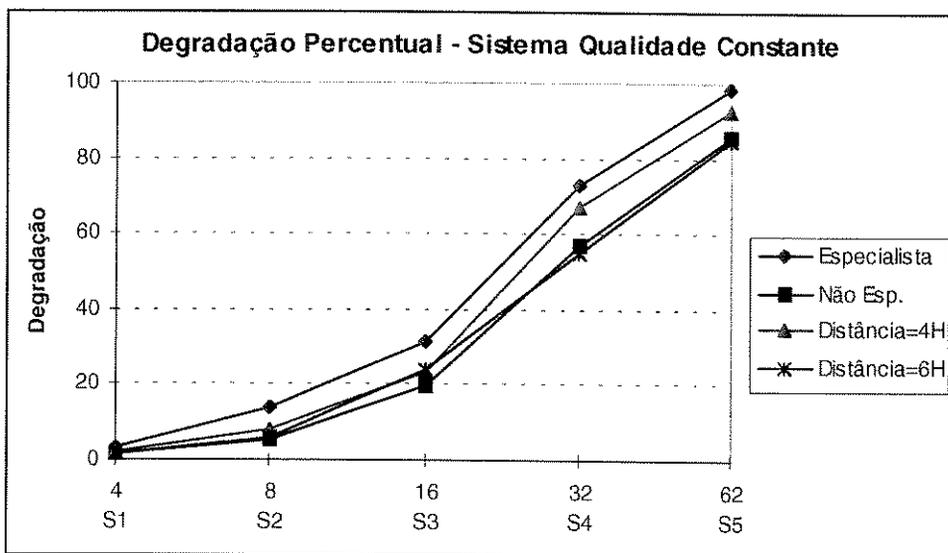


Figura 31: Desempenho subjetivo do sistema de qualidade constante - Média por grupo

Na cena 'Flower' pode-se observar o mascaramento relacionado aos níveis de degradação mais elevados. Se tomarmos apenas as notas que foram dadas quando os avaliadores estão mais distantes do monitor, esta cena recebeu nota 2,5 no nível de degradação mais elevado, 75% a mais do que a média geral para este nível.

6.3.3.4 Codecs MPEG-2

6.3.3.4.1 Relação MOS x Taxa

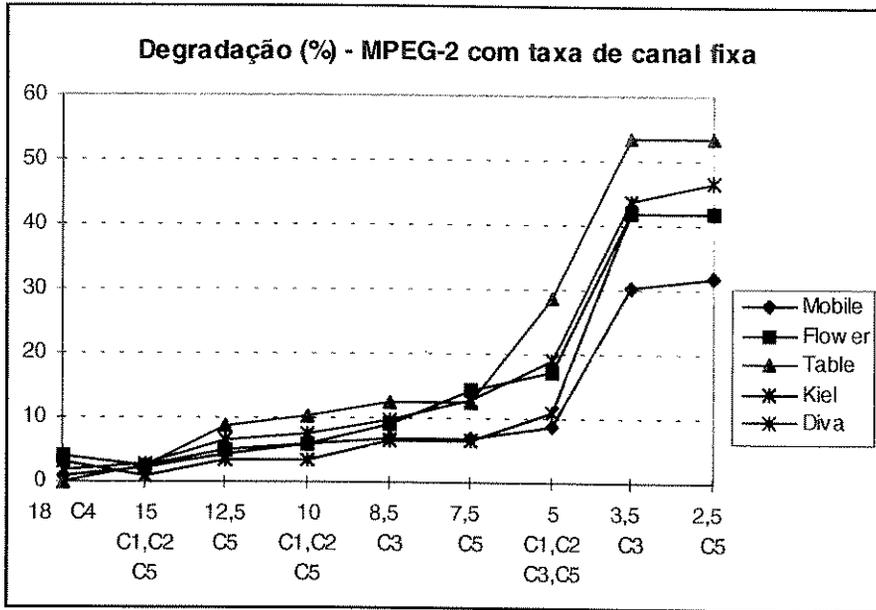


Figura 32: Desempenho subjetivo de codecs MPEG-2 - Média por cena.

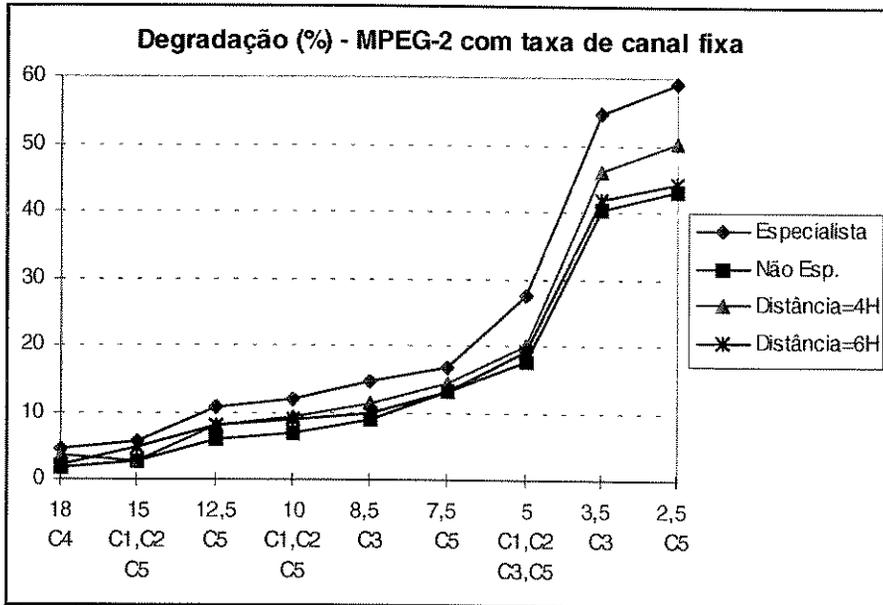


Figura 33: Desempenho subjetivo de codecs MPEG-2 - Média por grupo.

Para os codecs MPEG-2 é feita uma análise por cena, semelhante a que foi feita na seção anterior, ou seja, uma descrição do MOS calculado por mediação entre os diversos codecs em função da taxa. Como o interesse é avaliar o desempenho em relação à taxa, a mediação é feita tomando apenas as notas dadas para M=2. A influência deste parâmetro e as diferenças de desempenho entre codecs são analisadas separadamente.

Ao contrário do sistema de qualidade constante, como há restrição de taxa, há oscilação

de qualidade de imagem ao longo da cena. É uma característica de julgamento subjetivo que os avaliadores fiquem atentos às diversas nuances da codificação durante o tempo de observação da seqüência e dêem notas para o pior caso. Estas características de análise subjetiva explicam o aumento das variações das notas entre os diversos avaliadores verificada nos codecs MPEG-2 com relação ao sistema de qualidade constante.

A Figura 34 evidencia que quanto maior é a restrição imposta em termos de taxa, maior é a diferença de desempenho em relação ao conteúdo da cena. Em 15 Mb/s a diferença entre as cenas com relação ao MOS médio para esta taxa é de apenas 10,3%, cresce para 15,4% em 7,5 Mb/s e atinge o percentual de espalhamento máximo em 2,5 Mb/s com cerca de 51,3% em relação ao valor médio.

Tomando como referência o MOS médio obtido pelas seqüências originais ($MOS_{ORIGINAL}=4,82$), em relação aos valores apresentados na Figura 33, pode-se afirmar que as degradações somente são perceptíveis para taxas abaixo de 7,5 Mb/s no caso de especialistas e em taxas inferiores a 5 Mb/s no caso de avaliadores não especialistas.

6.3.3.4.2 Influência do parâmetro M

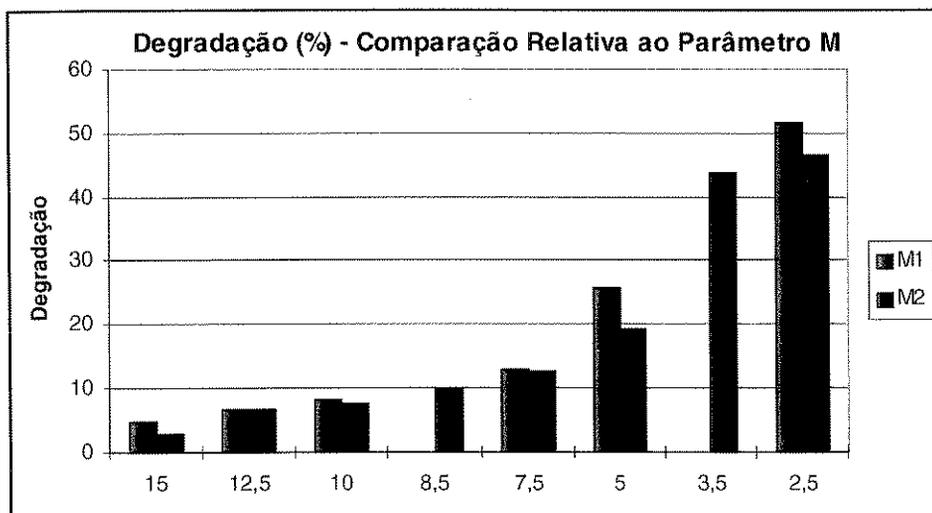


Figura 34: Comparação entre o desempenho com relação ao parâmetro M.

São mostradas na Figura 34 os resultados médios entre os diversos codecs obtidos com $M=1$ e $M=2$, separadamente. Nas taxas 8,5 e 3,5 Mb/s, somente foi analisado o desempenho do codec C3, que não pode ser configurado para operar com $M=1$.

A influência deste parâmetro é significativa somente em taxas inferiores a 7,5 Mb/s, região na qual, em função da restrição de taxa, qualquer ganho de desempenho objetivo se torna subjetivamente significativo.

6.3.3.4.3 Diferenças de desempenho entre codecs

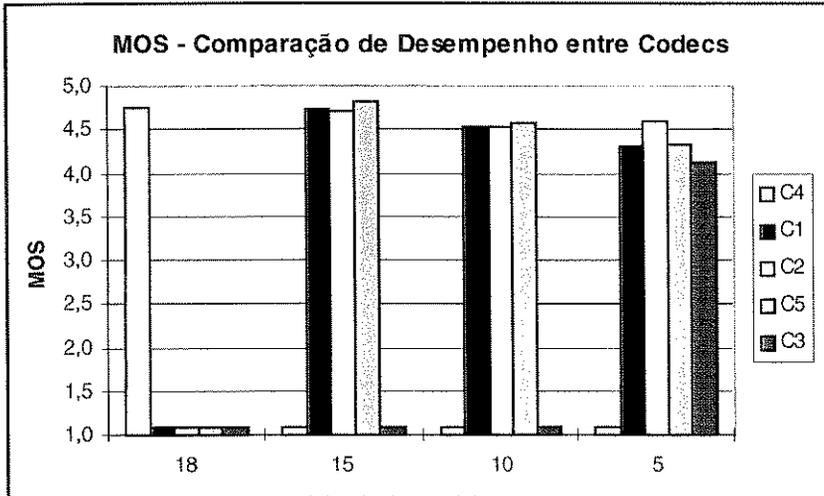


Figura 35: Comparação de desempenho subjetivo entre codecs MPEG-2, com M=2.

Os resultados obtidos na avaliação subjetiva definem características gerais do comportamento dos codecs com relação à taxa. Claramente não são apropriados para a diferenciação de sistemas com desempenho subjetivo semelhante. A Figura 35 apresenta os resultados médios de C4 com M=1, e dos demais codecs com M=2. Observe, entretanto, que nem todos os codecs foram avaliados em todas as taxas.

Testes subjetivos projetados com o objetivo de comparar e classificar equipamentos de diferentes fabricantes, por exemplo, devem ser projetados de forma que o avaliador possa compará-los e estabelecer relações diretamente. Dados comparativos, com elevado grau de sensibilidade e poder discriminatório, não podem ser inferidos com base em julgamentos feitos tomando sempre a mesma imagem como referência, como no caso do teste realizado. A comparação pura e simples entre os sistemas, entretanto, tem a limitação de que as notas obtidas passam a expressar as relações entre os equipamentos testados e não mais valores de degradação.

6.3.3.5 Codecs para sinal composto

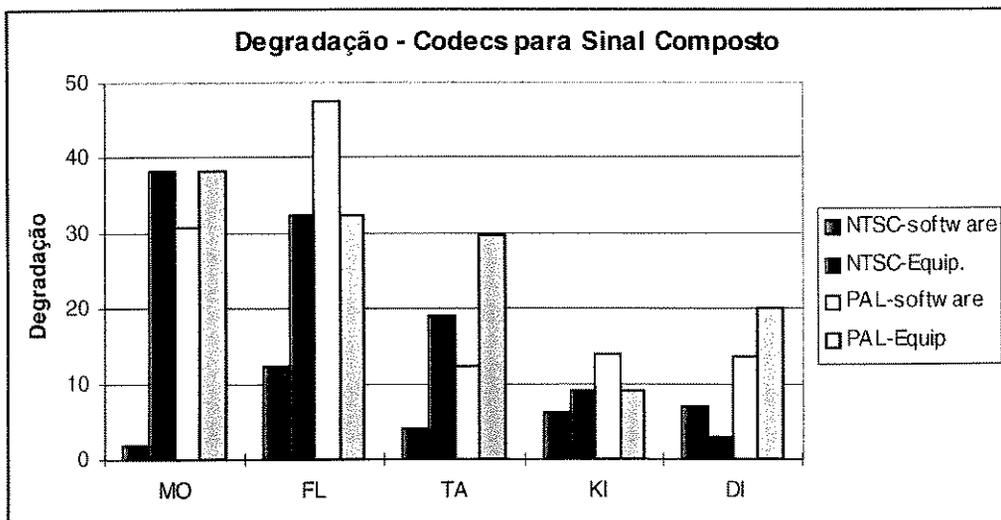


Figura 36: Desempenho Subjetivo dos Codecs - Média por Cena.

