



UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
FACULDADE DE ENGENHARIA DA COMPUTAÇÃO E
TELECOMUNICAÇÕES

**Quantização inter adaptativa baseada no sistema visual humano:
análise e aplicações no MPEG-1.**

Luan Assis Gonçalves

BELEM - PARÁ

2016



UNIVERSIDADE FEDERAL DO PARÁ
INSTITUTO DE TECNOLOGIA
FACULDADE DE ENGENHARIA DA COMPUTAÇÃO E
TELECOMUNICAÇÕES

Luan Assis Gonçalves

**Quantização inter adaptativa baseada no sistema visual humano:
análise e aplicações no MPEG-1.**

Trabalho de Conclusão de Curso apresentado para
obtenção do grau de Engenheiro em Engenharia
da Computação, do Instituto de Tecnologia,
da Faculdade de Engenharia da Computação e
Telecomunicações.

BELÉM - PARÁ

2016

**Quantização inter adaptativa baseada no sistema visual humano:
análise e aplicações no MPEG-1.**

Este trabalho foi julgado adequado em xx/xx/2016 para a obtenção do Grau de Engenheiro da Computação, aprovado em sua forma final pela banca examinadora que atribui o conceito _____.

Prof. Dr. Ronaldo de Freitas Zampolo
ORIENTADOR

Prof. Dr. Evaldo Gonçalves Pelaes
MEMBRO DA BANCA EXAMINADORA

Prof^a. Dr^a. Valquíria Gusmão Macedo
MEMBRO DA BANCA EXAMINADORA

Prof. Dr. Francisco Carlos Bentes Frey Müller
DIRETOR DA FACULDADE DE ENGENHARIA
DA COMPUTAÇÃO E TELECOMUNICAÇÕES

Aos anjos em minha vida
Helena e Antônio.

Agradecimentos



*"Qualquer um que nunca tenha cometido um erro nunca
tentou algo novo."*
Albert Einstein

Resumo



Palavras-chave: Compressão, Quantização, Codificação perceptual.

Abstract

Keywords: Compression, Quantization, Perceptual Coding.

Lista de Figuras

2.1	Reduções da fonte [1].	19
2.2	Codificação da fonte reduzida [1].	19
2.3	Fluxograma do codificador JPEG.	22
2.4	Fluxograma do decodificador JPEG.	22
2.5	Compensação de movimento pra frente.	25
2.6	Compensação de movimento bidirecional.	25
2.7	Fluxograma do codificador MPEG.	27
2.8	Sequência de imagens dentro de um GOP.	28
2.9	Fluxograma do decodificador MPEG.	28
3.1	Distribuição da frequência espacial no domínio da DCT.	35
3.2	Visão perspectiva da superfície de limiar espacial e temporal [2]. Cada curva representa a resposta a frequência espacial em uma frequência espacial fixa. . .	35
4.1	MSSIM.	39
4.2	Codec padrão vs Codec perceptual.	40
4.3	Codec padrão vs Codec perceptual.	41


Lista de Tabelas

2.1	Exemplo de codificação aritmética.	20
2.2	Tabela de quantização padrão.	23
2.3	Padrão zigzag.	23
2.4	Exemplo de subimagem.	23
2.5	Subimagem da tabela 2.4 quantizada.	23
4.1	Descrição dos vídeos.	37

Lista de Abreviaturas e Siglas

CD Compact Disc

DCT Discrete Cosine Transform

DPCM Differential pulse-code modulation 

DVD Digital Versatile Disc

HVS Human Visual System

IDCT Inverse Discrete Cosine Transform

ITU International Telecommunication Union

JPEG Joint Photographic Experts Group

MSE Mean Squared Error

MSSIM Mean Structural Similarity Index

MOS Mean Opinion Score

MPEG Moving Picture Experts Group

PSNR Peak Signal-to-Noise-Ratio

SNR Signal-to-Noise Ratio

SSIM Structural Similarity Index

TV Television

Sumário

1	Introdução	14
1.1	Motivação	14
1.2	Visão geral do trabalho	14
1.3	Organização do trabalho	15
2	Fundamentos de compressão de imagens e vídeos	16
2.1	Introdução	16
2.2	Redundância	16
2.3	Teoria da Informação	17
2.3.1	Primeiro teorema de Shannon	18
2.3.2	Entropia	18
2.4	Alguns métodos básicos de compressão	18
2.4.1	Métodos	19
2.5	Os padrões JPEG e MPEG	21
2.5.1	JPEG	21
2.5.2	MPEG-1	24
3	Qualidade visual	30
3.1	Introdução	30
3.2	Artefatos provenientes do processo de compressão	30
3.3	Avaliações subjetiva e objetiva	31
3.4	Classificação das métricas objetivas de avaliação de qualidade visual	32
3.4.1	Não perceptuais	33
3.4.2	Perceptuais	33
3.5	Quantização interadaptativa baseada no sistema visual humano	34

4	Procedimento experimental e resultados	37
4.1	Implementação do sistema	37
4.2	Conjunto de teste	37
4.3	Metodologia	38
4.4	Resultados	38
5	Considerações finais	42
5.1	Conclusões	42
5.2	Trabalhos futuros	42
	Referências Bibliográficas	43

Capítulo 1

Introdução

Motivação

Com a expansão de tecnologias como a **TV digital, streaming de vídeo e DVD (do inglês Digital Versatile Disc)**, a compressão de vídeo tornou-se elemento indispensável para as transmissões broadcast e mídias de entretenimento. Neste contexto, pesquisas relacionadas **à área** foram potencializadas a fim de otimizar a utilização da banda de transmissão, resultando no surgimento de novos padrões de compressão de imagens e vídeos.

Embora o poder de processamento computacional, armazenamento e a largura de banda disponível para transmissão de dados tenham evoluído, ainda assim, a necessidade de métodos que possam prover taxas consideráveis de compressão e baixas taxas de distorção é um desafio. Esta realidade é evidenciada em trabalhos como [3], [4], [5] e [6].

A situação apresentada anteriormente potencializa-se com a expansão dos dispositivos móveis. Em [7] é feita uma projeção de que nos **próximos 5 anos o** planeta terá mais dispositivos móveis conectados à internet do que habitantes, gerando um fluxo de cerca de 10 imagens e 1 vídeo por habitante ao dia.

Visão geral do trabalho


Na literatura há trabalhos que abordam de maneira profunda as bases teóricas da compressão de imagens e vídeos, enquanto outros tem como foco principal os aspectos práticos e de implementação. Na fase de graduação observa-se que os discentes muitas vezes são expostos a conceitos teóricos de teoria de informação e processamento de sinais e poucas vezes tem

contato com suas aplicações.

Portanto, este trabalho tem como objetivo abordar teoria e prática, fundamentando os conceitos de compreensão de imagens e vídeos através de implementações dos padrões *JPEG* (do inglês Joint Photographics Experts Group) e *MPEG-1* (do inglês Moving Picture Experts Group), bem como abordar os aspectos de qualidade visual.

Organização do trabalho

O restante deste documento está organizado da seguinte forma:

- Capítulo 2: são apresentados conceitos básicos relacionados à compressão de imagens e vídeos ~~e suas áreas afins, como teoria de informação e processamento de sinais. Por fim, uma visão geral dos padrões JPEG e MPEG-1 é apresentada;~~ , com ênfase nos
- Capítulo 3: neste capítulo, o foco é a avaliação da qualidade visual de imagens e vídeos depois do processo de descompressão ~~bem como~~ o apresentar um método que visa a melhoria da qualidade subjetiva, através da geração dinâmica das tabelas de quantização baseadas no sistema visual humano, proposto por [8], que será ~~analisado~~ neste trabalho. 
- Capítulo 4: ~~defini-se~~ o procedimento experimental e, por fim, os resultados obtidos são apresentados e discutidos;
- Capítulo 5: a partir da análise detalhada dos resultados, serão apresentadas as conclusões da técnica analisada, destacando suas vantagens e desvantagens.

