7

Referências bibliográficas

[1] SHANNON, C. E. A mathematical theory of communication. **Bell System Technical Journal,** v. 27, n. 17, p. 379 – 423, 623 – 656, July and October 1948.

[2] DRAGOTTI, P. L.; GASTPAR, M. **Distributed Source Coding: theory, algorithms, and applications.** Burlington, MA, USA: Academic Press, 2009.

[3] SLEPIAN, J. D.; WOLF, J. K. Noiseless coding of correlated information sources. **IEEE Transactions on Information Theory,** v.19, n. 4, pp. 471 – 480, Julho, 1973.

[4] WYNER, A.; ZIV, J. The rate-distorsion function for source coding with side information at the decoder. **IEEE Transactions on Information Theory,** v. 2, n. 1, pp. 1 – 10, Janeiro 1976.

[5] OPPENHEIM, A. V.; WILLSKY, A. S. **Signals & Systems.** 2.ed. New Jersey, USA: Prentice-Hall, 1997.

[6] OPPENHEIM, A. V.; SHAFER, R; BUCK, J. R. **Discrete-Time Signal Processing.** 2.ed. New Jersey, USA: Prentice-Hall, 1997.

[7] JAIN, A. K. **Fundamentals of Digital Image Processing.** New Jersey, USA: Prentice-Hall, 1997.

[8] TEKALP, A. M. **Digital Video Processing.** New Jersey, USA: Prentice-Hall, 1995.

[9] INTERNATIONAL ORGANIZATION FOR STANDARDZATION. **ISO/IEC International Standard 13818-2:1996:** Information technology – Generic coding of moving pictures and associated audio information: Video. Genebra, Suíça, 1996.

[10] RICHARDSON, I. E. G. **H.264 and MPEG-4 Video Compression: Video Coding for Next-Generation Multimedia.** UK: John Wiley & Sons, 2003.

[11] GOWDAK, D.; GOWDAK, L. H. **Atlas de Anatomia Humana.** São Paulo: FTD, 1989.

[12] LIM, J. S. **Two-dimensional Signal and Image Processing.** New Jersey, USA: Prentice-Hall, 1990.

[13] JAIN, A. K. **Fundamentals of Digital Image Processing.** New Jersey, USA: Prentice-Hall, 1989.

[14] INTERNATIONAL TELECOMMUNICATION UNION – RADIO COMMUNICATION SECTOR. **Recommendation ITU-R BT.601-5:** Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios. Genebra, Suíça, 1995.

[15] INTERNATIONAL TELECOMMUNICATION UNION – RADIO COMMUNICATION SECTOR. **Recommendation ITU-R BT.500-11:** Methodology for the subjective assessment of the quality of television pictures. Genebra, Suíça, 2002.

[16] TAN, T.; SULLIVAN, G.; WEDI, T. Recommended simulation commom conditions for coding efficiency experiments. **ITU-T Video Coding Experts Group ITU-T SG16 Q.16 Document VCEG-AE010**, 31st Meeting, Marrocos, Janeiro, 2007.

[17] COVER, T. M.; THOMAS, J. A. **Elements of Information Theory.** 2.ed. New Jersey, USA: John Wiley & Sons, 2006.

[18] PAPOULIS, A.; PILLAI, S. U. **Probability, Random Variables and Stochastic Processes.** 4.ed. New York, USA: McGraw-Hill, 2002.

[19] LIN, S.; COSTELLO, D. J. **Error Control Coding.** 2.ed. New Jersey, USA: Prentice-Hall, 2004.

[20] PRADHAN, S.; CHOU, J; RAMCHANDRAN, K. Duality between source coding and channel coding and its extension to the side information case. **IEEE Transactions on Information Theory,** v.14, n. 5, pp. 1181 – 1203, 2003.

[21] INTERNATIONAL TELECOMMUNICATION UNION – RADIO COMMUNICATION SECTOR. **Recommendation ITU-T H.263:** Video coding for low bit rate communications. Genebra, Suíça, 1996.

[22] INTERNATIONAL TELECOMMUNICATION UNION – RADIO COMMUNICATION SECTOR. **Recommendation ITU-T H.264:** Advanced Video Coding for Generic Audiovisual Services. Genebra, Suíça, 2003.

[23] GIBSON, J. D. et al. **Digital Compression for Multimedia: Principles and Standards.** Morgan Kaufmann, 2006.

[24] PRADHAN, S.; RAMCHANDRAN, K. Distributed source coding using syndromes (DISCUS): design and construction. **Proceedings of IEEE Data Compression Conference,** pp. 158 – 167, 1999.

[25] AARON, A.; GIROD, B. Compression with side information using turbo codes. **Proceedings of IEEE Data Compression Conference,** pp. 252 – 261, 2002.

[26] AARON, A.; ZHANG, R.; GIROD, B. Wyner-Ziv coding of motion video. **Proceedings of Asilomar Conference on Signals and Systems,** Novembro, 2002.

[27] AARON, A.; ZHANG, R.; GIROD, B. Transform-domain Wyner-Ziv codec for video. **Proceedings of SPIE Visual Communications and Image Processing,** v. 5308, pp. 520 – 528, San Jose, Janeiro, 2004.

[28] AARON, A.; VARODAYAN, D.; GIROD, B. Wyner-Ziv residual coding of video. **Proceedings of International Picture Coding Symposium,** Abril, 2006.

[29] AARON, A.; VARODAYAN, D.; GIROD, B. Rate-adaptative distributed source coding using low-density parity-checks codes. **Proceedings of Asilomar Conference on Signals, Systems and Computing,** Novembro, 2005.