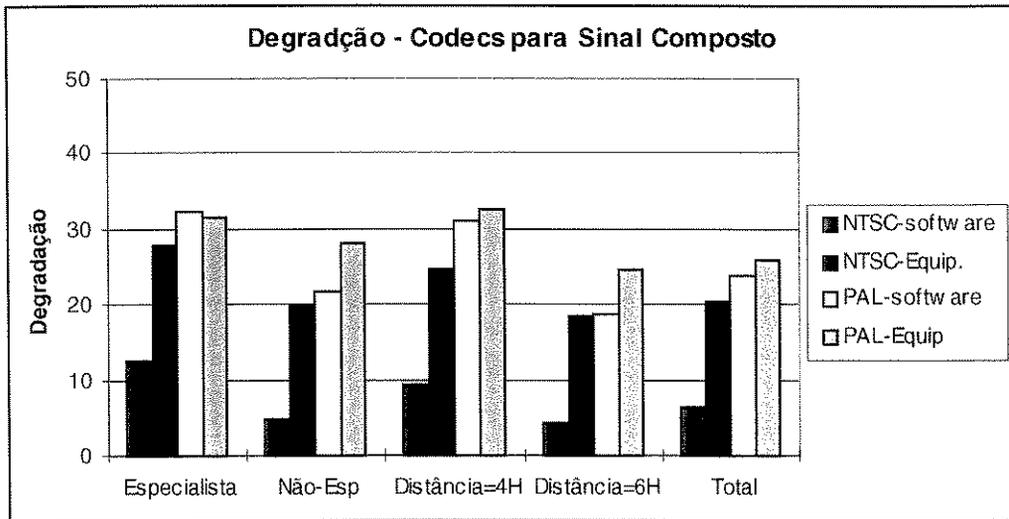


Figura 37: Desempenho Subjetivo dos Codecs - Média por Grupo.

O desempenho subjetivo médio dos pares codificador e decodificador para sinal composto é mostrado nas Figuras 36 e 37. A diferença de desempenho entre o protótipo em “software” e o equipamento só é significativa no caso NTSC. De fato, o protótipo tem desempenho 17% superior à média dos demais conversores. Para a seqüência ‘Diva’, entretanto, o desempenho do protótipo é inferior ao do equipamento comercial.

A equivalência, em termos de desempenho, entre o protótipo e o equipamento utilizado, observada no caso PAL-M, está relacionada às limitações no projeto dos filtros do protótipo em “software”. A conversão PAL apresentou acentuada degradação nos primeiros quadros da cena ‘Table’, entretanto, a pior média foi observada para a cena ‘Flower’, cujas degradações se encontram distribuídas por todo o trecho de dez segundos da seqüência.

6.3.3.6 Comparação entre NTSC e codecs de vídeo

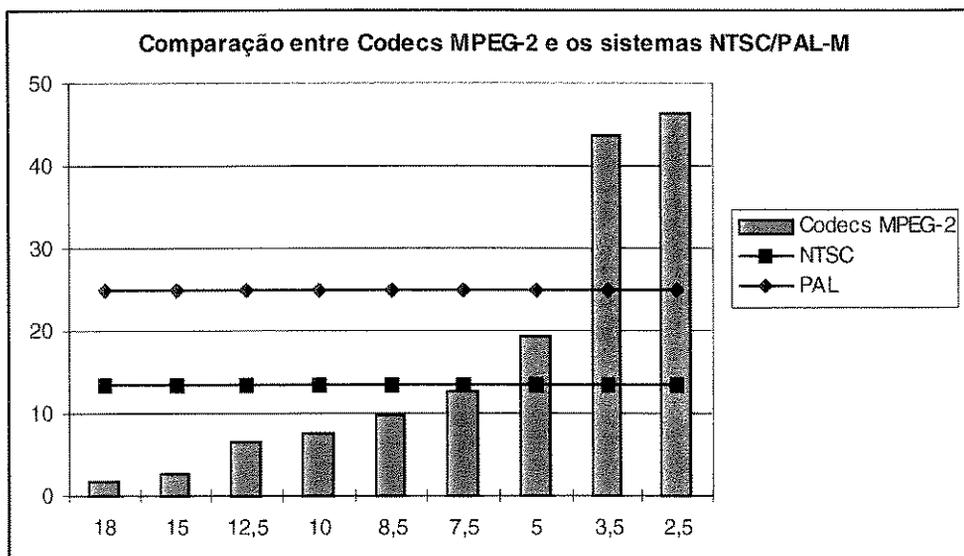


Figura 38: Comparação entre o desempenho subjetivo de codecs MPEG-2 e NTSC/PAL-M

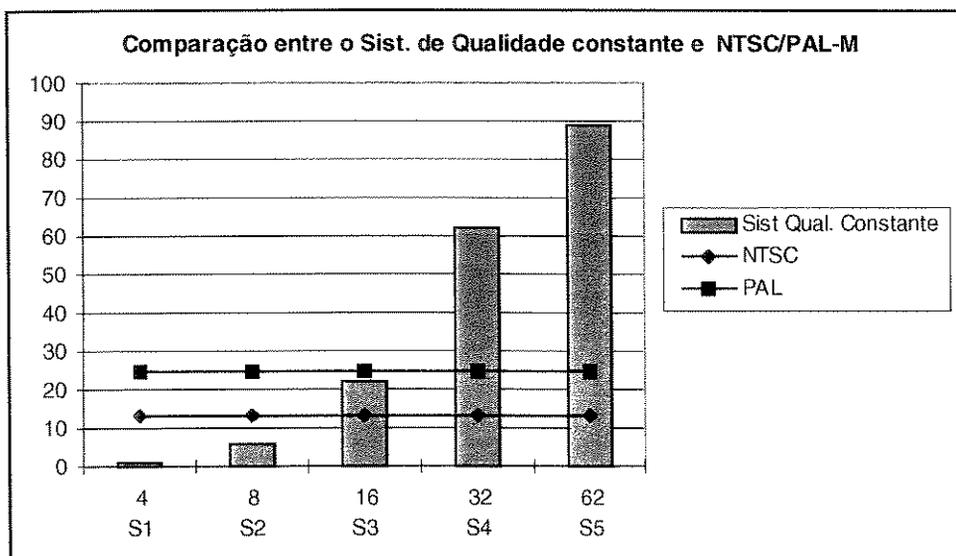


Figura 39: Comparação entre o desempenho subjetivo do sistema de qualidade constante e NTSC/PAL-M

As Figuras 38 e 39 contêm resultados comparativos entre os sistemas de compressão digital e sistemas NTSC/PAL-M com qualidade estúdio. Estes resultados são bastante relevantes dado que os sistemas comparados têm metodologias de avaliação objetiva bastantes distintas, inviabilizando a comparação direta.

Verifica-se que os codecs têm um desempenho superior ou equivalente aos sistemas onde há conversão NTSC em taxas acima de 7,5 Mbits/s e acima de 5 Mbit/s, no caso de conversão PAL-M. Abaixo destas taxas, as degradações introduzidas pelos algoritmos de compressão de dados (blocagem, “blurring”, “ringing”, etc) se tornam subjetivamente mais significativas do que as introduzidas por sistemas de conversão para sinais compostos (“cross-color”, “cross-luminance”, etc).

Da mesma forma, o sistema NTSC passa a ter desempenho superior ao sistema de qualidade constante para $Q > 8$ e, no caso do sistema PAL-M, para $Q > 16$, região de visibilidade das degradações introduzidas pela codificação digital.

6.4 Aplicação dos Resultados na Estimação de Qualidade Subjetiva através de Parâmetros Objetivos

Esta seção descreve a estimação de qualidade subjetiva conforme esquema da Figura 16 e com base no banco de dados subjetivos obtido. A base de dados subjetivos escolhidos corresponde as médias gerais, que envolvem todos os avaliadores.

6.4.1 Cálculo dos Parâmetros Objetivos

Amostras de luminância e de cromaticidade dos sinais de entrada e de saída após o registro espacial e temporal e da correção de ganho e “offset” são utilizadas para cálculo dos parâmetros.

Os parâmetros são o resultado da média em 12 quadros selecionados, da porção de regime da cena, da aplicação do estimador em uma das três regiões. O número 12 é escolhido por ser um múltiplo do comprimento de um GOP de todos os sistemas testados. O banco de dados com os parâmetros calculados para todos os estimadores e todas as cenas é apresentado no Anexo 9.3.

Cada parâmetro (D) é relacionado ao resultado subjetivo por meio de uma função logística conforme equação (6.2).

$$d_{est}^i = \frac{1}{1 + \left(\frac{D}{D_M}\right)^G} \quad (6.2)$$

onde i varia de um a doze, correspondendo aos doze estimadores selecionados.

Os valores de D_M e G são determinados de forma a minimizar o erro quadrático médio entre o valor obtido na avaliação subjetiva e a nota subjetiva estimada por meio da função logística (d^i)

$$e_i = E\left[\left(d_{subjetivo} - d_{esti}^i\right)^2\right] \quad (6.3)$$

Obtêm-se desta forma, 12 curvas definidas por D_M e G referentes a cada um dos doze estimadores. As curvas estabelecem preditores da qualidade subjetiva de um sistema com base em seu valor de degradação definido pelo parâmetro.

Esta estimativa de qualidade pode ou não, ser boa. Um critério para decidir sobre a qualidade da predição é através de uma medida de confiabilidade. A confiabilidade é definida pelo inverso do erro da regressão não linear dos parâmetros [53]. A Figura 40 ilustra as diferenças em termos de confiabilidade dos parâmetros mostrando a dispersão dos dados subjetivos face à curva de regressão obtida.

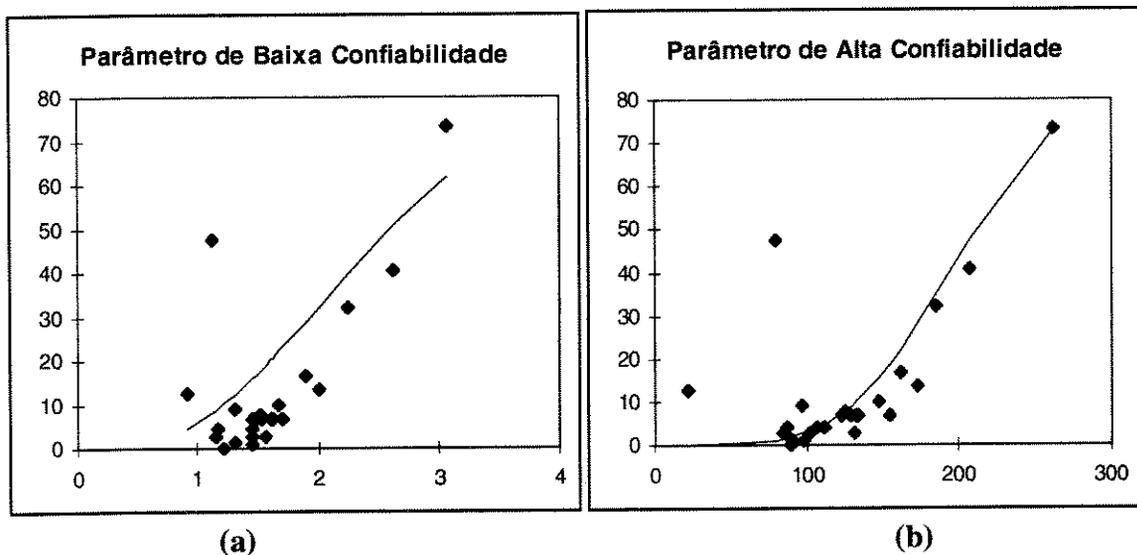


Figura 40: Ilustração de Confiabilidade dos parâmetros para a cena *Flower Garden*
 (a) medidos por Sobel negativo em áreas planas de luminância; (b) MSE em região de textura de crominância.

Numa segunda fase, estabelece-se uma regra para a combinação destas estimativas e

aobtenção da estimativa final. A ponderação entre os preditores é feita com base em sua confiabilidade. O valor de degradação final estimado é dado por:

$$d_{\text{final}} = \sum_{i=1}^{12} a_i d_{\text{est}}^i \quad (6.4)$$

Os coeficientes a_i são determinados no sentido de minimizar o erro quadrático médio.

$$\text{erro} = E\left[\left(d_{\text{subjetivo}} - d_{\text{final}}\right)^2\right] \quad (6.5)$$

satisfazendo as seguintes restrições:

$$\sum a_i = 1; \quad \frac{a_i}{a_k} = \frac{e_k}{e_i} \quad (6.6)$$

Por esta regra, o peso ao cada uma dos preditores é diretamente proporcional à sua confiabilidade. Esta combinação de preditores é feita em três etapas para possibilitar a obtenção de estimadores por região e, também, para as componentes de luminância e crominância separadamente. No primeiro estágio é obtida a estimativa de região por componente, numa segunda etapa são obtidas, separadamente, estimativas para a luminância e para a crominância e, finalmente, uma estimativa global na terceira etapa.

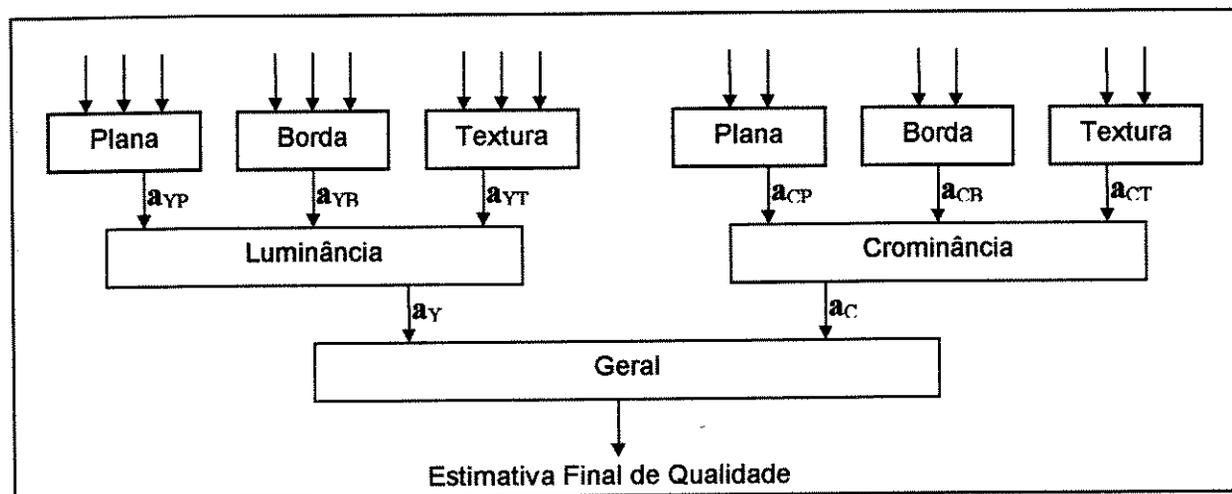


Figura 41: Algoritmo de combinação dos preditores

Conforme discutido em 4.3.3, devido à variabilidade, ruído e redundância dos testes subjetivos, a predição ótima, aquela que minimiza o erro quadrático médio entre a estimativa final e o dado subjetivo, é pouco robusta e, por isso, não é adequada.