Capítulo 2

Fundamentos ^{da} ~~de~~ compressão de imagens e vídeos

Introdução


Neste capítulo abordam-se as bases teóricas do processo de compressão de imagens e vídeos. Primeiramente, o conceito de redundância, presente em arrays 2D, será abordado, fazendo-se a relação com a teoria da informação. Em seguida, é discutida uma possível classificação dos tipos de compressão quanto a preservação do sinal original. Por fim, analisa-se o funcionamento dos padrões JPEG e MPEG-1.


Redundância

O processo de compressão de dados consiste em reduzir a quantidade de bits necessária para representar uma dada informação. Neste contexto, os conceitos de dados e informação são diferentes, em que os dados são os meios pelos quais as informações são transmitidas [1].

Seguindo esta linha de raciocínio, uma informação pode ser representada de infinitas maneiras. Desta forma, o questionamento a ser respondido quando se objetiva a compressão de dados é: qual representação forneceria o menor volume de dados sem que haja perda de informação?

A compressão é obtida através da eliminação dos dados redundantes presentes na representação de uma informação. Em se tratando de arrays bidimensionais, os principais tipos de redundância são:

- *Redundância de codificação*: surge quando a quantidade de bits utilizada para representar as **intensidades** presentes em um array 2D é superior à quantidade necessária. 
- *Redundância espacial e temporal*: devido à grande parte dos pixels presentes em um array 2D estarem espacialmente correlacionados, surge a redundância espacial. Já as sequências de vídeo estão sujeitas a mais outro tipo de redundância, a temporal, em que os pixels de quadros vizinhos encontram-se correlacionados, devido à grande semelhança entre ~~quadros próximos~~ ^{eles}.
- *Redundância psicovisual*: é originada a partir das características do sistema ~~visual humano~~ ^{e da percepção visuais humanos} (HVS, do inglês human visual system). Sua resposta aos estímulos visuais é uma função não linear de grandezas físicas, como intensidade luminosa e cores.

A quantificação do volume de dados redundantes presente em uma representação de imagem é necessária, a fim de quantificar a compressão obtida. Sendo assim, assumindo que b e b' são, respectivamente, o volume de dados presentes na representação real de uma imagem e o volume de dados presentes em uma representação comprimida da mesma. A redundância  *relativa* R é dada,

$$R = 1 - \frac{1}{c} \quad (2.1)$$

em que c é a *taxa de compressão*, definida.

$$c = \frac{b}{b'} \quad (2.2)$$

Teoria da Informação

Durante a década de 40, no período da Segunda Guerra Mundial, o processo de troca de informações tornou-se fundamental. Dessa forma, surgiu a necessidade de estabelecer um limite mínimo de volume de bits necessário para a transmissão de uma determinada informação, a fim de otimizar a utilização do canal disponível.

Neste contexto, Claude Elwood Shannon ficou conhecido com o “pai da teoria da informação” ao propor com sucesso uma medida de informação própria para medir incerteza sobre espaços desordenados.

Primeiro teorema de Shannon

Segundo a equação 2.3, pode-se provar, através do teorema da codificação sem perda [9],
que é possível representar a saída de um sistema sem memória com uma média H de unidades de informação por pixel,

$$\lim_{n \rightarrow \infty} \left[\frac{L_{avg,n}}{n} \right] = H \quad (2.3)$$

em que, $L_{avg,n}$ é o tamanho médio dos códigos necessários para representar todos os grupos de n símbolos e H é a entropia.

Entropia

A entropia, equação 2.4, é uma métrica utilizada em diversas áreas do conhecimento, como na química e física. Em se tratando de informações, a entropia representa o grau de incerteza de uma fonte.

Sendo (a_1, a_2, \dots, a_J) o conjunto de símbolos de uma determinada fonte, a entropia é dada pela informação média associada a cada símbolo.

$$H = - \sum_{j=1}^J P(a_j) \log_b P(a_j) \quad (2.4)$$

No caso de imagens digitais, em que a unidade de representação é o *bit*, temos que $b = 2$ na equação 2.4.

Alguns métodos básicos de compressão

De maneira geral, existem dois tipos de compressão, com e sem perda, com as seguintes características:

- Sem perda: objetiva comprimir uma determinada informação sem que a mesma seja afetada. Para isso, códigos diferentes do código natural são atribuídos aos símbolos.
- Com perda: objetiva alcançar um maior nível de compressão através da eliminação de elementos sem que a informação seja fortemente afetada, de forma que a mesma possa ser entendida.

s	$p(s)$	Faixa	$s1$	$s1s2$	$s1s2s4$
$s1$	0,2	[0,0, 0,2)	[0,0, 0,04)	[0,04, 0,048)	[0,072, 0,0736)
$s2$	0,2	[0,2, 0,4)	[0,04, 0,08)	[0,048, 0,056)	[0,0736, 0,0752)
$s3$	0,4	[0,4, 0,8)	[0,08, 0,16)	[0,056, 0,072)	[0,0752, 0,0784)
$s4$	0,2	[0,8, 1,0)	[0,16, 0,2)	[0,072, 0,08)	[0,0784, 0,08)

Tabela 2.1: Exemplo de codificação aritmética.

Por fim, escolhe-se um número dentro do intervalo atribuído para uma determinada mensagem que deverá representar a mesma. No caso da mensagem $s1s2s4$, foi atribuída a faixa $[0,072, 0,08)$ e cada valor dentro da mesma poderá ser escolhido para representar esta mensagem.

- Run-length: inicialmente produzida para a compressão para ser utilizada na tecnologia de FAX, cujas imagens são binárias.


A codificação run-length é executada linha a linha começando com o valor inicial (0 ou 1) seguido pelo número de repetições sucessivas. Quando houver a mudança de valor basta acrescentar o números de repetições sucessivas, pois sabe-se que o próximo valor é a negação do anterior.

2. Com perda:

- DPCM: Sinais, como o de voz e imagens, possuem um alto grau de correlação entre amostras.

Considerando que uma determinada amostra pode ser representada por

$$f(n) = \hat{f}(n) + e(n) \quad (2.5)$$

em que $\hat{f}(n)$ é uma aproximação da amostra original e $e(n)$ é erro associado a mesma, o erro médio quadrático de predição pode ser minimizado através de uma  melhor aproximação do sinal original.

Na codificação DPCM (do inglês Differential Pulse-Code Modulation), se $e(n) \rightarrow 0$, temos que a aproximação do sinal original pode ser representada por uma combinação linear descrita em

$$\hat{f}(n) = \sum_{i=1}^m \alpha_i f(n-i) \quad (2.6)$$

em que os coeficientes devem ser calculado através da minimização da expressão 2.7.

$$E\{e(n)^2\} = E \left\{ \left[f(n) - \sum_{i=1}^m \alpha_i f(n-i) \right]^2 \right\} \quad (2.7)$$

- Codificação baseada em transformada de blocos: é uma técnica de compressão que consiste em dividir uma imagem em blocos não sobrepostos de tamanhos iguais (geralmente 8×8). Uma transformada linear reversível, como a transformada de Fourier e a transformada cosseno, é utilizada para mapear estes blocos em um conjunto de coeficientes que por fim serão submetidos a um processo de quantização [11].

Os padrões JPEG e MPEG

Informações visuais exercem uma grande influência sob a percepção humana: cerca de 80 – 90% dos neurônios estão relacionados com o processamento de informações visuais [12]. Dessa forma, não é de se surpreender que imagens e vídeos sejam cada vez mais explorados digitalmente.

Seguindo essa tendência, intensificaram-se as buscas por métodos capazes de otimizar a utilização da banda de transmissão sem que a informação seja prejudicada. Neste contexto surgiram o JPEG [13] e MPEG-1 [14], [15], [16].

JPEG

Em meados da década de 80, a União Internacional de Telecomunicações (ITU, do inglês International Telecommunication Union) concentrou seus esforços para a criação de um padrão de compressão de imagens estáticas. Desta forma deu-se a origem do JPEG.