6.4.2 Caracterização da Base de Dados e da Base de Treinamento

Como o banco de dados subjetivos é limitado, alguns sistemas foram utilizados para “treinar” o sistema e o restante dos dados para teste das funções e dos pesos obtidos. Na definição da base de treinamento e base de teste os sistemas foram divididos em quatro classes conforme a Tabela 17.

Classe	Denominação	Caracterização	Base de Treino	Base de Teste
Classe 1	S1 a S5	Sistemas MPEG-2 com taxa variável e qualidade constante	2	3
Classe 2	C4	Sistema MPEG-2 422P@ML com taxa constante	1	
Classe 3	C1, C2, C3 e C5	Sistemas MPEG-2 MP@ML com taxa constante	9	9
Classe 4		Sistemas de conversão para sinais compostos	1	1

Tabela 17: Definição das classes de sistemas.

Esta divisão em classes é importante para que se possa representar cada um destes sistemas embora a maioria dos dados subjetivos sejam da classe 3. Na ponderação dos estimadores para obtenção da estimativa final a minimização do erro quadrático médio é feita por classes.

6.4.3 Apresentação dos Resultados

A combinação dos pesos entre os diversos preditores define o algoritmo de estimação de qualidade subjetiva. As Figuras 41 e 42 mostram o peso relativo das componentes de luminância e crominância e os pesos por regiões em cada uma das cinco cenas testadas. É interessante notar que os resultados obtidos têm boa correlação com o que é intuitivamente esperado das cenas.

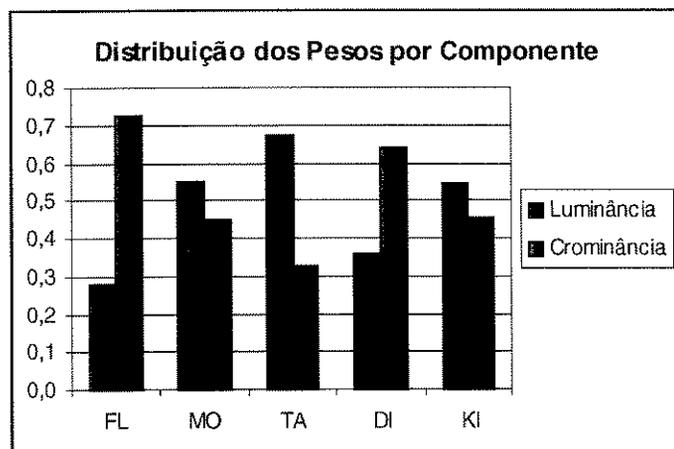


Figura 42: Distribuição dos pesos dos parâmetros objetivos - Resultado por componente.

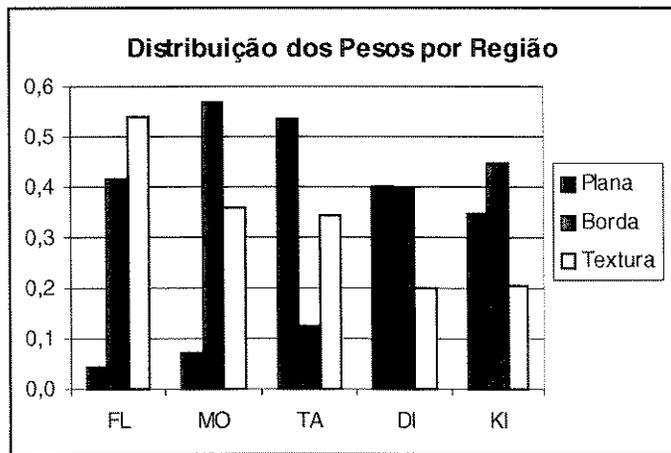


Figura 43: Distribuição dos pesos dos parâmetros objetivos - Resultado por região.

O desempenho da metodologia empregada é medido em termos do erro absoluto médio e do erro máximo. As Tabelas contendo os erros calculados para cada uma das cenas e discriminadas por classes são apresentadas no Anexo 9.4.

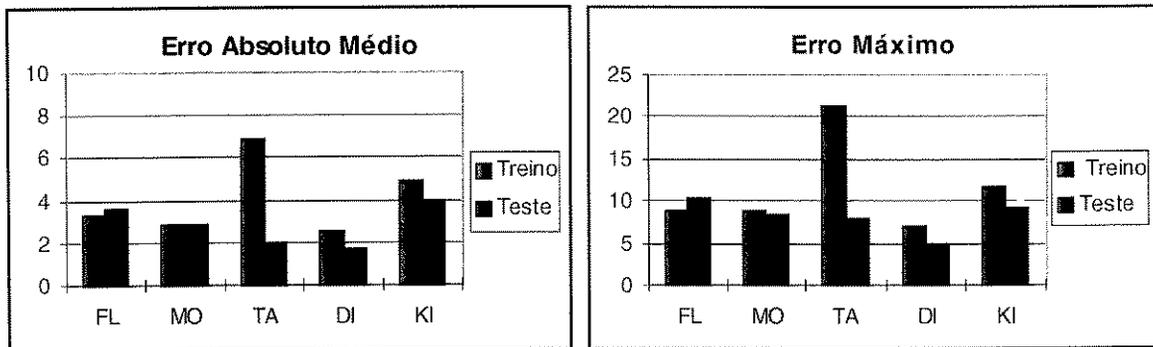


Figura 44: Erros de estimação de qualidade subjetiva.

Não há grandes discrepâncias entre as cenas de treino e de teste, exceto para a cena “Table Tennis”. Mesmo para esta cena a base de treino está bem escolhida e os erros absoluto e máximo dos sistemas de teste não ultrapassam os limites estabelecidos pela base de treino. Considerando todos os sistemas avaliados o erro absoluto médio não ultrapassa 7%, um erro bastante pequeno.

A utilização de medidas de avaliação objetiva baseadas em contexto tem um custo associado, elevada complexidade computacional da segmentação e aumento do número de estimadores. Para possibilitar comparações de desempenho, o mesmo procedimento foi aplicado utilizando estimadores baseados em medidas globais da imagem e utilizando apenas os estimadores de erro quadrático médio (MSE). As medidas globais utilizadas foram MSE, PSD e NSD para a componente de luminância e MSE para crominância. Medidas comparativas de erro por cena estão detalhadas no Anexo, na seção 9.4

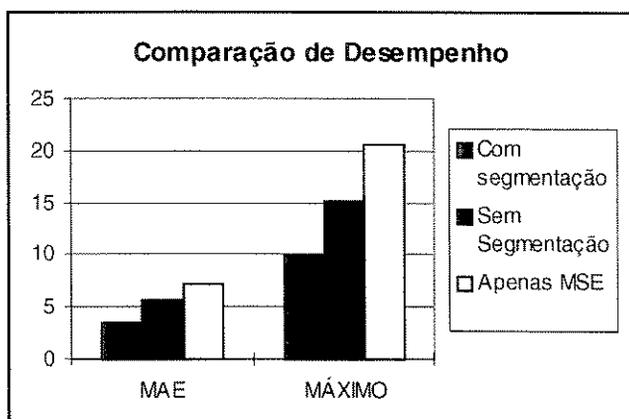


Figura 45: Comparação de desempenho - Ganho com Segmentação.

A utilização de medidas baseadas em contexto, podem proporcionar ganhos de estimação significativos em algumas cenas como em "Mobile & Calendar", para a qual a utilização de segmentação representa uma melhora de desempenho com relação à base de treinamento da ordem de 300%. Em outras cenas o ganho é pequeno ou não existe. Ainda assim, o aumento do erro absoluto médio é de 47% quando se utiliza medidas globais em relação a utilização de estimadores baseados em contexto. Essa diferença cresce para 72% quando somente a medida de erro quadrático médio por componente é utilizado.

Os resultados alcançados mostram que os parâmetros objetivos escolhidos medem as características de percepção fundamentais e são aplicáveis a larga gama de sistemas de vídeo digital. Ao lado dos gráfico relativos aos erros obtidos, a Figura 46 confirma esta conclusão. Nesta Figura é mostrada a dispersão dos dados objetivos frente a curva subjetiva desejada. Cada ponto assinalado corresponde a média de todas as cenas para os sistemas avaliados.

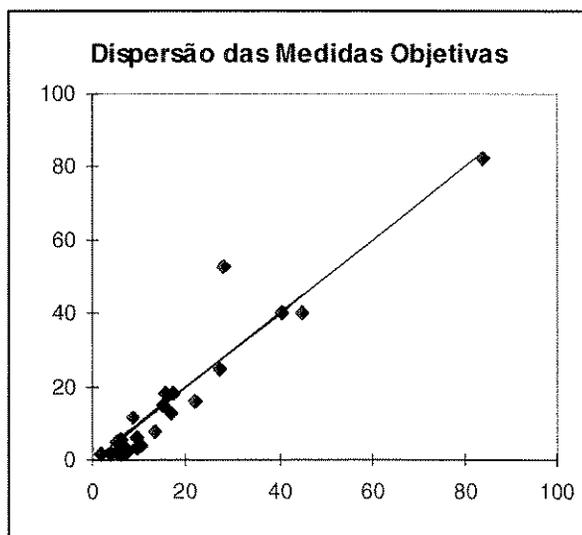


Figura 46: Gráfico de dispersão - Resultados objetivos x subjetivos.

Neste experimento foi aplicada a metodologia baseada na combinação linear dos estimadores analisada no capítulo 5, porém outras abordagens são possíveis. Uma possibilidade é a descrição dos estimadores através de conjuntos Fuzzy e a utilização de lógica Fuzzy para a sua combinação e estimação da qualidade subjetiva final.

7. Conclusões

Atualmente, a avaliação subjetiva de qualidade de imagem é a forma mais efetiva de se avaliar de forma confiável a qualidade de imagem, especialmente porque medidas amplamente utilizadas como erro quadrático médio e relação sinal ruído não provêem, em geral, boa indicação da qualidade percebida da imagem. A Recomendação ITU-R BT. 500 [02] é o padrão internacionalmente reconhecido para a avaliação subjetiva da qualidade de imagem em televisão. Este documento descreve métodos de escalonamento e condições de visualização para avaliação da qualidade percebida da imagem de forma padronizada. O objetivo destes métodos é coletar um conjunto de opiniões e notas que reflitam a impressão de qualidade dos observadores. A norma ANSI T1 801.03 [21], por sua vez, introduz novas medidas objetivas para quantificar o efeito da compressão digital, mas ainda não é um documento definitivo no sentido de fornecer um método estável e confiável para a avaliação de codecs de vídeo.

Os estudos e experimentos realizados mostram a necessidade de casamento entre o tipo de degradação presente, os objetivos da avaliação e a metodologia empregada. Este trabalho focalizou o estudo de novas maneiras de conduzir e projetar um experimento de avaliação subjetiva e métodos empíricos que permitam a extrapolação dos resultados obtidos sistemas ou configurações não testadas subjetivamente. As idéias apresentadas constituem adaptações e adequações das metodologias usuais para avaliação de televisão e as novas tecnologias e serviços.

Em avaliação subjetiva, todos os elementos relacionados ao projeto do teste, tais como: condições de visualização, escolha dos avaliadores, condições de referência, duração dos trechos de vídeo em teste, seleção do range de degradação e da escala de avaliação, tem impacto direto sobre o resultado obtido. Uma etapa fundamental do projeto de um experimento de avaliação é a escolha das cenas naturais a serem envolvidas nos testes subjetivos e objetivos. O estudo de criticalidade das cenas e as simulações realizadas demonstraram que a medida de criticalidade definida é uma boa forma de caracterizar o material de teste. Esta medida é também uma forma de especificar o material de teste de uma avaliação seja ela subjetiva ou objetiva em termos de parâmetros objetivos. Adicionalmente, na caracterização do material de teste, podem ser incluídas a informação temporal e a informação espacial da cena, conforme definido na norma ANSI T1 801.03 [21].

O ponto de partida de um teste é sempre a forma como as cenas serão analisadas. Com base na análise comparativa dos resultados de diversos experimentos, concluiu-se que o modo de analisar as cenas irá depender não apenas dos objetivos do teste, conforme recomendação do ITU [01], mas também do nível de degradação presente. Comparação entre as diversas escalas reforça que cada uma têm aplicações em determinada faixa de degradações. Se as degradações forem muito semelhantes, devem ser aplicados métodos comparativos e escalas que facilitem a estimativa de pequenas diferenças como, por exemplo, numéricas e gráficas. Métodos para

caracterização de degradações e de sistemas em largo “range” de degradação, devem ser projetados de forma que os estímulos estejam bem distribuídos ao longo da escala de notas e, preferencialmente, envolver um estímulo de referência. Neste caso, escalas puramente numéricas não são indicadas pois estão dissociadas do sentido de qualidade.

Um aspecto importante é o tratamento dado à construção de escalas de categorias de qualidade. Essas categorias são obtidas a partir de termos que representem cinco níveis de qualidade de imagem. Os aspectos lingüísticos deste tipo de escala e seu impacto nas notas em diferentes países não está bem definido na Recomendação. Especificamente, no caso do português falado no Brasil, os resultados experimentais mostraram que a escala de categorias de qualidade proposta não é totalmente aplicável. Os resultados experimentais evidenciaram tendência a inversão dos termos relativos a baixa qualidade. É um aspecto relevante quando se trata da realização de testes em português e de sua interpretação em outros idiomas. Desta forma, as escalas de degradação são preferíveis às escalas de qualidade dada à influência das diferenças individuais no significado dos termos.

Esta tese mostra a importância dos fatores psicológicos em especial no que diz respeito à influência da duração do trecho de vídeo e do contexto (demais cenas testadas). O entendimento e a quantificação destes efeitos são indispensáveis no projeto e melhoria das técnicas de avaliação. Os métodos de longa duração constituem em uma nova classe de métodos de avaliação com aplicação na simulação das reais condições de visualização do telespectador e da receptividade aos novos serviços. Têm a desvantagem de requerer aparato especial de aquisição de dados na sala de avaliação compensada pela possibilidade de monitoração da qualidade em tempo real.

Ao contrário do que prevê a recomendação, no experimento realizado, não são exploradas todas as condições em teste para todos os sistemas e cenas, mas apenas um subconjunto escolhido de forma a maximizar as informações depreendidas do testes. O ganho no número de situações em teste obtido tem como contrapartida uma pequena diminuição da confiabilidade das medidas, mas que pode ser compensada pelo aumento do número de avaliadores. Na análise dos resultados, adicionalmente, às ferramentas estatísticas usuais são aplicados critérios de suavização das notas determinados com base em conhecimento prévio do comportamento de sistemas de codificação MPEG-2 e na análise de redundância do teste.

A grande vantagem da avaliação subjetiva, é a obtenção de resultados válidos para os sistemas convencionais e sistemas de compressão sobre larga gama de aplicações para cenas paradas e cenas em movimento. Em contrapartida, uma grande variedade de métodos e parâmetros de teste, precisam ser considerados requerendo um arranjo meticuloso do “setup” e controle e diversos avaliadores devem ser envolvidos, o que a torna bastante dispendiosa e complexa. Os custos associado à avaliação subjetiva motivam o desenvolvimento de medidas objetivas mas com boa correlação com resultados subjetivos. A construção de modelos objetivos de qualidade de vídeo envolveu a condução simultânea de testes objetivos e subjetivos para determinar como os parâmetros objetivos podem ser usados para prever as respostas subjetivas dos avaliadores. Este procedimento determina a precisão dos parâmetros objetivos e identifica a porção do banco de dados subjetivos que pode ser explicada através de parâmetros objetivos.

De fato, as medidas objetivas escolhidas foram capazes de explicar uma parte surpreendentemente grande da variância subjetiva.

O método de estimação definido se baseia em diferenças entre imagens de entrada e de saída e na segmentação da imagem em três regiões: plana, borda e textura. Assim, os parâmetros definidos neste trabalho são altamente dependentes dos registros espacial e temporal entre os sinais de entrada e de saída do sistema sob avaliação e, por isso, preferencialmente aplicáveis a sistemas com canal e interfaces de entrada e saída digitais. O modelo de predição das notas subjetivas foi definido pelo cálculo da relação entre medidas objetivas e o resultado do experimento de avaliação subjetiva. A relação entre os parâmetros objetivos e os níveis de degradação subjetivos é aproximada por uma função logística. A estimativa de qualidade final é obtida por combinação linear dos parâmetros ponderado por sua confiabilidade estatística. Os resultados obtidos desta forma são melhores do que os resultados obtidos por medidas globais (sem segmentação) e mais robustos do que predição ótima.

Embora não contemple aplicações específicas, o modelo pode ser adaptado para levar em conta as diferenças entre os telespectadores, simplesmente tomando-se como base de dados subjetivos as notas apenas da população que se deseja caracterizar, por exemplo, avaliadores especialistas e não especialistas. O modelo não leva em conta, entretanto, as mudanças na expectativa da população ao longo do tempo. Estas mudanças são particularmente relevantes em sistemas de vídeo digital onde o avanço da tecnologia e o decréscimo dos custos são rápidos. Por esta razão, o desenvolvimento deste tipo de modelo de estimação de qualidade pressupõe definição precisa da aplicação e da população alvo do serviço.

A extensão do método proposto, deve permitir monitorar a qualidade do vídeo em serviço, de forma não intrusiva, muito útil em aplicações de detecção de falhas, monitoração automática de qualidade e otimização dinâmica de recursos de rede finitos. Refinamentos posteriores destes métodos de comprimir a informação sobre a qualidade de vídeo, prometem produzir um método para medir a qualidade do vídeo bom o suficiente para, em muitos casos, substituir os experimentos subjetivos. No atual contexto tecnológico, entretanto, a informação subjetiva ainda é fundamental em todas as etapas de desenvolvimento dos serviços, bem como de outras ferramentas de avaliação.

8. Referências Bibliográficas

- [01] ITU-R Handbook; Subjective assessment methodology in television; 1996.
- [02] ITU-R Recommendation BT.500-7; Methodology for the subjective assessment of the quality of television pictures; 1995.
- [03] ITU-R Recommendation BT.601-5; Studio encoding parameters of digital television for standard 4:3 and wide screen 16:9 aspect ratios; 1995.
- [04] ITU-R Recommendation BT. 710-2; Subjective assessment methods for image quality in high-definiton television; 1994.
- [05] ITU-R Recommendation BT.802-1; Test pictures and sequences for subjective assessments of digital codecs conveying signals produced according to Recommendation 601; 1994.
- [06] ITU-R Recommendation BT. 813; Methods for objective picture quality assessment in relation to impariments from digital coding of television signals; 1992.
- [07] ITU-T Recommendation H.262; Generic coding of moving pictures and associated audio - information: video; 1995.
- [08] ITU-T Recommendation P.910; Subjective video quality assessment methods for multimedia applications; 1996.
- [09] ITU-T Recommendation P.930; Principles of reference impairment system for video, 1996.
- [10] ITU-T Recommendation T.81; Digital compression and coding of continuous-tone still images - requirements and guidelines; 1992.
- [11] ITU-T Report 1082-1; Studies toward the unification of picture assessment methodology; 1990.
- [12] ITU-T Report 1205; Analisis and presentation of the results of television subjective testes; 1990.
- [13] ITU-T Report 1206; Methods for picture quality assessment in relation to impairment from digital coding of television signals; 1990.
- [14] ITU-T, SG15, Doc No. AVC-491; Test Model 5; 1993.
- [15] Standard SMPTE 244M, Television digital component recording - 19 mm type D1 tape record; 1990.
- [16] Abraham, D.; Ardito, M.; Boch, L.; Messina, A. Stroppiana, M. e Visca, M.; Attempsts at correlation between DSCQS and objective measurements; *EBU Technical Review*, pp. 21-30 primavera, 1997.

- [17] Aldridge, R., Davidoff, J. Ghanbari, M., Hands, D. e Pearson, D.; Recency effect in subjective assessment of digitally-coded television pictures; *Conference Publication No 410, Image Processing and Its Applications*, pp. 336-339 julho, 1995.
- [18] Allnatt, J. W.; Television measurements through psychophysics to subjective picture quality; *The Radio and Electronic Engineer*, vol. 49 no. 12 pp. 611-619 dezembro, 1979.
- [19] Alpert, Thierry e Evain, Jean Pierre; Subjective quality evaluation - The SSCQE and DSCQE methodologies; *EBU Technical Review*, 12-20 spring, 1997.
- [20] Anderberg, Michael R.; *Clusters Analysis for Applications*. New York: Academic Press, 1973.
- [21] ANSI T1.801.03-1996 - Telecommunications - Digital transport of one-way video signals - Parameters for objective performance assessment, 1996.
- [22] Ardito, Maurizio e Visca, Massimo; Correlation between objective and subjective measurements for video compressed systems; *International Broadcasting Convention , Conference Publication No 413 IEE*, pp. 7-12 setembro, 1995.
- [23] Ardito, Maurizio e Visca, Massimo; Correlation between objective and subjective measurements for video compressed systems; *SMPTE Journal*, pp. 768-773 dezembro, 1996.
- [24] Beakley, G., Cressy, C. e Van Pelt, J.; Video Compression: The Need for Testing; *SMPTE Journal*, pp. 742-750 novembro, 1995.
- [25] Bjøntegaard, Gisle; Lillevold, Karl Olav e Danielsen, Robert; A comparison of different coding formats for digital coding of video using MPEG-2; *IEEE Transactions on Image Processing*, vol. 5 no. 8 pp. 1271-1276 agosto, 1996.
- [26] Cemark, Gregory W.; Terare, Sandra K. ; Tweedy Ernest P. e Stoddard James C.; Consumer acceptance of MPEG-2 video at 3.0 to 8.3Mb/s; *Proceedings SPIE*, vol. 2917 pp. 53-61, 1996.
- [27] Corbett, J. M.; Effect of observer adaptation on the results of television quality-grading tests; *Proceedings IEE*; vol. 117 no. 3 pp. 512-514, 1970.
- [28] Dalton, Chris J.; Concatenation of compression codecs - The need for objective evaluations; *EBU Technical Review*, pp. 31-40 ,1997.
- [29] Dalton, Chris J.; Objective assessment of single and concatenated compression codecs; *SMPTE Journal*, pp. 698-701 novembro, 1996.
- [30] Gardiner, P. N.; Ghanbari, M.; Pearson, D. E. e Tan, K. T.; Development of a perceptual distortion meter for digital video; *International Broadcasting Convention, Conference Publication No 447 IEE*, pp. 493-497 setembro, 1997.
- [31] Gasman, C.; Andronov, V.; Bukina, A. e Vasilyev, O.; Subjective assessment of compression systems by trained and untrained observers; *International Broadcasting Convention, Conference Publication No 447 IEE*, pp. 476-481 setembro, 1997.

- [32] Goodman J.S. e Pearson D. E.; Multidimensional scaling of multiply-impaired television pictures; *IEEE Transactions on Systems Man, and Cybernetics*, vol. SMC-9 no. 6 pp. 353-356 junho, 1979.
- [33] Hamberg, Roelof e de Ridder Huib; Continuous assessment of image quality; *SMPTE Journal*, pp. 123-128 fevereiro, 1997.
- [34] Hamberg, Roelof e de Ridder, Huib; Continuous assessment of perceptual image quality; *Journal of the Optical Society of America A*; vol. 12 no. 12 pp. 2573-2577 dezembro, 1995.
- [35] Haskell, Barry G.; Puri, Atul; Netravali, Arun N.; Digital video: An introduction to MPEG-2, *International Thompson Publishing*, 1996.
- [36] Hess, Robert F. e Badcock David R.; Metric for separation discrimination by the human visual system; *Journal of the Optical Society of America A*, vol. 12 no. 1 pp. 3-16 janeiro, 1995.
- [37] Hidaka, Tsuneyoshi e Ozawa Keiji; Subjective assessment of redundancy-reduced moving images for interactive application: test methodology and report; *Signal Processing: Image Communication*, pp. 201-219, 1990.
- [38] Jayant, N. S. e Noll, Peter; Digital Coding of Waveforms Principles and Applications to Speech and Video. *New Jersey: Prentice Hall*, 1984.
- [39] Jones, Bronwen L. e McManus, Pamela R.; Graphic scaling of qualitative terms; *SMPTE Journal*, pp. 1166-1171 novembro, 1986.
- [40] Jones, Bronwyn e Marks, Lawrence E.; Picture quality assessment: A comparison of ratio and ordinal scales; *SMPTE Journal*, pp. 1244-1248 dezembro, 1985.
- [41] Le Gall, Didier J.; The MPEG video compression algorithm; *Signal Processing: Image Communication*, vol. 4 no. 2 pp. 129-140, 1992.
- [42] Lewis, N. W. e Allnatt, J. W.; Subjective and objective impairments in television pictures; *Conference Publication IEE*, vol. 46 pp. 2.1.1-5, 1968.
- [43] Lewis, N. W. e Allnatt, J. W.; Subjective quality of television pictures with multiple impairments; *Electronics Letters*, pp. 187-188, 1965.
- [44] Lim, Jae S.; Two-dimensional Signal and Image Processing; capítulo 8. *Prentice Hall Signal Processing Series*, 1990.
- [45] Limb, John O.; Distortion criteria of the human viewer; *IEEE Transactions on Systems, Man, and Cybernetics*; vol. SMC-9 no. 12 pp. 778-793 dezembro, 1979.
- [46] Lookabaugh, Tom; Matching MPEG-2 to applications; *SMPTE Journal*, pp. 4-8 janeiro, 1995.
- [47] Lubin, J.; A human vision system model for objective picture quality measurements; *International Broadcasting Convention, Conference Publication No 447 IEE*, pp. 498-503 setembro, 1997.

- [48] Lupker, Stephen J. e Hearty Paul J.; Analising the effects of multiple sources of impairment in television signals; *Proceedings 3rd International Colloquim on Advanced Television Systems HDTV'87 - Canada*, pp. 1.2.1-17, 1987.
- [49] Miceli, S. e Orlando, A.; Subjective assessment elaboration for determining the distortion tolerances of a telecommunication system; *Proc. IEEE Eurocon'77, Veneza, Itália*; pp. 2.4.3 1-5, 1977.
- [50] Nakasu, E., Aoki, K., Yajima, R., Kanatsugu, Y. e Kubota, K.; A statistical analysis of MPEG-2 picture quality for television broadcasting; *SMPTE Journal*, pp. 702-711 novembro, 1996.
- [51] Narita, Nagato; Subjective-evaluation method for quality of coded images; *IEEE Transactions on Broadcasting*, vol. 40 no. 1 pp. 7-13 março, 1994.
- [52] Pessoa, Antonio C. F.; Uma metodologia para a avaliação objetiva de codecs de vídeo; *XV Simpósio Brasileiro de Telecomunicações - Recife*, setembro, 1997.
- [53] Pessoa, Antonio C. França; Video quality assessment using objective parameters based on image segmentation; *ITU Contribution Document - Study Group 12*, dezembro 1997
- [54] Prosser, R. D.; Allnatt, J. W. e Lewis, N. W.; Quality grading of impaired television pictures; *Proceedings IEE*; vol. 111 no. 3 pp. 491-502 março, 1964.
- [55] Ridder, Huib de e Hamberg, Roelof; Continuous assessment of image quality; *SMPTE Journal*, pp. 123-128 fevereiro, 1997.
- [56] Ridder, Huib de e Majoor, Gijberta M.M.; Numerical category scaling: an efficient method for assessing digital image coding impairments; *Proceedings SPIE Human Vision and Electronic Imaging: Models, Methods and Applications*, vol. 1249 pp. 65-77, 1990.
- [57] Riddre, Huib de; Minkowski-metrics as a combination rule for digital-image coding impairments; *Proceedings SPIE Human Vision, Visual Processing, and Digital Display III*, vol.1666 pp. 16-26, 1992.
- [58] Sallio P. e Kretz, F.; A comparison of two methods for the subjective evaluation of television pictures. Representation of the results in common units; *EBU Technical Review*, no.192 pp. 59-69 abril, 1982.
- [59] Seferidis, V., Ghanbari, M. e Pearson, D. E.; Forgiveness effect in subjective assessment of packet video; *Electronics Letters*, vol. 28 no. 21 pp. 2013-2014 outubro, 1992.
- [60] Siocos, C. A.; Subjective impairment of television pictures - Considerations regarding the summability of impairment units; *Journal of the SMPTE*, vol. 81 pp. 917-920 dezembro, 1972.
- [61] Teunissen, Kees e Westerink, Joyce H. D. M.; A multidimensional evaluation of the perceptual quality of television sets; *SMPTE Joutnal*, pp. 31-38 janeiro, 1996.
- [62] Teunissen Kees; The validity of CCIR quality indicators along a graphical scale; *SMPTE Journal*, pp. 144-149 março, 1996.