Este padrão consiste em uma combinação de duas técnicas de compressão, com e sem perda (quantização e codificação de entropia). Como pode ser notado nas figuras 2.3 e 2.4, o codificador e o decodificador ~~baseados em entropia~~, respectivamente, do *JPEG baseline* [1], são destacados pela cor azul.

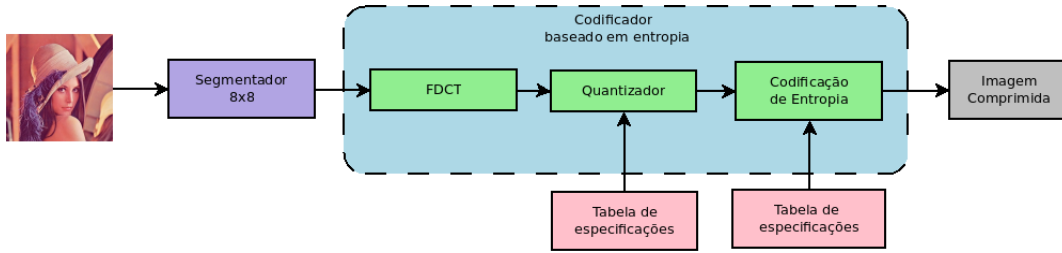


Figura 2.3: Fluxograma do codificador JPEG.

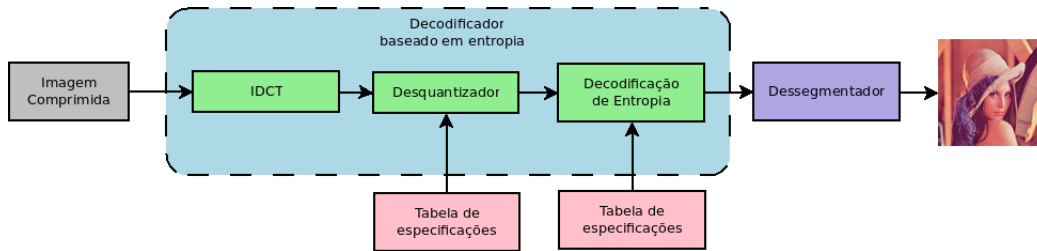


Figura 2.4: Fluxograma do decodificador JPEG.

JPEG baseline

No sistema “baseline”[1] o processo de compressão é composto por quatro passos sequenciais: segmentação, cálculo da transformada cosseno discreta, quantização e determinação dos códigos de tamanho variável para cada símbolo.

Inicialmente, a imagem é subdividida em blocos 8×8 . Depois que os blocos são encontrados, seus valores são deslocados, subtraindo 2^{k-1} unidades, em que 2^k é o número máximo de níveis de intensidade. Então, aplica-se a transformada cosseno ¹ discreta

$$D(i, j) = \frac{1}{\sqrt{2N}} C(i) C(j) \sum_{x=0}^{N-1} \sum_{y=0}^{N-1} p(x, y) \cos \left[\frac{(2x+1)i\pi}{2N} \right] \cos \left[\frac{(2y+1)j\pi}{2N} \right] \quad (2.8)$$

$$C(u) = \begin{cases} \frac{1}{\sqrt{2}} & \text{if } u = 0 \\ 1 & \text{if } u > 0 \end{cases} \quad (2.9)$$

2D [17], seguida pelo processo de quantização,

$$\hat{T}(u, v) = \text{round} \left(\frac{T(u, v)}{Z(u, v)} \right) \quad (2.10)$$

¹Onde $p(x, y)$ é o valor do pixel na posição (x, y) e N é a ordem do bloco (oitava ordem).

em que $\hat{T}(u, v)$ é o bloco quantizado (tabela 2.5), $T(u, v)$ é o bloco transformado e $Z(u, v)$ é a tabela de quantização (tabela 2.2).

16	11	10	16	24	40	51	61
12	12	14	19	26	58	60	55
14	13	16	24	40	57	69	56
14	17	22	29	51	87	80	62
18	22	37	56	68	109	103	77
24	35	55	64	81	104	113	92
49	64	78	87	103	121	120	101
72	92	95	98	112	100	103	99

Tabela 2.2: Tabela de quantização padrão.

0	1	5	6	14	15	27	28
2	4	7	13	16	26	29	42
3	8	12	17	25	30	41	43
9	11	18	24	31	40	44	53
10	19	23	32	39	45	52	54
20	22	33	38	46	51	55	60
21	34	37	47	50	56	59	61
35	36	48	49	57	58	62	63

Tabela 2.3: Padrão zigzag.

52	55	61	66	70	61	64	73
63	59	66	90	109	85	69	72
62	59	68	113	144	104	66	73
63	58	71	122	154	106	70	69
67	61	68	104	126	88	68	70
79	65	60	70	77	63	58	75
85	71	64	59	55	61	65	83
87	79	69	68	65	76	78	94

Tabela 2.4: Exemplo de subimagem.

-26	-3	-6	2	2	0	0	0
1	-2	-4	0	0	0	0	0
-3	1	5	-1	-1	0	0	0
-4	1	2	-1	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0

Tabela 2.5: Subimagem da tabela 2.4 quantizada.

Por fim, os coeficientes dos blocos quantizados são reordenados no padrão zigzag (tabela 2.3), resultadando em um arrey 1D,

[-26 -3 1 -3 -2 -6 2 -4 1 -4 1 1 5 0 2 0 0 -1 2 0 0 0 0 0 -1 -1 EOB]


e codificados com base nas tabelas de Huffman, pré definidas, apresentadas em [13], as quais agrupam derivações dos métodos DPCM e run-length², mensioandos na seção 2.4.1.



O processo de decodificação é obtido através da simples inversão da ordem das operações.




²A palavra código EOB significa que os coeficientes são iguais a 0 daquele ponto até o último coeficiente AC.

MPEG-1

O MPEG-1 (H.261 [18]) é um método versátil de compressão de vídeos com perda, pois pode ser aplicado a uma grande variedade de formatos de entradas. Porém, foi otimizado para aplicações que suportam taxas contínuas de transferência de bits de 1.5Mbps (CD, do inglês Compact Disc). 

Como mencionado na seção 2.2, os vídeos estão sujeitos a redundância temporal devido a alta semelhança entre frames vizinhos. Por isso, antes de falar do MPEG-1 abordaremos o processo de estimação e compensação de movimentos baseados em blocos, a fim de eliminar informações irrelevantes entre frames.  


Estimação e compensação de movimentos baseados em blocos

Objetivando-se reduzir a redundância temporal entre imagens consecutivas poderia se pensar em armazenar apenas a diferença entre duas imagens, porém este processo pode ser otimizado através da utilização de macroblocos. Por isso pode-se dizer que este método é uma variação da codificação DPCM. 

A utilização do conceito de macroblocos para a encontrar a diferença entre imagens consecutivas possibilita a minimização do erro médio absoluto (MAE, do inglês mean absolute error)

$$MAE(i, j) = \frac{1}{MN} \sum_{k=0}^{M-1} \sum_{l=0}^{N-1} |C(x+k, y+l) - R(x+i+k, y+j+l)| \quad (2.11)$$

da mesma da mesma. Dessa forma obtendo uma maior redução de volume de dados.

Inicialmente consideremos uma imagem atual em que temos um macrobloco de dimensões $m \times n$ e uma imagem anterior com um macrobloco de mesmas dimensões com a menor diferença possível. Estes dois macroblocos possuem um deslocamento relativo designado por “vetor de deslocamento” e esta diferença é designada por “erro de predição”. A este processo dá-se o nome de compensação de movimento pra frente. (fig:2.6). 

Também pode-se estender este raciocínio, tomando como base uma imagem atual, uma anterior e uma posterior a fim de obter a menor diferença possível. Este processo chama-se compensação de movimento bidirecional. (fig:2.6).

A diferença fundamental entre os dois modelos de compensação de movimento é que o bidirecional trabalha com três erros de predição (pra trás, pra frente ou interpolativo), ao em vez de um.