- [63] Webster, Arthur A., Jones, Coleen T. e Pinson, Margaret H.; An objective video quality assessment system based on human perception; *Proceeding SPIE*, vol. 1913 pp. 15-26 fevereiro, 1993.
- [64] Westerink, Joyce H. D. M. e Roufs, Jaques A. J.; Subjective image quality as a function of viewing distance, resolution, and picture size; *SMPTE Journal*; pp. 113-119 fevereiro, 1989.
- [65] Wilkinson, J. H. e Stone, J.J.; Cascading diferent types of video compression systems; *SMPTE European Conference, Colonia-Alemanha*, setembro, 1994.
- [66] Wolf, Stephen; Objective and subjective measures of MPEG video quality: summary of experimental results; *ITU-T Delayed Contribution D.101*, USA, 1997
- [67] Wolf, Stephen; Selections from the draft American National Standard: Digital Transport of one-way signals - parameters for objective performance assessment; *ITU Contribution Document -COM 12-66*, USA, 1997
- [68] Zetzche C. e Hauske, G.; **Principal** features of human vision in the context of image quality models; pp. 102-103.
- [69] Zou, Wiliam Y.; Performance evaluation: From NTSC to digitally compressed video; *SMPTE Journal*, pp. 795-800 dezembro, 1994.

9. Anexos

9.1 Tabela de Resultados: Análise Gráfica dos Descritores de Qualidade

TABELA 9.1.1: Média Geral

Esta tabela é obtidas a partir da média e desvio padrão das notas dadas por todas as pessoas envolvidas. Os cinco termos efetivamente empregados na avaliação, de acordo com ITU-R Recomendação BT 500 [02] são mostrados em **negrito**.

Descritor	Média	Desvio Padrão
Excelente	4,82	0,19
Ótimo	4,58	0,23
Muito Bom	4,31	0,28
Bom	3,86	0,56
Razoável	3,06	0,49
Médio	3,22	0,59
Regular	3,22	0,43
Pobre	1,61	0,70
Ruim	2,13	0,55
Muito Ruim	2,02	0,50
Medíocre	1,70	0,48
Péssimo	1,29	0,38
	Desvio Padrão	0,45

TABELA 9.1.2: Média Discriminada por Idade

Esta tabela é obtidas a partir das médias calculadas separadamente para observadores acima e abaixo de 25 anos de idade, o que divide a população de avaliadores aproximadamente na metade. Os cinco termos efetivamente empregados na avaliação, de acordo com ITU-R Recomendação BT 500 [02] são mostrados em **negrito**.

Descritor	> 25 anos	< 25 anos
Excelente	4,86	4,78
Ótimo	4,62	4,53
Muito Bom	4,34	4,28
Bom	3,96	3,76
Razoável	3,10	3,01
Médio	3,40	3,04
Regular	3,23	3,22
Pobre	1,79	1,44
Ruim	2,19	2,08
Muito Ruim	1,89	2,16
Medíocre	1,55	1,85
Péssimo	1,19	1,39

9.2 Base de Dados Subjetivos

TABELA 9.2.1: Médias Gerais

Esta tabela é construída a partir da média das notas dadas por todos os avaliadores. Em destaque são mostrados os sistemas que foram utilizados para estimação de qualidade subjetiva por meio de medidas objetivas.

		MO	FL	TA	KI	DI	Média
Referência	Original	4,83	4,86	4,82	4,75	4,85	4,82
18 Mbits/s	C4 (M=1)	4,80	4,83	4,66	4,75	4,73	4,75
15 Mbits/s	C5 (M=1)	4,37	4,77	4,63	4,43	4,85	4,61
	C5 (M=2)	4,61	4,77	4,72	4,62	4,85	4,71
	C1 (M=1)	4,65	4,75	4,78	4,52	4,71	4,68
	C1 (M=2)	4,68	4,75	4,78	4,63	4,80	4,73
	C2 (M=2)	4,53	4,82	4,68	4,70	4,80	4,71
12 Mbits/s	C5 (M=1)	4,37	4,70	4,63	4,43	4,71	4,57
	C5 (M=2)	4,37	4,70	4,63	4,43	4,71	4,57
10 Mbits/s	C5 (M=1)	4,27	4,60	4,57	4,43	4,70	4,51
	C5 (M=2)	4,34	4,60	4,57	4,43	4,71	4,53
	C1 (M=1)	4,37	4,70	4,48	4,28	4,71	4,51
	C1 (M=2)	4,37	4,70	4,59	4,28	4,71	4,53
	C2 (M=2)	4,37	4,60	4,63	4,37	4,71	4,54
8,5 Mbits/s	C3 (M=2)	4,27	4,60	4,48	4,28	4,60	4,45
7,5 Mbits/s	C5 (M=1)	3,93	4,60	4,22	4,28	4,60	4,33
	C5 (M=2)	3,93	4,60	4,27	4,28	4,60	4,34
5 Mbits/s	C5 (M=1)	3,18	4,22	3,81	3,60	4,10	3,78
	C5 (M=2)	3,40	4,60	4,22	3,60	4,40	4,04
	C1 (M=1)	3,50	4,52	4,10	3,50	3,80	3,89
	C1 (M=2)	3,93	4,52	4,10	3,81	4,30	4,13
	C2 (M=2)	3,56	4,48	4,09	3,53	4,60	4,05
	C3 (M=2)	3,74	4,52	4,22	3,75	4,40	4,13
3,5 Mbits/s	C3 (M=2)	2,88	3,69	3,22	2,74	3,23	3,15
2,5 Mbits/s	C5 (M=1)	2,30	3,63	3,00	2,11	3,23	2,85
	C5 (M=2)	2,42	3,63	3,22	2,74	3,23	3,05
Sistema de Qualidade Constante	S1 (Q=4)	4,74	4,86	4,76	4,66	4,84	4,77
	S2 (Q=8)	4,33	4,56	4,56	4,66	4,84	4,59
	S3 (Q=16)	3,14	4,34	3,81	4,06	4,55	3,98
	S4 (Q=32)	1,76	3,30	2,34	2,02	2,84	2,45
	S5 (Q=62)	1,15	2,03	1,34	1,40	1,23	1,43
Conversões	NTSC (soft)	4,75	4,38	4,67	4,52	4,58	4,58
	NTSC (equip)	3,36	3,61	4,09	4,41	4,74	4,04
	PAL (soft)	3,65	3,02	4,35	4,23	4,32	3,91
	PAL (equip)	3,36	3,61	3,69	4,41	4,08	3,83
	Espacial	3,24	3,34	1,77	1,83	3,04	2,64
	Temporal	2,30	1,46	1,37	1,30	2,25	1,73

TABELA 9.2.2: Médias dos Avaliadores Especialistas

Esta tabela é construída a partir da média das notas dadas pelos avaliadores classificados como especialistas, devido a sua experiência na avaliação de qualidade de imagem. Em destaque são mostrados os sistemas que foram utilizados para estimação de qualidade subjetiva por meio de medidas objetivas.

		MO	FL	TA	KI	DI	Média
Referência	Original	4,75	4,92	4,80	4,88	4,89	4,85
18 Mbits/s	C4 (M=1)	4,59	4,83	4,67	4,77	4,50	4,67
15 Mbits/s	C5 (M=1)	3,89	4,71	4,56	4,34	4,69	4,44
	C5 (M=2)	4,56	4,71	4,73	4,38	4,69	4,61
	C1 (M=1)	4,46	4,76	4,79	4,40	4,66	4,61
	C1 (M=2)	4,46	4,76	4,79	4,58	4,71	4,66
	C2 (M=2)	4,39	4,78	4,67	4,57	4,73	4,63
12 Mbits/s	C5 (M=1)	3,88	4,71	4,56	4,34	4,66	4,43
	C5 (M=2)	3,88	4,71	4,56	4,34	4,66	4,43
10 Mbits/s	C5 (M=1)	3,86	4,57	4,56	4,31	4,65	4,39
	C5 (M=2)	3,86	4,57	4,56	4,31	4,66	4,39
	C1 (M=1)	3,88	4,71	4,29	4,34	4,65	4,37
	C1 (M=2)	3,88	4,71	4,50	4,34	4,65	4,42
	C2 (M=2)	3,88	4,60	4,36	4,27	4,66	4,36
8,5 Mbits/s	C3 (M=2)	3,86	4,57	4,29	4,27	4,42	4,28
7,5 Mbits/s	C5 (M=1)	3,00	4,57	3,98	4,27	4,42	4,05
	C5 (M=2)	3,75	4,57	3,98	4,27	4,42	4,20
5 Mbits/s	C5 (M=1)	2,30	4,21	3,72	3,40	3,77	3,48
	C5 (M=2)	2,75	4,57	3,98	3,75	4,10	3,83
	C1 (M=1)	2,75	4,52	3,91	3,10	3,56	3,57
	C1 (M=2)	3,	4,52	3,91	3,88	3,75	3,81
	C2 (M=2)	3,00	4,39	3,97	3,45	4,22	3,80
	C3 (M=2)	3,00	4,48	3,82	3,44	3,88	3,72
3,5 Mbits/s	C3 (M=2)	2,30	3,68	2,86	2,23	2,65	2,74
2,5 Mbits/s	C5 (M=1)	2,20	3,59	2,30	1,25	2,65	2,40
	C5 (M=2)	2,20	3,59	2,42	2,00	2,65	2,57
Sistema de Qualidade Constante	S1 (Q=4)	4,75	4,92	4,75	4,30	4,88	4,72
	S2 (Q=8)	3,75	4,61	4,50	4,25	4,50	4,32
	S3 (Q=16)	2,80	4,25	3,44	3,30	4,42	3,64
	S4 (Q=32)	1,50	2,30	2,50	1,67	2,25	2,04
	S5 (Q=62)	1,10	1,00	1,10	1,00	1,11	1,06
Conversões	NTSC (soft)	4,69	4,10	4,70	4,30	4,04	4,36
	NTSC (equip)	2,83	3,34	3,79	4,30	4,60	3,77
	PAL (soft)	3,17	2,80	4,17	3,83	4,04	3,60
	PAL (equip)	2,83	3,34	3,79	4,30	3,92	3,64
	Espacial	3,38	2,90	1,28	1,50	1,25	2,06
	Temporal	2,83	1,00	1,50	1,00	2,56	1,78

TABELA 9.2.3: Média dos Avaliadores Não Especialistas

Esta tabela é construída a partir da média das notas dadas pelos avaliadores classificados como não especialistas. Em destaque são mostrados os sistemas que foram utilizados para estimação de qualidade subjetiva por meio de medidas objetivas.

		MO	FL	TA	KI	DI	Média
Referência	Original	4,90	4,88	4,88	4,70	4,88	4,85
18 Mbits/s	C4 (M=1)	4,87	4,83	4,68	4,70	4,81	4,78
15 Mbits/s	C5 (M=1)	4,48	4,81	4,71	4,44	4,87	4,66
	C5 (M=2)	4,61	4,81	4,76	4,61	4,87	4,73
	C1 (M=1)	4,60	4,73	4,83	4,48	4,79	4,69
	C1 (M=2)	4,75	4,73	4,83	4,61	4,79	4,74
	C2 (M=2)	4,57	4,88	4,74	4,70	4,88	4,76
12 Mbits/s	C5 (M=1)	4,48	4,66	4,71	4,44	4,78	4,61
	C5 (M=2)	4,48	4,66	4,71	4,44	4,78	4,61
10 Mbits/s	C5 (M=1)	4,45	4,63	4,69	4,44	4,78	4,60
	C5 (M=2)	4,45	4,63	4,69	4,44	4,78	4,60
	C1 (M=1)	4,48	4,66	4,64	4,09	4,77	4,53
	C1 (M=2)	4,48	4,66	4,71	4,09	4,77	4,54
	C2 (M=2)	4,44	4,62	4,71	4,42	4,78	4,60
8,5 Mbits/s	C3 (M=2)	4,44	4,62	4,64	4,09	4,70	4,50
7,5 Mbits/s	C5 (M=1)	3,90	4,62	4,43	4,09	4,70	4,35
	C5 (M=2)	3,90	4,62	4,43	4,09	4,70	4,35
5 Mbits/s	C5 (M=1)	3,36	4,26	3,98	3,50	4,32	3,88
	C5 (M=2)	3,58	4,62	4,42	3,50	4,43	4,11
	C1 (M=1)	3,71	4,54	4,37	3,44	3,90	3,99
	C1 (M=2)	3,90	4,54	4,37	3,80	4,40	4,20
	C2 (M=2)	3,60	4,60	4,27	3,51	4,54	4,10
	C3 (M=2)	3,84	4,62	4,43	3,72	4,65	4,25
3,5 Mbits/s	C3 (M=2)	3,02	3,72	3,73	2,82	3,19	3,29
2,5 Mbits/s	C5 (M=1)	2,30	3,72	3,26	2,17	3,19	2,93
	C5 (M=2)	2,50	3,72	3,73	2,82	3,19	3,19
Sistema de Qualidade Constante	S1 (Q=4)	4,67	4,83	4,88	4,70	4,85	4,79
	S2 (Q=8)	4,45	4,53	4,70	4,70	4,85	4,65
	S3 (Q=16)	3,22	4,36	4,20	4,03	4,61	4,08
	S4 (Q=32)	1,88	3,55	2,70	2,20	3,00	2,67
	S5 (Q=62)	1,17	2,32	1,47	1,48	1,30	1,55
Conversões	NTSC (soft)	4,79	4,50	4,79	4,53	4,71	4,66
	NTSC (equ)	3,48	3,62	4,14	4,43	4,77	4,09
	PAL (soft)	3,86	3,09	4,67	4,10	4,37	4,02
	PAL (equip)	3,48	3,62	3,18	4,43	4,15	3,77
	Espacial	3,20	3,45	2,17	1,86	3,38	2,81
	Temporal	1,94	1,55	1,33	1,33	2,07	1,64

TABELA 9.2.4: Notas Médias à Distância de 4H

Esta tabela é construída a partir da média das notas dadas por todos os avaliadores na sessão em que se sentaram a uma distância de 4H do monitor (H=altura do monitor). Em destaque são mostrados os sistemas que foram utilizados para estimação de qualidade subjetiva por meio de medidas objetivas.