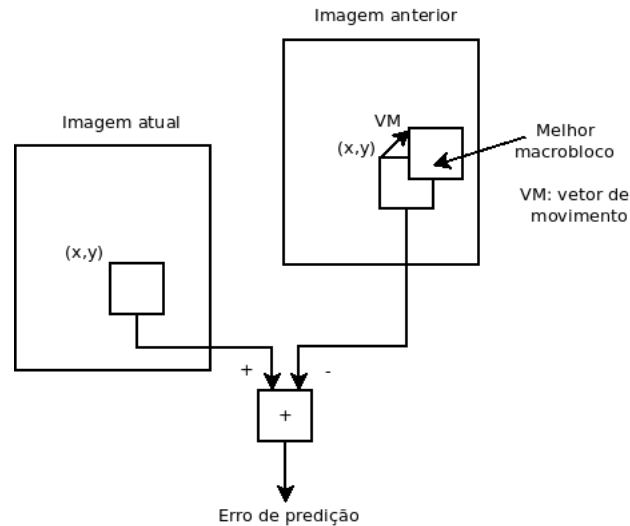


Figura 2.5: Compensação de movimento pra frente.

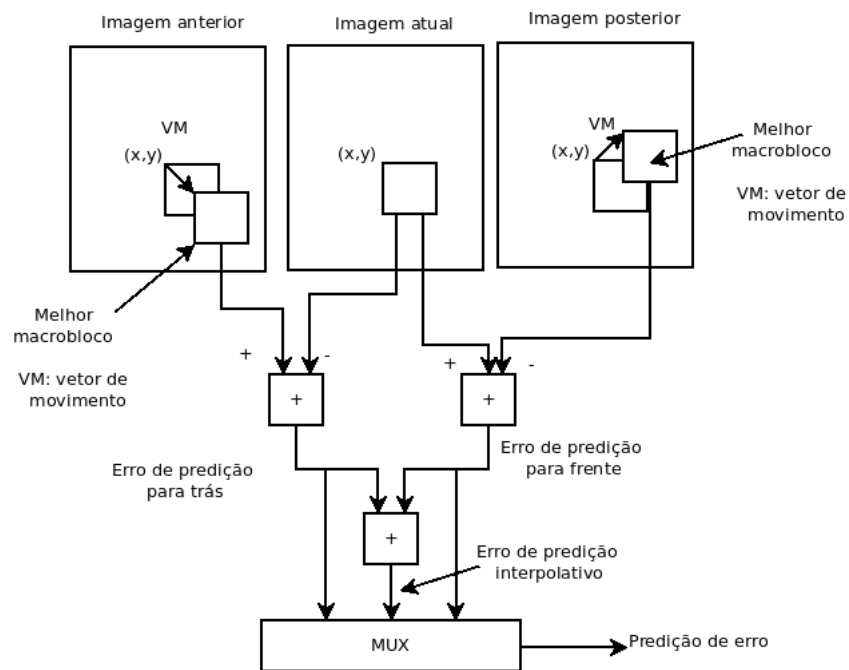


Figura 2.6: Compensação de movimento bidirecional.

Métodos de busca

Há muitas formas de determinar os vetores de deslocamento dos macroblocos, algumas possibilitam uma maior precisão ao encontrar o melhor candidato outras possibilitam uma maior velocidade na determinação do mesmo.


Dentre os métodos básicos, podemos citar: 

1. **Busca completa:** apesar de se o métodos mais simples, garante que o melhor candidato seja encontrado, obtendo o menor valor possível do MAE.

Consiste na definição de uma área de busca de p pixels para os lados, para cima e para baixo a partir do canto superior esquerdo do macrobloco. Cada um dos $(2p + 1)^2$ pixels da área de busca serão formados os possíveis macroblocos a serem submetidos ao critério de minimização do MAE.

2. **Busca unidimensional paralela hierárquica:** este método não garante o menor valor possível para o MAE, porém apresenta um ganho considerável em velocidade.

Este algoritmo de busca é descrito da seguinte forma:

- (a) Para uma área de busca $[-p, p]$, definida de forma similar a apresentada no método de busca completa, defini-se $S = 2^{\lfloor \log_2 p \rfloor}$ e assumindo que a posição de origem (d_i, d_j) do macrobloco seja $(0, 0)$.
- (b) Em paralelo, deve-se calcular:
 - Para o eixo i:  encontrar dentre as três posições $(d_i - S, d_j)$, (d_i, d_j) e $(d_i, d_j + S)$, qual delas gera o menor MAE. Por fim substituir d_i pela posição encontrada.
 - Para o eixo j: encontrar dentre as três posições $(d_i - S, d_j)$, (d_i, d_j) e $(d_i, d_j + S)$, qual delas gera o menor MAE. Por fim substituir d_j pela posição encontrada e $S = \frac{S}{2}$.

O passo (b) deve ser repetido sucessivas vezes até que $S = 0$. O vetor resultante (d_i, d_j) será o vetor de deslocamento do macrobloco.



Algoritmo

O padrão H.261 não reconhece entradas entrelaçadas, por isso utiliza-se a denominação de “imagens” em vez de “frames”. Há três tipos de imagens que podem ser classificadas em dois métodos de compressão:

1. Intra imagem:

- I (intra): não leva em consideração imagens vizinhas.

2. inter imagens³:

- P (predita): imagem processada com base na imagem I anterior. 
- B (bidirecionalmente predita): imagem processada com base na imagem I anterior e na imagem P posterior, ou vice sersa. 

Para o codificador (fig: 2.7), inicialmente os canais Cb e Cr são disimados segundo o padrão 4:2:0 [19] e defini-se a ordem dos tipos de imagens dentro de um GOP (do inglês group of images), de acordo com as necessidades. A mais comum é mostrada na figura 2.8.

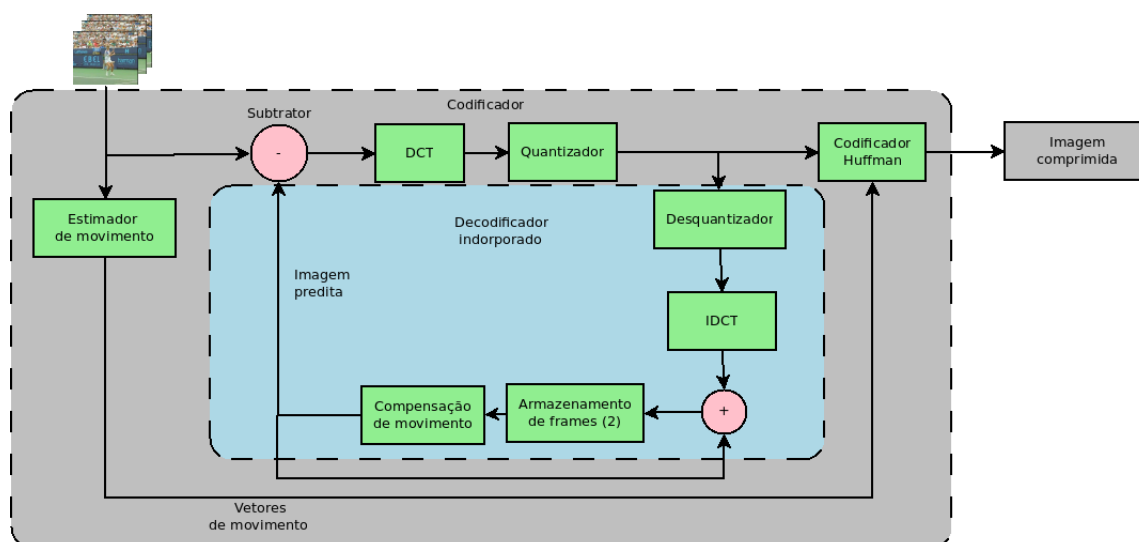


Figura 2.7: Fluxograma do codificador MPEG.

Depois que cada tipo de imagem foi processada, as imagens I, P e B são codificadas conforme descrito no padrão JPEG, porém as duas últimas utilizam a codificação de huffman descrita em [14] para armazenar os vetores de deslocamento.

Para o decodificador (fig:2.9), deve-se trabalhar por grupos de imagens, recuperando as imagens segundo o decodificador JPEG e recuperando as imagens ($P_1, P_2, P_3, \dots, P_n$) originais na seguinte ordem:

1. P_i (imagem I): não há compensação de movimento. Deve ser reconstruído através do decodificador JPEG;

³As imagens codificadas dessa forma utilizam a metodologia apresentada no item 2.5.2.1.

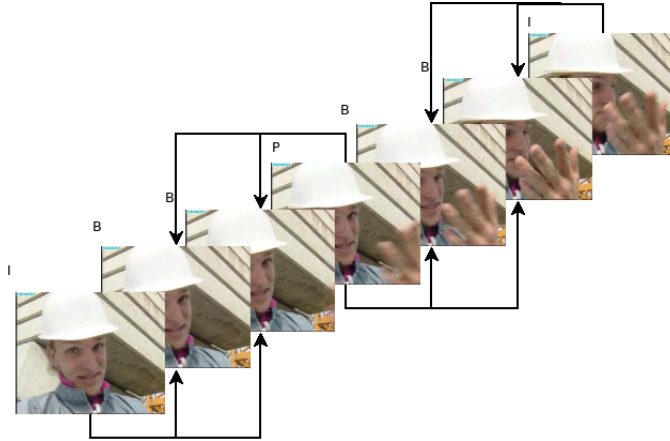


Figura 2.8: Sequência de imagens dentro de um GOP.

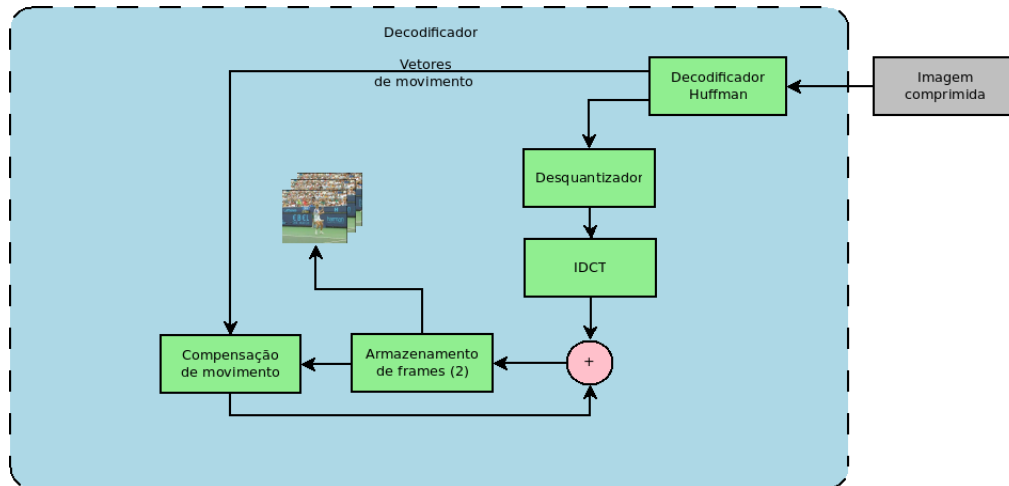


Figura 2.9: Fluxograma do decodificador MPEG.

2. P_{i+3} (imagem P): primeiramente o erro de predição deve ser recuperado através do decodificador JPEG. Em seguida, cada macrobloco é submetido a IDCT e a compensação de movimento com base em P_i .
3. P_{i+1} (imagem B): primeiramente o erro de predição deve ser recuperado através do decodificador JPEG. Em seguida, cada macrobloco é submetido a IDCT e a compensação de movimento com base em P_i e P_{i+3} .
4. P_{i+2} (imagem B): repete-se os passos descritos em P_{i+1} .
5. P_{i+6} (imagem I): repete-se os passos descritos em P_i .

6. P_{i+4} (imagem B): repete-se os passos descritos em P_{i+1} , com base em P_{i+3} e P_{i+6} .
7. P_{i+5} (imagem B): repete-se os mesmos passos de P_{i+4} .

Capítulo 3

Qualidade visual

Introdução

O processo de evolução da espécie humana atribuiu uma grande importância às informações visuais. Não é de se espantar que existam grandes esforços voltados para pesquisas que objetivam encontrar métodos capazes de avaliar, bem como melhorar a qualidade visual de imagens e vídeos recuperados através do processo de decodificação.

O conhecimento das características do sistema visual humano (HVS, do inglês human visual system) é fundamental para o desenvolvimento de metodologias eficazes no processo de melhoria da qualidade visual. Embora o nosso conhecimento do HVS seja muito limitado, há de se reconhecer que já nos proporcionaram bons resultados, como em [8] e [20]. Neste contexto, os processos de captura, exibição, armazenamento e transmissão deveram ser adaptados a fim de gerar representações mais exatas das imagens reais.

Este capítulo tem como objetivo apresentar os aspectos principais da qualidade visual bem como apresentar maneiras de quantificá-la, com foco em compressão de imagens e vídeos digitais.

Artefatos provenientes do processo de compressão

Na seção 2.5 foi mentionado que o JPEG e o MPEG-1 realizam uma quantização dos blocos transformado, a fim de eliminar as redundâncias espacial e temporal. Em alguns sistemas este processo é o principal responsável pelo surgimento de distorções, apesar de não ser o único fator capaz de afetar a qualidade visual. Alguns dos tipos de artefatos mais comuns em

seqüências de vídeos são listados a seguir:

- Efeito de blocagem: ocorre devido a quantização independente de blocos individuais, gerando descontinuidade nos limites dos blocos adjacentes.
- Embaçamento: consequência da supressão das componentes de alta frequênciadurante o processo de quantiação, manifestando-se através da perda da reslução espacial.
- Vazamento de cores: devido aos processos de subamostragem de crominância seguido pela quantização, ocorre o vazamento de cores entre áreas de com crominância muito diferentes.
- Efeito de imagem base da DCT: ocorre quando uma única componente da DCT é comi-
nante em um macro bloco, resultando na ênfase de uma imagem base da DCT.
- Efeito ressonante: está relacionado com o fenômeno de Gibb's [1], logo é mais evidente em áres de grande contraste. Resultando da inregularidade de reconstrução das componentes de baixa frequência devido a quantização.
- Aliasing: ocorre quando a frequência de amostragem está abaixo da frequência de Nyquist [1], tanto espacialmente quanto temporalmente.

Avaliações subjetiva e objetiva

A melhor forma de conseguir avaliar a qualidade de informações visuais é através da opinião de observadores. A nota média de opiniões (MOS do inglês Mean Opinion Score), métrica subjetiva que necessita da opinião de um grupo de observadores, é tida como uma das melhores métricas de qualidade visual, porém custosa e precisa de um longo espaço de tempo para que possa gerar resultados confiáveis.

Frequentemente o erro médio quadrático (MSE, do inglês Mean Square Error) é utilizado peloss métodos avaliação de qualidade de imagens, devido a sua fácil implementação. No entanto, esses métodos aprensntam baixo nível de correlação com as métricas subjetivas, pois o MSE não leva em consideração características espaciais [21].

Logo, as pesquisas de avaliação objetiva de imagens buscam contornar os inconvenientes presentes na avaliação subjetiva, através de modelos matemáticos que simulam a percepção visual humana.



Classificação das métricas objetivas de avaliação de qualidade visual

Para que uma métrica ~~objetiva~~ seja realmente útil para a avaliação de qualidade visual é necessário que ela seja capaz suprir as limitações do MSE. Objetivando-se contornar essa situação, surgiram vários métodos que, agrosso modo, podem ser classificados em três categorias:

- Referência completa: utiliza a imagem original (considerada sem distorção) para avaliar uma imagem distorcida. Portanto, proporciona resultados mais precisos em relação a similaridade e fidelidade entre as duas imagens.
- Sem referência: utilizada quando não é possível ter acesso a imagem original, logo a avaliação da imagem distorcida deve ser feita as “cegas”, o que faz disso uma tarefa difícil. Por isso esta categoria também é conhecida como de referência cega.
- Referência reduzida: neste caso a imagem de referência não está completamente disponível e sim algumas características, que são embutidas no sistema que irá avaliar a imagem distorcida.