		MO	FL	TA	KI	DI	Média
Referência	Original	4,86	4,87	4,94	4,71	4,89	4,86
18 Mbits/s	C4 (M=1)	4,86	4,79	4,57	4,71	4,63	4,71
15 Mbits/s	C5 (M=1)	4,39	4,82	4,62	4,56	4,88	4,65
	C5 (M=2)	4,68	4,87	4,82	4,56	4,88	4,76
	C1 (M=1)	4,63	4,71	4,75	4,62	4,73	4,69
	C1 (M=2)	4,77	4,71	4,89	4,62	4,85	4,77
	C2 (M=2)	4,61	4,80	4,73	4,58	4,89	4,72
12 Mbits/s	C5 (M=1)	4,39	4,71	4,62	4,34	4,66	4,54
	C5 (M=2)	4,39	4,71	4,62	4,34	4,66	4,54
10 Mbits/s	C5 (M=1)	4,30	4,59	4,62	4,23	4,66	4,48
	C5 (M=2)	4,39	4,71	4,56	4,23	4,66	4,51
	C1 (M=1)	4,39	4,63	4,44	4,19	4,59	4,45
	C1 (M=2)	4,39	4,65	4,62	4,19	4,59	4,49
	C2 (M=2)	4,34	4,66	4,60	4,18	4,66	4,49
8,5 Mbits/s	C3 (M=2)	4,30	4,59	4,44	4,18	4,59	4,42
7,5 Mbits/s	C5 (M=1)	4,03	4,59	4,28	4,18	4,39	4,29
	C5 (M=2)	4,03	4,59	4,31	4,18	4,39	4,30
5 Mbits/s	C5 (M=1)	3,10	4,32	4,00	3,50	4,06	3,80
	C5 (M=2)	3,10	4,59	4,27	3,50	4,14	3,92
	C1 (M=1)	3,22	4,55	4,25	3,41	3,83	3,85
	C1 (M=2)	4,03	4,55	4,28	3,75	4,18	4,16
	C2 (M=2)	3,45	4,52	4,27	4,18	4,39	4,16
	C3 (M=2)	3,66	4,42	4,26	3,77	4,39	4,10
3,5 Mbits/s	C3 (M=2)	2,83	3,53	3,43	2,42	3,20	3,08
2,5 Mbits/s	C5 (M=1)	2,18	3,53	3,19	2,42	3,20	2,90
	C5 (M=2)	2,38	3,53	2,89	2,42	3,20	2,89
Sistema de Qualidade Constante	S1 (Q=4)	4,64	4,86	4,94	4,56	4,87	4,78
	S2 (Q=8)	4,15	4,50	4,64	4,56	4,87	4,55
	S3 (Q=16)	3,12	4,50	3,79	3,86	4,50	3,96
	S4 (Q=32)	1,25	3,35	2,32	2,00	2,45	2,27
	S5 (Q=62)	1,00	1,56	1,35	1,27	1,21	1,28
Conversões	NTSC (soft)	4,61	4,15	4,56	4,46	4,68	4,49
	NTSC (equip)	3,42	3,27	3,85	4,25	4,77	3,91
	PAL (soft)	3,33	2,90	4,45	3,54	4,09	3,66
	PAL (equip)	3,42	3,27	3,05	4,25	4,00	3,60
	Espacial	3,09	3,19	1,88	1,78	2,86	2,56
	Temporal	2,36	1,50	1,18	1,27	2,00	1,66

TABELA 9.2.5: Notas Médias à Distância de 6H

Esta tabela é construída a partir da média das notas dadas por todos os avaliadores na sessão em que se sentaram a uma distância de 6H do monitor (H=altura do monitor). Em destaque são mostrados os sistemas que foram utilizados para estimação de qualidade subjetiva por meio de medidas objetivas.

		MO	FL	TA	KI	DI	Média
Referência	Original	4,89	4,89	4,91	4,75	4,92	4,87
18 Mbits/s	C4 (M=1)	4,74	4,87	4,78	4,73	4,83	4,79
15 Mbits/s	C5 (M=1)	4,30	4,67	4,67	4,45	4,75	4,57
	C5 (M=2)	4,54	4,67	4,71	4,45	4,75	4,62
	C1 (M=1)	4,64	4,83	4,77	4,43	4,75	4,68
	C1 (M=2)	4,64	4,83	4,83	4,55	4,75	4,72
12 Mbits/s	C2 (M=2)	4,46	4,87	4,72	4,75	4,75	4,71
	C5 (M=1)	4,27	4,67	4,67	4,43	4,74	4,56
10 Mbits/s	C5 (M=2)	4,27	4,67	4,67	4,43	4,74	4,56
	C5 (M=1)	4,26	4,52	4,59	4,43	4,71	4,50
8,5 Mbits/s	C5 (M=2)	4,26	4,52	4,59	4,43	4,71	4,50
	C1 (M=1)	4,27	4,67	4,65	4,42	4,74	4,55
	C1 (M=2)	4,27	4,67	4,65	4,42	4,74	4,55
	C2 (M=2)	4,27	4,54	4,67	4,42	4,73	4,52
7,5 Mbits/s	C3 (M=2)	4,27	4,52	4,59	4,42	4,65	4,49
5 Mbits/s	C5 (M=1)	3,94	4,52	4,39	4,27	4,65	4,35
	C5 (M=2)	3,96	4,52	4,42	4,27	4,65	4,36
3,5 Mbits/s	C5 (M=1)	3,04	4,37	3,84	3,55	4,14	3,79
	C5 (M=2)	3,77	4,37	4,39	3,61	4,61	4,15
	C1 (M=1)	3,68	4,48	4,14	3,41	3,71	3,88
	C1 (M=2)	3,94	4,48	4,14	3,90	4,39	4,17
	C2 (M=2)	3,68	4,38	4,16	3,75	4,38	4,07
	C3 (M=2)	3,73	4,52	4,37	3,69	4,40	4,14
2,5 Mbits/s	C3 (M=2)	2,92	3,81	3,66	2,58	3,30	3,25
Sistema de Qualidade Constante	C5 (M=1)	2,44	3,81	3,07	2,58	3,30	3,04
	C5 (M=2)	2,46	3,81	3,66	2,58	3,30	3,16
	S1 (Q=4)	4,83	4,89	4,82	4,61	4,86	4,80
	S2 (Q=8)	4,46	4,71	4,67	4,61	4,75	4,64
	S3 (Q=16)	3,17	4,14	4,04	3,83	4,59	3,95
Conversões	S4 (Q=32)	2,14	3,25	2,96	2,00	3,38	2,75
	S5 (Q=62)	1,33	2,50	1,46	1,47	1,25	1,60
	NTSC (soft)	4,86	4,71	4,91	4,50	4,56	4,71
	NTSC (equip)	3,29	4,04	4,22	4,54	4,71	4,16
	PAL (soft)	3,91	3,13	4,73	4,42	4,56	4,15
	PAL (equip)	3,29	3,79	3,83	4,54	4,15	3,92
Espacial	Espacial	3,44	3,50	1,79	1,73	3,18	2,73
	Temporal	2,25	1,42	1,61	1,25	2,50	1,81

9.3 Base de Dados Objetivos

Os parâmetros são calculados conforme definição no Capítulo 5, em cada uma das regiões da imagem (plana, borda ou textura) e para cada componente (luminância e crominância) separadamente. Os valores apresentados correspondem a média em doze quadros da cena num trecho de regime (distante de corte de cena).

TABELA 9.3.1: Base de Dados Objetivos da Cena "Flower Garden"

Luminância

Características		Plana			Borda			Textura		
		MSE	PSD	NSD	MSE	PSD	NSD	MSE	PSD	NSD
18 Mbits/s	C4 (M=1)	2,74	1,43	1,46	19,86	5,27	4,96	34,44	8,39	6,96
15 Mbits/s	C5 (M=1)	2,39	1,34	1,56	5,82	3,00	3,01	27,74	7,20	6,19
	C5 (M=2)	2,14	1,32	1,46	5,50	2,95	2,91	25,59	7,02	5,91
	C1 (M=2)	4,88	1,15	1,15	16,64	5,33	3,08	25,30	6,90	5,43
	C2 (M=2)	2,82	1,31	1,31	9,31	3,97	3,24	20,66	7,06	4,87
12 Mbits/s	C5 (M=1)	2,31	1,41	1,46	7,36	3,36	3,30	34,76	8,21	6,66
	C5 (M=2)	2,31	1,41	1,46	7,36	3,36	3,30	34,76	8,21	6,66
10 Mbits/s	C5 (M=1)	2,98	1,58	1,62	10,64	3,99	3,88	54,09	10,08	8,00
	C5 (M=2)	2,66	1,54	1,48	10,43	3,98	3,79	48,76	9,68	7,59
	C1 (M=2)	5,32	2,07	1,17	21,93	6,22	3,73	36,90	8,78	6,50
	C2 (M=2)	3,44	1,57	1,46	15,93	5,03	4,18	40,75	9,51	6,70
7,5 Mbits/s	C5 (M=1)	3,66	1,78	1,71	16,75	4,96	4,76	72,12	11,86	8,84
	C5 (M=2)	3,17	1,68	1,54	16,11	4,90	4,55	70,00	11,82	8,60
5 Mbits/s	C5 (M=1)	4,79	2,03	1,89	29,52	6,57	5,94	121,08	15,66	10,45
	C5 (M=2)	4,05	1,88	1,63	27,42	6,43	5,59	108,91	14,98	9,93
	C1 (M=1)	6,56	2,32	1,31	41,69	8,58	5,37	80,37	13,22	8,72
	C2 (M=2)	4,74	1,85	1,68	37,76	7,46	6,08	105,22	14,76	10,00
2,5 Mbits/s	C5 (M=1)	7,41	2,65	2,25	70,64	10,51	8,06	239,38	23,38	12,32
	C5 (M=2)	6,56	2,51	1,93	69,52	10,55	7,74	233,93	22,85	12,32
Sistema de Qualidade Constante	S1 (Q=4)	1,84	1,48	1,22	5,30	3,17	2,72	6,37	3,99	3,10
	S2 (Q=8)	3,20	1,88	1,52	15,78	5,05	4,48	20,71	6,82	5,30
	S3 (Q=16)	5,65	2,46	2,01	42,57	8,10	6,79	60,96	11,49	8,09
	S4 (Q=32)	9,27	3,29	2,62	103,06	12,74	9,61	164,09	18,68	11,46
	S5 (Q=62)	13,55	4,06	3,07	210,42	19,32	12,42	369,05	28,12	14,80
Conversões	NTSC	4,79	1,81	0,91	55,35	8,35	5,63	115,10	14,14	9,23
	PAL-M	5,76	1,95	1,13	94,64	11,63	7,57	191,02	18,89	11,34

Crominância

Características		Plana		Borda		Textura	
		MSE	MSE	MSE	MSE	MSE	MSE
18 Mbits/s	C4 (M=1)	0,93	1,35	7,28	8,57	22,14	20,09
	C5 (M=1)	1,21	1,55	17,42	16,83	86,91	46,78
15 Mbits/s	C5 (M=2)	1,21	1,55	17,06	16,45	84,54	44,57
	C1 (M=2)	1,29	1,61	27,29	25,83	127,96	51,06
	C2 (M=2)	1,78	1,73	25,65	23,07	125,34	50,36
12 Mbits/s	C5 (M=1)	1,22	1,55	17,87	17,30	89,97	47,54
	C5 (M=2)	1,22	1,55	17,87	17,30	89,97	47,54
10 Mbits/s	C5 (M=1)	1,39	1,78	19,37	18,97	102,11	54,96
	C5 (M=2)	1,30	1,64	18,88	18,41	97,07	51,65
	C1 (M=2)	1,36	1,72	28,52	27,17	132,46	54,21
	C2 (M=2)	1,85	1,86	27,11	24,66	133,38	55,09
7,5 Mbits/s	C5 (M=1)	1,56	1,98	21,19	20,94	111,19	59,58
	C5 (M=2)	1,42	1,78	20,52	20,23	106,43	55,99
5 Mbits/s	C5 (M=1)	1,88	2,29	24,72	24,55	131,85	68,25
	C5 (M=2)	1,63	2,01	23,29	23,18	122,20	63,11
	C1 (M=1)	1,49	1,89	32,14	31,12	147,26	62,53
	C2 (M=2)	2,00	2,13	30,77	28,52	154,55	65,33
2,5 Mbits/s	C5 (M=1)	2,64	2,88	33,43	32,74	173,01	83,96
	C5 (M=2)	2,28	2,52	31,48	31,05	163,71	80,37
Sistema de Qualidade Constante	S1 (Q=4)	1,38	1,74	19,82	19,73	79,08	42,47
	S2 (Q=8)	1,98	2,28	25,12	25,30	97,74	54,21
	S3 (Q=16)	2,86	3,05	33,81	33,81	131,24	70,64
	S4 (Q=32)	4,40	3,70	46,03	43,96	184,96	89,55
	S5 (Q=62)	6,14	4,49	59,85	53,34	261,87	109,16
Conversões	NTSC	2,24	1,84	34,92	32,22	161,09	60,27
	PAL-M	2,36	2,01	43,06	37,94	207,53	71,63