Outra classificação possível seria em relação a utilização das características do HVS. Neste caso, existem duas categorias:

- Perceptuais: essas são formulações matemáticas de certa complexidade inspiradas em características fisiológicas e psicovisuais da visão que mensuram de forma automática a qualidade da imagem, com expressiva correlação com a percepção humana.
- Não perceptuais: não utilizam características do SVH nas suas formulações e tem como virtude a baixa complexidade computacional. Porém, possuem baixa correlação com as métricas subjetivas.

A seguir serão apresentadas as métricas não perceptuais e depois as perceptuais utilizadas neste trabalho.

Não perceptuais

Erro médio quadrático (Mean Square Error - MSE)

O MSE é uma métrica bastante utilizada que, que consiste no valor esperado dos quadrados dos erros,

$$MSE = E [(y(i, j) - x(i, j))^2] \quad (3.1)$$



em que y é a imagem distorcida e x é a imagem original.



Razão Sinal-Ruído (Signal-to-Noise Ratio - SNR)

Quantifica o quanto um sinal foi distorcido através do cálculo da energia do erro da imagem distorcida.

A SNR é comumente meidda em dB da seguinte forma,

$$SNR = 10 \log_{10} \frac{\sum_{i,j} [x(i, j)]^2}{\sum_{i,j} [x(i, j) - y(i, j)]^2} \quad (3.2)$$



em que y é a imagem distorcida e x é a imagem original.

Razão Sinal-Ruído de Pico (Peak Signal-to-Noise Ratio - PSNR)

Comumente utilizada para medir a qualidade da reconstrução da imagem ou vídeo após uma compressão com perdas.

A PSNR é comumente meidda em dB da seguinte forma,

$$PSNR = 10 \log_{10} \frac{NK^2}{\sum_{i,j} [x(i, j) - y(i, j)]^2} \quad (3.3)$$



onde $x(i, j)$ representa o sinal de referência, $y(i, j)$ representa o sinal de teste, N representa o número total de pixels da imagem e K representa o valor máximo que um bit pode atingir. No caso de uma imagem de 8 bits/pixel o valor de K é 255.



Perceptuais

Índice de Similaridade Estrutural (Structural Similarity Index - SSIM)

O SSIM talvez seja a métrica perceptual mais utilizada, devido a sua baixa complexidade de implementação em relação as demais e obter bom aproximações do HVS. Esta métrica avalia

o quanto da estrutura da imagem de teste diferencia da estrutura da imagem de referência.

O SSIM é definido como

$$SSIM = [l(x, y)^\alpha c(x, y)^\beta s(x, y)^\gamma] \quad (3.4)$$

em que $\alpha > 0$, $\beta > 0$ e $\gamma > 0$ são responsáveis pelo ajuste da importância relativa das componentes $l(x, y)$, $c(x, y)$ e $s(x, y)$ correspondem as componentes de luminância, contraste e estrutura, respectivamente, definidas como

$$l(x, y) = \frac{2\mu_x\mu_y + c_1}{\mu_x^2\mu_y^2 + c_1} \quad (3.5)$$

$$c(x, y) = \frac{2\sigma_x\sigma_y + c_2}{\sigma_x^2\sigma_y^2 + c_2} \quad (3.6)$$

$$s(x, y) = \frac{2\sigma_{xy} + c_3}{\sigma_x\sigma_y + c_3} \quad (3.7)$$

em que μ_x e σ_x são a média e a variância da imagem x ; μ_y e σ_y são a média e a variância da imagem y ; σ_{xy} é a covariância entre as imagens x e y . Os valores c_1 , c_2 e c_3 são inseridos com o objetivo de evitar instabilidades.

Esta métrica é aplicada localmente, deslocando, horizontalmente e verticalmente, uma janela de tamanho $B \times B$ sobre a imagem. Pode-se calcular a média do SSIM (MSSIM) para que se tenha uma índice de qualidade geral da imagem.

$$MSSIM = \frac{1}{p} \sum_{j=1}^p SSIM_j \quad (3.8)$$

Quantização interadaptativa baseada no sistema visual humano

A grosso modo, o sistema visual humano consegue captar pequenas variações de brilho em áreas relativamente grandes, mas não consegue captar com a mesma facilidade quando estas variações estão presentes em componentes de alta frequência. Os sistemas de compressão obtêm vantagens desse fato ao quantizar fortemente as componentes de alta frequência e de maneira mas branda as componentes de baixa frequência. Como pode ser visto na figura 3.1, a DCT agrupa as componentes de baixa frequência no canto superior esquerdo e as de alta frequência no inferior direito do macrobloco transformado, possibilitando a produção de tabela de quantização adequadas para o propósito mencionado anteriormente.

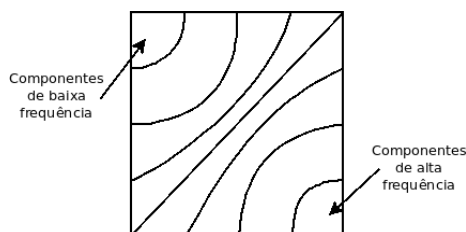


Figura 3.1: Distribuição da frequência espacial no domínio da DCT.

O padrão H.261 [14] utiliza o mesmo método de compressão para as imagens intra e inter codificadas. Atualmente vários métodos que levam em consideração as características do HVS tem sido propostos para melhorar a qualidade do vídeo reconstruído.

Em [8], um algoritmo de quantização inter imagens é proposto. Este algoritmo baseia-se no modelo apresentado por D. H. Kelly [2], o qual afirma que o sistema visual humano é mais sensível as variações de contraste nas frequências espaciais intermediárias, ao passo que é menos sensível em frequências baixas e altas (figura 3.2). Gerando previamente uma série de matrizes de quantização.

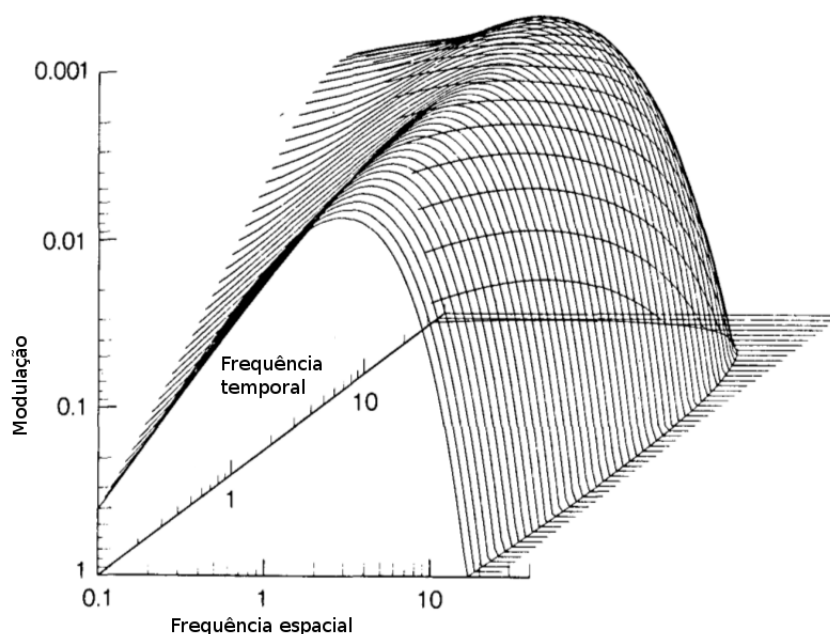



Figura 3.2: Visão perspectiva da superfície de limiar espacial e temporal [2]. Cada curva representa a resposta a frequência espacial em uma frequência espacial fixa.