TABELA 9.3.2: Base de Dados Objetivos da Cena "Mobile & Calendar"**Luminância**

Características		Plana			Borda			Textura		
		MSE	PSD	NSD	MSE	PSD	NSD	MSE	PSD	NSD
18 Mbits/s	C4 (M=1)	13,46	2,33	3,24	48,65	6,64	7,49	66,69	8,96	8,66
15 Mbits/s	C5 (M=1)	9,36	2,16	2,86	16,37	4,25	4,85	39,82	7,19	6,48
	C5 (M=2)	8,19	2,09	2,60	13,68	3,93	4,41	33,97	6,64	6,14
	C1 (M=2)	7,48	2,46	2,46	21,43	4,71	4,85	33,66	6,79	6,14
	C2 (M=2)	8,53	2,49	2,49	20,14	5,03	4,51	36,57	7,34	5,87
12 Mbits/s	C5 (M=1)	9,93	2,26	2,85	18,84	4,54	5,12	43,26	7,41	6,74
	C5 (M=2)	9,93	2,26	2,85	18,84	4,54	5,12	43,26	7,41	6,74
10 Mbits/s	C5 (M=1)	14,13	2,61	3,43	33,12	5,88	6,68	73,30	9,69	8,10
	C5 (M=2)	12,22	2,47	3,14	26,92	5,35	6,01	58,90	8,60	7,60
	C1 (M=2)	10,57	2,30	2,75	35,00	5,92	6,05	54,46	8,47	7,43
	C2 (M=2)	11,83	2,35	2,98	36,06	6,38	6,05	67,00	9,38	7,61
7,5 Mbits/s	C5 (M=1)	18,41	3,03	3,79	52,85	7,38	8,06	112,22	12,08	9,24
	C5 (M=2)	15,67	2,84	3,45	42,47	6,66	7,19	88,53	10,53	8,76
5 Mbits/s	C5 (M=1)	25,36	3,71	4,28	93,77	9,90	10,04	180,75	15,79	10,46
	C5 (M=2)	21,33	3,47	3,88	73,46	8,71	9,09	145,57	13,54	10,27
	C1 (M=1)	18,80	2,99	3,62	84,35	8,99	9,11	128,26	12,69	10,28
	C2 (M=2)	20,94	3,07	4,05	92,70	9,71	9,59	166,37	13,99	11,08
2,5 Mbits/s	C5 (M=1)	36,65	4,65	4,92	182,84	14,45	12,79	263,08	19,93	12,24
	C5 (M=2)	35,49	4,76	4,69	175,42	13,87	12,88	252,98	18,78	12,72
Sistema de Qualidade Constante	S1 (Q=4)	4,46	1,84	1,89	6,64	3,03	3,03	6,72	3,57	3,12
	S2 (Q=8)	9,77	2,42	2,67	22,55	4,94	5,36	23,07	5,93	5,45
	S3 (Q=16)	19,59	3,51	3,57	68,25	8,19	8,84	70,81	9,67	8,61
	S4 (Q=32)	35,73	4,80	4,61	183,27	13,65	13,48	191,02	15,47	12,26
	S5 (Q=62)	55,22	6,15	5,61	386,44	21,15	18,07	401,87	23,66	15,28
Conversões	NTSC	13,00	1,92	2,69	163,34	9,58	10,18	172,22	11,94	9,33
	PAL-M	23,23	2,28	4,14	266,74	13,35	13,89	271,69	16,00	11,48

Crominância

Características		Plana		Borda		Textura	
		MSE	MSE	MSE	MSE	MSE	MSE
18 Mb/s	C4 (M=1)	5,17	4,20	19,77	17,95	21,29	18,12
15 Mb/s	C5 (M=1)	8,47	7,76	41,79	38,82	38,11	32,82
	C5 (M=2)	7,55	6,87	39,73	36,90	35,82	30,55
	C1 (M=2)	8,45	9,21	77,28	76,57	53,59	49,90
	C2 (M=2)	9,42	9,93	78,72	82,43	53,47	50,36
12 Mb/s	C5 (M=1)	8,17	7,53	42,37	39,45	37,94	32,59
	C5 (M=2)	8,17	7,53	42,37	39,45	37,94	32,59
10 Mb/s	C5 (M=1)	10,55	9,93	49,67	46,67	45,19	39,82
	C5 (M=2)	8,89	8,26	45,93	43,06	41,12	35,73
	C1 (M=2)	9,27	10,00	80,55	79,63	57,03	53,22
	C2 (M=2)	10,52	10,99	82,62	86,12	57,69	54,71
7,5 Mb/s	C5 (M=1)	12,39	11,91	57,42	54,46	51,89	46,35
	C5 (M=2)	10,16	9,49	51,53	48,54	45,72	40,47
5 Mb/s	C5 (M=1)	15,71	15,60	71,96	69,04	62,67	57,29
	C5 (M=2)	12,36	11,97	60,96	58,49	53,84	48,54
	C1 (M=1)	11,94	12,39	89,97	88,53	66,08	62,53
	C2 (M=2)	13,74	13,87	96,62	100,25	69,36	66,54
2,5 Mb/s	C5 (M=1)	22,81	22,55	98,87	96,85	79,45	73,80
	C5 (M=2)	19,77	19,37	91,22	88,12	73,97	67,62
Sistema de Qualidade Constante	S1 (Q=4)	7,89	7,15	42,47	39,18	33,66	27,67
	S2 (Q=8)	11,22	10,43	55,86	52,37	43,16	36,90
	S3 (Q=16)	17,34	16,56	81,11	77,64	59,99	53,71
	S4 (Q=32)	26,86	27,23	122,77	123,05	86,91	81,11
	S5 (Q=62)	43,66	45,82	189,27	194,57	126,50	121,64
Conversões	NTSC	15,35	11,64	104,01	100,02	72,12	60,41
	PAL-M	17,42	18,03	133,38	141,93	85,92	84,93

TABELA 9.3.3: Base de Dados Objetivos da Cena "Table Tennis"

Luminância

Características		Plana			Borda			Textura		
		MSE	PSD	NSD	MSE	PSD	NSD	MSE	PSD	NSD
18 Mbits/s	C4 (M=1)	6,67	2,32	2,17	14,03	4,15	3,93	10,79	3,37	3,28
15 Mbits/s	C5 (M=1)	5,89	2,21	2,27	5,82	2,75	2,86	9,27	3,14	3,27
	C5 (M=2)	5,72	2,18	2,26	5,64	2,75	2,76	8,96	3,10	3,24
	C1 (M=2)	5,08	1,99	1,99	10,38	3,70	3,06	7,71	2,99	2,84
	C2 (M=2)	5,76	1,85	1,85	6,76	3,37	2,55	7,91	3,15	2,63
12 Mbits/s	C5 (M=1)	7,30	2,37	2,33	7,50	3,11	3,09	11,33	3,41	3,40
	C5 (M=2)	7,30	2,37	2,33	7,50	3,11	3,09	11,33	3,41	3,40
10 Mbits/s	C5 (M=1)	9,31	2,57	2,40	10,50	3,53	3,62	14,97	3,72	3,72
	C5 (M=2)	9,08	2,55	2,37	10,16	3,56	3,49	14,56	3,68	3,68
	C1 (M=2)	7,55	2,33	2,24	15,35	4,45	3,72	11,70	3,41	3,28
	C2 (M=2)	8,30	2,60	2,01	11,15	4,03	3,20	12,39	3,60	3,13
7,5 Mbits/s	C5 (M=1)	11,78	2,74	2,49	14,56	4,07	4,11	20,23	4,11	4,05
	C5 (M=2)	11,43	2,73	2,42	13,87	4,06	3,95	19,46	4,05	3,97
5 Mbits/s	C5 (M=1)	14,59	3,03	2,54	20,80	4,86	4,74	26,43	4,54	4,43
	C5 (M=2)	14,29	3,05	2,43	19,82	4,81	4,52	26,01	4,60	4,24
	C1 (M=1)	10,97	2,88	2,11	24,16	5,65	4,42	18,80	4,23	3,60
	C2 (M=2)	12,62	2,91	2,26	21,63	5,26	4,41	21,78	4,32	3,86
2,5 Mbits/s	C5 (M=1)	20,56	4,16	2,69	36,06	6,48	6,03	42,37	5,82	5,10
	C5 (M=2)	20,94	4,61	2,27	39,27	6,95	6,09	44,78	6,37	4,85
Sistema de Qualidade Constante	S1 (Q=4)	5,07	2,26	1,91	5,81	3,03	2,62	5,86	2,88	2,50
	S2 (Q=8)	11,33	2,85	2,26	16,15	4,65	4,01	16,11	3,91	3,47
	S3 (Q=16)	19,01	3,96	2,44	39,27	7,24	5,88	35,08	5,32	4,41
	S4 (Q=32)	25,77	6,92	1,77	91,85	11,86	8,64	62,82	9,79	4,43
	S5 (Q=62)	28,45	8,35	1,30	189,27	18,57	11,92	95,08	14,50	3,96
Conversões	NTSC	6,58	2,24	1,70	73,80	9,08	5,30	26,13	4,33	3,26
	PAL-M	9,84	2,70	1,70	117,24	12,14	6,86	42,08	5,62	3,62

Crominância

Características		Plana		Borda		Textura	
		MSE	MSE	MSE	MSE	MSE	MSE
18 Mbits/s	C4 (M=1)	1,61	1,04	5,62	5,56	2,75	2,51
15 Mbits/s	C5 (M=1)	2,20	1,42	17,54	16,60	4,92	5,50
	C5 (M=2)	2,18	1,40	17,38	16,45	4,84	5,41
	C1 (M=2)	2,25	1,52	26,37	25,83	5,56	8,00
	C2 (M=2)	2,33	2,13	25,53	26,07	5,57	8,75
12 Mbits/s	C5 (M=1)	2,23	1,46	17,99	17,10	5,08	5,76
	C5 (M=2)	2,23	1,46	17,99	17,10	5,08	5,76
10 Mbits/s	C5 (M=1)	2,32	1,59	19,10	18,24	5,48	6,33
	C5 (M=2)	2,29	1,54	18,84	17,95	5,40	6,24
	C1 (M=2)	2,27	1,53	27,42	26,92	5,92	8,36
	C2 (M=2)	2,39	2,24	26,31	26,92	5,88	9,14
7,5 Mbits/s	C5 (M=1)	2,40	1,68	20,28	19,50	5,99	7,06
	C5 (M=2)	2,34	1,63	19,96	19,15	5,84	6,94
5 Mbits/s	C5 (M=1)	2,48	1,81	21,88	21,29	6,61	7,89
	C5 (M=2)	2,43	1,76	21,53	20,94	6,50	7,87
	C1 (M=1)	2,31	1,63	29,05	28,71	6,38	9,12
	C2 (M=2)	2,60	2,59	28,58	29,25	6,67	10,43
2,5 Mbits/s	C5 (M=1)	2,69	2,10	25,89	26,07	7,75	9,82
	C5 (M=2)	2,68	2,10	26,37	26,80	7,89	10,35
Sistema de Qualidade Constante	S1 (Q=4)	2,29	1,55	20,42	19,59	5,21	6,00
	S2 (Q=8)	2,56	1,86	25,89	25,77	6,56	8,26
	S3 (Q=16)	2,87	2,28	35,90	37,59	8,47	12,00
	S4 (Q=32)	3,00	2,90	49,21	54,59	10,64	17,42
	S5 (Q=62)	3,36	3,81	64,73	73,63	13,97	26,31
Conversões	NTSC	1,95	1,24	38,91	39,64	6,82	10,99
	PAL-M	2,00	1,43	43,76	47,32	7,41	13,87

TABELA 9.3.4: Base de Dados Objetivos da Cena "Diva with Noise"

Luminância

Características		Plana			Borda			Textura		
		MSE	PSD	NSD	MSE	PSD	NSD	MSE	PSD	NSD
18 Mbits/s	C4 (M=1)	8,36	2,80	2,80	21,93	5,66	5,25	43,96	7,38	6,86
15 Mbits/s	C5 (M=1)	7,73	2,75	2,68	11,35	4,14	3,97	38,11	6,37	5,33
	C5 (M=2)	7,85	2,81	2,63	11,83	4,33	3,96	37,76	6,28	5,41
	C1 (M=2)	7,08	2,45	2,45	14,86	4,99	4,16	30,77	6,73	5,23
	C2 (M=2)	6,98	2,46	2,46	12,16	4,59	3,49	39,18	7,37	4,83
12 Mbits/s	C5 (M=1)	9,08	3,07	2,72	15,14	4,94	4,46	48,87	7,07	5,86
	C5 (M=2)	9,08	3,07	2,72	15,14	4,94	4,46	48,87	7,07	5,86
10 Mbits/s	C5 (M=1)	11,15	3,52	2,95	20,28	5,86	5,17	65,78	8,33	6,39
	C5 (M=2)	10,94	3,51	2,84	20,37	5,90	5,11	67,16	8,41	6,37
	C1 (M=2)	9,96	3,33	2,84	25,59	6,70	5,34	58,62	8,80	6,53
	C2 (M=2)	9,77	3,19	2,88	20,33	5,91	4,54	74,49	9,35	5,99
7,5 Mbits/s	C5 (M=1)	14,19	4,18	3,16	29,86	7,39	6,13	95,08	10,31	6,93
	C5 (M=2)	13,87	4,15	3,04	30,14	7,41	6,06	97,74	10,61	6,81
5 Mbits/s	C5 (M=1)	18,84	5,27	3,33	47,98	9,83	7,43	140,95	14,29	7,08
	C5 (M=2)	18,20	5,18	3,18	49,33	9,93	7,31	141,93	14,60	6,94
	C1 (M=1)	16,91	4,77	3,36	57,95	10,69	7,54	148,28	14,13	7,55
	C2 (M=2)	16,64	4,33	3,60	47,22	9,25	6,86	174,61	14,80	7,04
2,5 Mbits/s	C5 (M=1)	27,23	7,39	3,34	94,42	15,12	9,36	203,74	19,10	7,88
	C5 (M=2)	26,55	7,47	3,24	94,21	15,09	9,30	201,41	19,28	7,59
Sistema de Qualidade Constante	S1 (Q=4)	3,48	2,01	1,85	4,91	2,94	2,51	5,68	3,53	2,70
	S2 (Q=8)	7,21	2,83	2,43	13,71	4,81	4,08	18,20	5,76	4,39
	S3 (Q=16)	13,49	4,15	3,13	36,40	8,14	6,44	54,96	9,14	6,50
	S4 (Q=32)	23,18	6,41	3,81	87,51	13,67	9,64	147,26	14,31	8,47
	S5 (Q=62)	35,00	9,26	4,39	167,14	21,18	12,69	269,20	25,28	7,61
Conversões	NTSC	9,62	1,94	1,67	21,88	5,93	4,15	35,16	7,21	4,15
	PAL-M	12,53	2,50	2,34	46,03	8,52	6,60	78,54	10,89	5,40

Crominância

Características		Plana		Borda		Textura	
		MSE	MSE	MSE	MSE	MSE	MSE
18 Mbits/s	C4 (M=1)	3,53	4,48	12,00	12,48	24,38	22,34
15 Mbits/s	C5 (M=1)	4,43	6,31	18,45	19,41	43,16	33,89
	C5 (M=2)	4,34	6,27	18,54	19,50	43,96	34,92
	C1 (M=2)	4,27	6,24	31,05	32,51	53,84	39,27
	C2 (M=2)	4,01	6,73	29,25	32,51	57,82	42,86
12 Mbits/s	C5 (M=1)	4,74	6,92	20,47	21,58	50,83	39,73
	C5 (M=2)	4,74	6,92	20,47	21,58	50,83	39,73
10 Mbits/s	C5 (M=1)	5,56	8,17	23,50	24,78	58,62	44,17
	C5 (M=2)	5,36	7,96	23,34	24,72	61,53	46,03
	C1 (M=2)	4,93	7,53	35,00	36,74	67,62	50,71
	C2 (M=2)	4,65	7,87	32,37	35,73	76,22	55,35
7,5 Mbits/s	C5 (M=1)	6,53	9,71	28,25	29,86	73,13	52,37
	C5 (M=2)	6,27	9,46	28,25	29,86	77,28	54,84
5 Mbits/s	C5 (M=1)	7,89	12,11	36,90	39,09	92,49	62,82
	C5 (M=2)	7,64	11,56	37,16	39,36	96,62	65,33
	C1 (M=1)	6,70	10,74	47,32	49,44	109,67	78,18
	C2 (M=2)	6,04	10,47	41,79	44,88	117,78	75,35
2,5 Mbits/s	C5 (M=1)	10,02	16,41	54,96	58,09	115,63	79,08
	C5 (M=2)	9,75	15,82	54,84	57,95	118,87	79,81
Sistema de Qualidade Constante	S1 (Q=4)	4,38	6,05	19,77	20,80	29,79	23,12
	S2 (Q=8)	6,11	9,08	31,12	32,44	48,99	39,64
	S3 (Q=16)	8,24	12,86	49,90	52,49	89,97	64,73
	S4 (Q=32)	11,43	18,45	71,79	76,75	124,76	80,37
	S5 (Q=62)	16,11	27,74	99,33	108,91	137,11	91,22
Conversões	NTSC	4,66	7,25	43,16	46,03	78,54	38,11
	PAL-M	5,20	8,98	52,73	58,90	79,08	49,55

TABELA 9.3.5: Base de Dados Objetivos da Cena "Kielharbour"

Luminância

Características		Plana			Borda			Textura		
		MSE	PSD	NSD	MSE	PSD	NSD	MSE	PSD	NSD
18 Mbits/s	C4 (M=1)	11,15	3,28	3,24	41,31	7,69	6,67	63,25	10,40	8,59
15 Mbits/s	C5 (M=1)	7,50	2,95	2,76	21,24	5,65	5,09	57,42	9,97	8,51
	C5 (M=2)	6,64	2,82	2,64	19,06	5,37	4,81	51,89	9,60	8,11
	C1 (M=2)	7,02	2,41	2,41	34,76	8,51	4,31	42,76	8,63	6,63
	C2 (M=2)	8,24	2,69	2,69	23,66	6,36	4,63	46,14	9,45	6,56
12 Mbits/s	C5 (M=1)	8,69	3,11	2,87	27,61	6,40	5,55	72,12	11,32	9,06
	C5 (M=2)	8,69	3,11	2,87	27,61	6,40	5,55	72,12	11,32	9,06
10 Mbits/s	C5 (M=1)	13,34	3,79	3,23	42,27	7,89	6,62	105,94	13,65	10,40
	C5 (M=2)	12,19	3,61	3,16	39,82	7,63	6,40	101,64	13,23	10,25
	C1 (M=2)	11,97	3,71	2,94	53,71	10,28	5,64	77,28	11,75	8,54
	C2 (M=2)	14,19	3,74	3,30	47,22	8,50	6,31	100,25	13,34	9,13
7,5 Mbits/s	C5 (M=1)	18,37	4,45	3,42	63,11	9,68	7,60	159,99	16,93	11,76
	C5 (M=2)	17,54	4,32	3,44	60,27	9,32	7,46	153,49	16,35	11,66
5 Mbits/s	C5 (M=1)	26,67	5,58	3,41	105,70	12,68	8,96	266,12	22,53	13,22
	C5 (M=2)	25,30	5,34	3,50	101,88	12,23	8,90	257,68	21,11	13,61
	C1 (M=1)	25,24	5,37	3,45	116,97	14,34	8,16	194,12	18,06	11,99
	C2 (M=2)	28,78	5,17	4,04	124,19	13,08	9,47	269,20	20,71	13,61
2,5 Mbits/s	C5 (M=1)	46,57	8,65	2,73	225,99	20,45	10,61	429,62	29,94	14,38
	C5 (M=2)	48,65	9,12	2,84	237,73	20,06	11,63	436,60	28,49	15,82
Sistema de Qualidade Constante	S1 (Q=4)	4,77	2,62	2,18	6,56	3,72	2,90	7,23	4,16	3,21
	S2 (Q=8)	12,68	3,66	3,21	22,86	6,32	4,99	27,36	7,51	5,69
	S3 (Q=16)	27,61	5,42	4,01	69,52	10,42	7,87	92,06	13,00	9,27
	S4 (Q=32)	47,00	8,44	4,05	177,86	16,75	11,20	260,06	21,35	13,32
	S5 (Q=62)	64,28	11,52	3,75	359,82	25,71	14,49	572,90	32,40	17,25
Conversões	NTSC	10,05	2,93	2,41	89,97	10,52	5,83	165,99	14,45	9,55
	PAL-M	15,10	3,50	2,65	135,86	13,10	7,08	259,47	19,16	11,24

Crominância

Características		Plana		Borda		Textura	
		MSE	MSE	MSE	MSE	MSE	MSE
18 Mbits/s	C4 (M=1)	1,74	1,53	8,42	7,48	11,12	8,51
15 Mbits/s	C5 (M=1)	2,30	2,05	12,56	11,04	19,77	15,14
	C5 (M=2)	2,20	1,93	12,02	10,50	18,41	13,87
	C1 (M=2)	2,02	1,85	13,93	12,14	18,67	13,87
	C2 (M=2)	2,80	1,85	15,21	13,12	19,50	14,29
12 Mbits/s	C5 (M=1)	2,36	2,06	13,40	11,70	20,19	14,90
	C5 (M=2)	2,36	2,06	13,40	11,70	20,19	14,90
10 Mbits/s	C5 (M=1)	2,77	2,47	15,63	13,68	23,45	17,50
	C5 (M=2)	2,58	2,24	14,97	13,03	21,78	16,11
	C1 (M=2)	2,23	2,05	16,37	14,19	21,29	15,78
	C2 (M=2)	3,12	2,15	17,95	15,56	22,29	16,33
7,5 Mbits/s	C5 (M=1)	3,21	2,81	17,91	15,63	25,77	18,67
	C5 (M=2)	2,88	2,48	17,10	14,86	23,88	17,38
5 Mbits/s	C5 (M=1)	3,65	3,38	21,78	18,97	27,99	20,09
	C5 (M=2)	3,22	2,81	20,56	17,75	26,31	18,75
	C1 (M=1)	2,81	2,52	21,58	18,45	25,83	18,45
	C2 (M=2)	3,74	2,69	23,23	20,09	26,80	19,50
2,5 Mbits/s	C5 (M=1)	4,79	4,44	30,84	26,01	30,70	21,58
	C5 (M=2)	4,61	3,82	30,63	25,89	30,00	20,80
Sistema de Qualidade Constante	S1 (Q=4)	2,39	2,05	12,53	10,89	14,79	11,25
	S2 (Q=8)	3,11	2,62	17,95	15,56	20,05	15,00
	S3 (Q=16)	4,17	3,37	25,18	21,48	26,25	18,54
	S4 (Q=32)	5,56	3,98	33,89	28,38	29,86	20,99
	S5 (Q=62)	6,82	4,46	41,79	33,97	31,34	21,68
Conversões	NTSC	3,52	2,18	20,05	15,71	19,15	12,14
	PAL-M	3,71	2,31	22,44	18,12	21,93	14,46

9.4 Resultados de Estimação de Qualidade Subjetiva com base em Medidas Subjetivas

TABELA 9.4.1: Pesos dos Estimadores

	Luminância	Crominância	Plana	Borda	Textura
Flower Garden	0,28	0,72	0,04	0,42	0,54
Mobile & Calendar	0,55	0,45	0,07	0,57	0,36
Table Tennis	0,67	0,33	0,53	0,12	0,34
Diva with Noise	0,36	0,64	0,40	0,40	0,20
Kielharbour	0,55	0,45	0,35	0,45	0,20

TABELA 9.4.2: Erro Quadrático Médio

As tabelas a seguir fornecem o erro quadrático médio entre as notas calculadas por meio das medidas objetivas propostas e os resultados da avaliação subjetiva. As classes citadas seguem a definição da Tabela 17.

Resultados obtidos pela aplicação dos estimadores na imagem segmentada

Treino	Classe 1	Classe 2	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	4,75	0,55	24,18	11,16	10,16	18,37
Mobile & Calendar	11,71	1,86	15,87	7,71	9,29	13,53
Table Tennis	25,12	1,82	114,82	1,49	35,81	83,61
Diva with Noise	8,96	6,54	9,51	32,00	14,25	10,93
Kielharbour	3,13	3,20	48,21	0,29	13,71	34,13

Teste	Classe 1	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	15,13	10,91	104,77	10,91	43,61
Mobile & Calendar	3,83	8,47	72,37	8,47	28,22
Table Tennis	1,51	10,47	2,44	10,47	4,81
Diva with Noise	8,59	2,06	23,09	2,06	11,25
Kielharbour	35,04	23,54	0,16	23,54	19,58

Resultados obtidos pela aplicação dos estimadores em toda a imagem (sem segmentação)

Treino	Classe 1	Classe 2	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	4,74	0,02	28,89	8,25	10,47	21,37
Mobile & Calendar	82,40	0,22	247,26	33,17	90,76	186,43
Table Tennis	3,03	0,67	80,24	28,88	28,21	58,29
Diva with Noise	3,13	8,70	27,66	33,50	18,25	22,88
Kielharbour	3,60	2,01	50,06	0,54	14,05	35,40

Teste	Classe 1	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	5,30	29,44	40,58	29,44	25,11
Mobile & Calendar	25,54	42,15	11,47	42,15	26,39
Table Tennis	26,44	9,67	27,15	9,67	21,09
Diva with Noise	381,72	12,68	42,13	12,68	145,51
Kielharbour	449,22	18,23	1,68	18,23	156,38

Resultados obtidos utilizando somente o medidor de erro quadrático médio

Treino	Classe 1	Classe 2	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	17,21	10,15	75,89	39,62	35,72	59,01
Mobile & Calendar	81,74	0,46	233,09	30,86	86,54	176,35
Table Tennis	23,89	7,18	331,91	83,98	111,74	240,47
Diva with Noise	4,25	7,74	26,78	48,71	21,87	23,54
Kielharbour	3,28	7,04	57,29	1,73	17,34	40,84

Teste	Classe 1	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	13,38	25,87	39,46	25,87	26,24
Mobile & Calendar	108,82	14,50	9,41	14,50	44,24
Table Tennis	24,82	12,38	14,25	12,38	17,15
Diva with Noise	357,98	5,95	45,28	5,95	136,40
Kielharbour	416,76	28,68	1,23	28,68	148,89

Tabela 9.4.3: Erro Absoluto Médio

As tabelas a seguir fornecem o erro percentual absoluto entre as notas calculadas por meio das medidas objetivas propostas e os resultados da avaliação subjetiva. As classes citadas seguem a definição da Tabela 17.

Resultados obtidos pela aplicação dos estimadores na imagem segmentada

Treino	Classe 1	Classe 2	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	2,18	0,74	3,85	3,34	2,53	3,31
Mobile & Calendar	2,88	1,36	3,01	2,78	2,51	2,84
Table Tennis	4,33	1,35	8,68	1,22	3,89	6,87
Diva with Noise	2,94	2,56	2,16	5,66	3,33	2,58
Kielharbour	1,75	1,79	6,43	0,54	2,63	4,90

Teste	Classe 1	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	3,71	2,84	10,24	5,59	3,61
Mobile & Calendar	1,88	2,56	8,51	4,32	2,86
Table Tennis	1,16	2,26	1,56	1,66	1,95
Diva with Noise	2,21	1,24	4,81	2,75	1,74
Kielharbour	5,08	4,02	0,40	3,17	3,99

Resultados obtidos pela aplicação dos estimadores em toda a imagem (sem segmentação)

Treino	Classe 1	Classe 2	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	1,96	0,14	3,78	2,87	2,19	3,15
Mobile & Calendar	6,98	0,47	14,31	5,76	6,88	11,46
Table Tennis	1,58	0,82	8,19	5,37	3,99	6,39
Diva with Noise	1,26	2,95	3,80	5,79	3,45	3,50
Kielharbour	1,86	1,42	6,67	0,74	2,67	5,07

Teste	Classe 1	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	1,69	3,01	6,37	3,69	2,97
Mobile & Calendar	4,18	3,72	3,39	3,76	3,80
Table Tennis	3,58	1,98	5,21	3,59	2,59
Diva with Noise	13,15	2,59	6,49	7,41	5,33
Kielharbour	16,76	3,81	1,30	7,29	6,61

Resultados obtidos utilizando somente o medidor de erro quadrático médio

Treino	Classe 1	Classe 2	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	3,68	3,19	6,29	6,29	4,86	5,65
Mobile & Calendar	6,70	0,68	13,55	5,55	6,62	10,89
Table Tennis	4,84	2,68	15,25	9,16	7,98	12,21
Diva with Noise	1,57	2,78	4,11	6,98	3,86	3,84
Kielharbour	1,78	2,65	6,81	1,32	3,14	5,29

Teste	Classe 1	Classe 3	Classe 4	Média por Classe	Média Geral
Flower Garden	2,78	3,21	6,28	4,09	3,35
Mobile & Calendar	8,43	3,26	3,07	4,92	4,44
Table Tennis	3,48	2,57	3,77	3,28	2,88
Diva with Noise	13,01	2,05	6,73	7,26	4,94
Kielharbour	16,38	4,37	1,11	7,29	6,89

TABELA 9.4.4: Erro Máximo

As tabelas a seguir fornecem o erro máximo entre as notas calculadas por meio das medidas objetivas propostas e os resultados da avaliação subjetiva, para cada uma das três estratégias de aplicação dos estimadores.

Treino	Com Segmentação	Sem Segmentação	Somente MSE
Flower Garden	8,90	10,90	17,19
Mobile & Calendar	8,81	27,55	24,47
Table Tennis	21,20	12,67	34,72
Diva with Noise	6,92	11,88	11,88
Kielharbour	11,75	12,38	14,64

Teste	Com Segmentação	Sem Segmentação	Somente MSE
Flower Garden	10,24	15,49	13,99
Mobile & Calendar	8,51	16,65	14,56
Table Tennis	7,94	8,79	8,45
Diva with Noise	4,83	33,32	32,07
Kielharbour	9,08	33,80	32,06