Essas matrizes são geradas com base no modelo de D. H. Kelly [2], que caracteriza as

limitações do HVS em relação as respostas estaciais e temporais ao propor uma equação que representa a resposta do HVS a frequência temporal em função da velocidade do macrobloco e da frequência espacial

$$G(\alpha, v) = [6.1 + 7.3|\log(v/3)|^3] v \alpha^2 \exp[-2\alpha(v+2)/45.9] \quad (3.9)$$

em que v velocidade do macrobloco medida em graus e α é a frequência espacial medida em  ciclos por segundo. Ambas são definidas por

$$v = (v_H^2 + v_T^2)^{1/2} \quad \text{e} \quad \alpha_{ij} = \alpha_i + \alpha_j \quad (3.10)$$

 em que m_H e m_T são os valores absolutos das componentes do vetor de deslocamento do macrobloco¹.

$$v_H = \frac{m_H}{w} \times f \quad \text{e} \quad v_T = \frac{m_T}{h} \times f \quad (3.11)$$

$$\alpha_i = \frac{m_H}{m} \times \frac{w}{m} \times c_i \quad \text{e} \quad \alpha_j = \frac{m_T}{m} \times \frac{h}{m} \times c_j \quad (3.12)$$

$$c_i = 0.5i \quad \text{e} \quad c_j = 0.5j \quad i, j = 1, 0...7 \quad (3.13)$$

em que w e h são as dimensões da imagem e m é o tamanho do macrobloco.

Desta forma, espera-se melhorar a qualidade subjetiva do vídeo decodificado ao quantizar individualmente cada macrobloco respeitando as características do HVS segundo o modelo proposto em [8]

$$Q_{HVS}(i, j) = (m_H + m_T)/p \left(1 - \frac{G(\alpha_{ij}, v)}{G_{max}} \right) \quad (3.14)$$

em que $Q_{HVS}(i, j)$ é uma componenta da matriz de quantização, p é um parâmetro de ajuste e G_{max} é o valor máximo de Q_{HVS} .

Por fim, o modelo proposto pode ser obtido somando a matriz de quanização baseada no HVS a matriz de quantização plana.

$$Q = Q_{flat} + Q_{HVS} \quad (3.15)$$

¹Logo, este método de quantização perceptual só é aplicado as imagens preditas (P, B), cuja qualidade depende diretamente do fator de qualidade das imagens I.

Capítulo 4

Procedimento experimental e resultados

Implementação do sistema

Para a realização dos experimentos práticos, o codificador e o decodificador, referentes a informações visuais, do padrão H.261 foram implementados em Python [22], bem como o método de quantização perceptual interadaptativa (seção 3.5).

Conjunto de teste

A análise do método de quantização perceptual [8] utilizado neste trabalho consiste em compará-lo com o utilizado no codificador H.261 padrão com diferentes condições. Para este procedimento quatro vídeos com diferentes intensidades de atividades temporais, obtidos em [23], serão utilizados (tabela 4.1), de forma a analisar a compressão e a qualidade visual alcançada em vídeos com diferentes atividades temporais.

Vídeos	Atividade Temporal	Duração (seg)
Akiyo	Baixa	12
Ice	Média	8
Stefan	Média	3
Foreman	Alta	12

Tabela 4.1: Descrição dos vídeos.

Metodologia

Para analisar em quais situações a quantização perceptual interadaptativa seria vantajosa, os codificadores com e sem a mesma foram submetidos a diferentes fatores de qualidade, de forma a gerar uma variação de taxa em ambos os casos.

Neste trabalho, o modelo de quantização perceptual utiliza deslocamento de busca de 15 pixels, apartir da origem, na horizontal e na vertical, $p = 2$ e $Q_{flat}(i, j) = 10$ para $i, j = 0, 1, \dots, 7$.

Por fim, os vídeos decodificados seram submetidos a uma avaliação visual objetiva e comparados com os resultados subjetivos obtidos em [8].

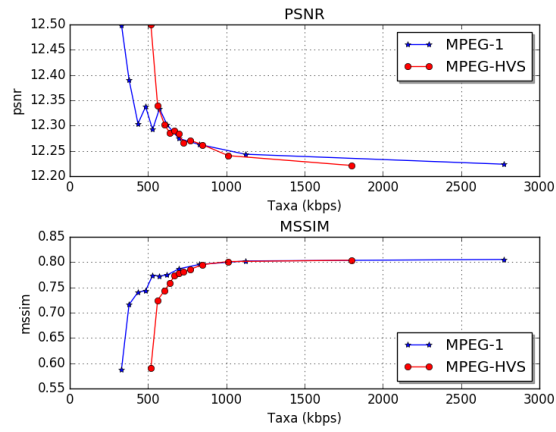
Resultados

Os resultados mostram que para baixos fatores de qualidade, o codificador perceptual apresenta um aumento nos valores de MSSIM apesar do aumento de taxa em relação ao codec padrão. Porém, para altos fatores de qualidade os valores de MSSIM dos dois codificadores, tendem a se igualar, enquanto que a taxa do codificador perceptual apresenta uma considerável redução em relação ao padrão (figura 4.1(a) a 4.1(d)).

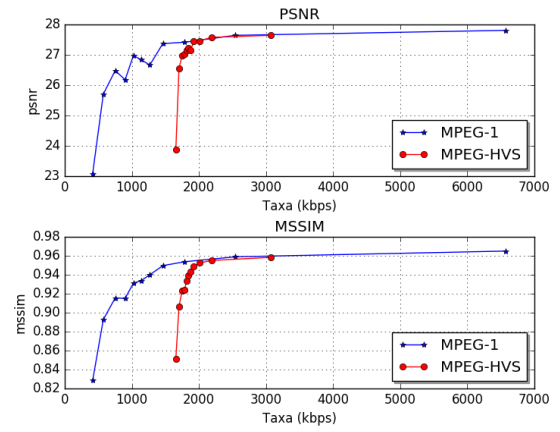
É importante observar que os nem sempre os valores do PSNR e MSSIM vão apresentar a mesma tendência, como na figura 4.1(a). Este fato atesta que o PSNR, diferentemente do MSSIM, não é capaz de avaliar as características espaciais das imagens, devendo ser usado em conjunto com outras métricas para garantir uma análise mais apurada.

Levando em consideração que a faixa de valores dos vetores de deslocamento do MPEG-1 é de $[-16; 16]$ e a do MPEG-3, usando em [8], é de $[-31, 5; 31, 5]$, era de se esperar que a quantização perceptual apresentasse um aumento de taxa para baixos fatores de qualidade quando comparado com o codec padrão, pois os valores de Q_{HVS} dependem dos valores dos vetores de deslocamento (equação 3.14).

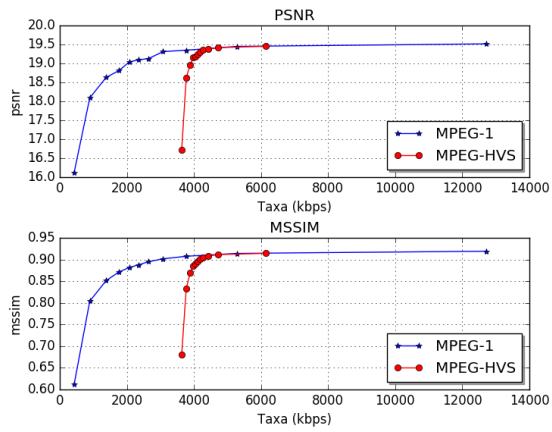
Visualmente, pode-se atestar que com exceção dos vídeos com baixa atividade temporal (Akiyo) o método de quantização perceptual gera uma diminuição do efeito de blocagem ao alocar mais bits para as frequências mais sensíveis ao sistema visual humano em baixas taxas (figuras 4.2(a) a 4.3(d)).



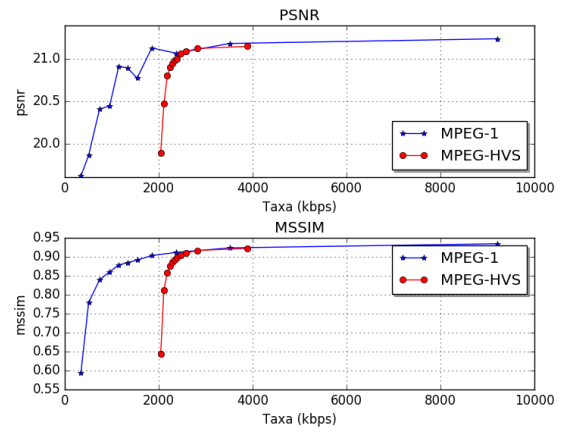
(a) Akiyo.



(b) Ice.



(c) Stefan.



(d) Foreman.

Figure 4.1: MSSIM.



(a) Codec padrão (10%): Akiyo.



(b) Codec perceptual (10%): Akiyo.



(c) Codec padrão (10%): Ice.



(d) Codec perceptual (10%): Ice.

Figura 4.2: Codec padrão vs Codec perceptual.



(a) Codec padrão (10%): Stefan.



(b) Codec perceptual (10%): Stefan.



(c) Codec padrão (10%): Foreman.



(d) Codec perceptual (10%): Foreman.

Figura 4.3: Codec padrão vs Codec perceptual.

Capítulo 5

Considerações finais

Conclusões

Com base nos resultados obtidos, o processo de quantização baseada no sistema visual humano, analisado neste trabalho, é capaz de melhorar substancialmente a qualidade perceptual dos vídeos decodificados, com exceção de vídeos com baixa atividade temporal. Logo a melhoria perceptual evidenciada subjetivamente em [8] foi reproduzida neste trabalho e atestada através de uma avaliação objetiva.

É bem verdade que quando este método é aplicado no MPEG-1, o nível de compressão alcançado é vantajoso apenas para altos fatores de qualidade, porém quando aplicado a codificadores que possibilitam áreas de busca maiores pode-se alcançar uma redução na taxa em relação ao codec padrão, como apresentado em [8].

Portanto, conclui-se que a viabilidade, em relação a volume de dados e qualidade visual, do método de quantização inter adaptativa baseada no sistema visual humano é dependente do tamanho da área de busca do codec utilizado.

Trabalhos futuros

A partir da teoria e dos resultados apresentados neste trabalho, objetiva-se aplicar a técnica analisada e outros codecs, bem como compará-la com outras técnicas que objetivam a melhoria da qualidade perceptual dos vídeos decodificados.

Referências Bibliográficas

- [1] R. C. Gonzalez and R. E. Woods, *Digital Image Processing (3rd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2006.
- [2] D. H. Kelly, “Motion and vision. ii. stabilized spatio-temporal threshold surface,” *J. Opt. Soc. Am.*, vol. 69, no. 10, pp. 1340–1349, Oct 1979. [Online]. Available: <http://www.osapublishing.org/abstract.cfm?URI=josa-69-10-1340>
- [3] A. Redondi, M. Cesana, and M. Tagliasacchi, “Low bitrate coding schemes for local image descriptors,” 2012, pp. 124–129, cited By 6. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84870605750&partnerID=40&md5=36a3c310d83eb80c36005d94ec72ffaa>
- [4] C. Singh, N. Singh, and R. Tripathi, “Optimization of standards for video compression tools over wireless networks,” 2012, pp. 114–118, cited By 0. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84862111611&partnerID=40&md5=c04186a826548dc0706b41c12308327f>
- [5] H. Song and C.-C. Jay Kuo, “Rate control for low-bit-rate video via variable-encoding frame rates,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 11, no. 4, pp. 512–521, 2001, cited By 88. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-0035307517&partnerID=40&md5=b2d06ea8d0e2cac993c70f68745c1039>
- [6] B. Vizzotto, B. Zatt, M. Shafique, S. Bampi, and J. Henkel, “A model predictive controller for frame-level rate control in multiview video coding,” 2012, pp. 485–490, cited By 1. [Online]. Available: <http://www.scopus.com/inward/record.url?eid=2-s2.0-84868110949&partnerID=40&md5=1ef2b60451ea8a334c78401dca4ddff0>

- [7] “Em 5 anos, o planeta terá mais dispositivos móveis ligados à Internet do que habitantes.” [Online]. Available: <http://startupi.com.br/2013/02/em-5-anos-o-planeta-tera-mais-dispositivos-moveis-ligados-a-internet-do-que-pessoas/>
- [8] J. Li, J. Koivusaari, J. Takala, M. Gabbouj, and H. Chen, “Human visual system based adaptive inter quantization.”
- [9] C. Shannon, “A mathematical theory of communication,” *Bell System Technical Journal*, vol. 27, pp. 379–423, 623–656, 1948.
- [10] D. A. Huffman, “A method for the construction of minimum-redundancy codes,” *Proceedings of the Institute of Radio Engineers*, vol. 40, no. 9, pp. 1098–1101, September 1952.
- [11] V. Bhaskaran and K. Konstantinides, *Image and Video Compression Standards: Algorithms and Architectures*, 2nd ed. Norwell, MA, USA: Kluwer Academic Publishers, 1997.
- [12] R. A. Young, “Oh say, can you see? the physiology of vision,” *Proc. SPIE*, vol. 1453, pp. 92–123, 1991. [Online]. Available: <http://dx.doi.org/10.1117/12.44348>
- [13] I. Telegraph and T. C. Committee, *CCITT Recommendation T.81: Terminal Equipment and Protocols for Telematic Services : Information Technology - Digital Compression and Coding of Continuous-tone Still Images - Requirements and Guidelines*. International Telecommunication Union, 1993. [Online]. Available: <https://books.google.com.br/books?id=XCcXHwAACAAJ>
- [14] T. S. S. of ITU., *ITU-T Recommendation H.261: Line Transmission of Non-Telephone Signals : Video Codec for Audiovisual Services at P X 64 Kbits*. International Telecommunication Union, 1993. [Online]. Available: <https://books.google.com.br/books?id=WG3jHgAACAAJ>
- [15] J. Bialkowski, M. Barkowsky, and A. Kaup, “Fast video transcoding from h.263 to h.264/mpeg-4 avc,” *Multimedia Tools Appl.*, vol. 35, no. 2, pp. 127–146, Nov. 2007. [Online]. Available: <http://dx.doi.org/10.1007/s11042-007-0126-7>
- [16] T. Wiegand, G. Sullivan, G. Bjontegaard, and A. Luthra, “Overview of the h.264/avc video coding standard,” *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 560–576, July 2003.

- [17] K. Cabeen and P. Gent, “Image compression and the discrete cosine transform,” *College of the Redwoods*, 1998.
- [18] T. S. S. of ITU., *ITU-T Recommendation H.261: Line Transmission of Non-Telephone Signals : Video Codec for Audiovisual Services at P X 64 Kbits*. International Telecommunication Union, 1993. [Online]. Available: <https://books.google.com.br/books?id=WG3jHgAACAAJ>
- [19] C. Poynton, “Chroma subsampling notation,” *Retrieved June*, vol. 19, p. 2004, 2002.
- [20] S. Chikkerur, V. Sundaram, M. Reisslein, and L. J. Karam, “Objective video quality assessment methods: A classification, review, and performance comparison,” *Broadcasting, IEEE Transactions on*, vol. 57, no. 2, pp. 165–182, 2011.
- [21] Z. Wang and A. Bovik, *Modern Image Quality Assessment*, ser. Synthesis lectures on image. Morgan & Claypool Publishers, 2006. [Online]. Available: <https://books.google.com.br/books?id=F6lYVwyZJz4C>
- [22] “PythonBrasil - PythonBrasil,” 00000. [Online]. Available: <http://wiki.python.org.br/>
- [23] “Xiph.org :: Derf’s Test Media Collection,” 00002. [Online]. Available: <https://media.xiph.org/video/derf/>