[30] PURI, R.; RAMCHANDRAN, K. PRISM: A new robust video coding architecture based on distributed compression principles. **Allerton Conference on Communications, Control and Computing,** 2002.

[31] PURI, R.; RAMCHANDRAN, K. PRISM: A reversed multimedia coding paradigm. **Proceedings of IEEE International Conference on Image Processing,** v. 1, pp. I – 617 – 20, Setembro, 2003.

[32] PURI, R.; MAJUMDAR, A.; RAMCHANDRAN, K. PRISM: A video coding paradigm with motion estimation at the decoder. **Proceedings of IEEE International Conference on Image Processing,** v. 16, n. 10, pp. 2436 – 2448, 2007.

[33] GALLAGER, R. G. **Low-density parity-check codes.** Cambridge, USA: MIT Press, 1963.

[34] CLARKE, R. **Transform Coding of Images.** San Diego, USA: Academic Press, 1990.

[35] RAO, K. R.; YIP, P. **Discrete Cosine Transform: Algorithms, Advantages and Applications.** Boston, USA: Academic Press, 1990.

[36] AARON, A.; VARODAYAN, D.; GIROD, B. Rate-Adaptative Codes for Distributed Source Coding. **EURASIP Signal Processing Journal, Special Section on Distributed Source Coding,** v. 86, n. 11, Novembro, 2006.

[37] HE, Z. Bringing wireless video sensor networks into practice. **SPIE,** Bellingham, USA, Abril, 2007. Disponível em: <http://spie.org/x14634.xml?ArticleID=x14634>. Acesso em: 22 jun. 2010.

[38] PRESSMAN, R.S. **Engenharia de Software.** 6.ed. São Paulo: McGraw-Hill, 2006.

[39] SOMMERVILLE, I. **Engenharia de Software.** 8.ed. São Paulo: Pearson Education, 2007.

[40] LARMAN, C. **Utilizando UML e padrões: uma introdução à análise e projeto orientados a objetos e ao desenvolvimento iterativo.** 3.ed. Porto Alegre: Bookman, 2007.

[41] DEITEL, P.; DEITEL, H. **Java como programar.** 8.ed. São Paulo: Pearson Education, 2010.

[42] HORSTMANN, C.S.; CORNELL, G. **Core Java 2.** 7.ed. Rio de Janeiro: Alta Books, 2005.

[43] MARTINEZ, J.L.; FERNÁNDEZ-ESCRIBANO, G.; KALVA, H.; FERNAND, W. A. C.; GARRIDO, A. Wyner-Ziv to H.264 video transcoder for mobile telephony. **Proceedings of IEEE International Conference on Consumer Electronics,** Las Vegas, USA: Janeiro, 2009.

[44] WILBURN, B.; JOSHI, N.; VAISH, V.; ANTUNEZ, E.; BARTH, A.; ADAMS, A.; HOROWITZ, M.; LEVOY, M. High imaging using large camera arrays. **Proceedings of International Conference on Computer Graphics and Interactive Techniques,** Los Angeles, USA: Janeiro, 2005.

[45] XIONG, Z.; CHENG, S. Distributed Source Coding for Sensor Networks. **IEEE Signal Processing Magazine,** v. 21, n. 5, Setembro, 2004.

[46] CARAPEZZA, E. M. Sensors, and Command, Control, Communications, and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense VIII. **Proceedings of SPIE,** Março, 2004.

[47] FELIN (Fantassin à Équipements et Liaisons Intégrés) - Future Infantry Soldier System, France. Disponível em <http://www.army-technology.com/projects/felin>. Acesso em 22 jun. 2010.

[48] ARTIGAS, X.; ASCENSO, J.; DALAI, M.; KLOMP, S; KUBASOV, D; OUARET, M. The DISCOVER codec: Architecture, Techniques and Evaluation. **Picture Coding Symposium 2007**, Lisboa, Portugal: Novembro, 2007.

[49] TIOBE Programming Community Index. Disponível em <http://www.tiobe.com/index.php/content/paperinfo/tpci/index.html>. Acesso em 05 fev. 2011.

[50] Unified Modeling Language. Disponível em <http://www.uml.org>. Acesso em 05 fev. 2011.

[51] Oracle Virtual Box. Disponível em

<http://www.oracle.com/technetwork/server-storage/virtualbox/overview>.

Acesso em 06 fev. 2011.

[52] HARTE, L. **Introduction to MPEG: MPEG-1, MPEG-2 and MPEG-4.** Fuquay-Varina, USA: Althos Publishing, 2006.

APÊNDICE A
Fundamentos de Codificação de Vídeo

A.1
Vídeo Digital

O vídeo digital bem como a digitalização de todos os outros tipos de mídia é uma tendência universal. O vídeo digital tem muitas vantagens, entre as quais podemos citar [8]:

- Possibilidade de transmissão sem degradação ou com aceitável e até controlável razão sinal-ruído;
- Capacidade de melhoria e flexibilização da resolução na recepção;
- Capacidade de manter a qualidade de reconstrução, reduzindo ao máximo a quantidade de dados armazenados ou transmitidos;
- Capacidade de implementação de multimídia, fazendo a interação da informação do vídeo com dados como tabelas, gráficos, jogos, etc;
- Capacidade de implementação de interação com o usuário;
- Capacidade de implementação de técnicas de tratamento do vídeo através de programas computacionais.

Cronologicamente, essa tendência à digitalização do vídeo se deu de forma massiva com o advento do DVD (*Digital Versatile Video*) e a mobilização comercial mundial de suas empresas criadoras, a saber, Toshiba, Philips, Sony e Time Warner, para universalizar o formato da mídia física e dos padrões utilizados, como por exemplo, o padrão MPEG-2 [9] que ficou muito conhecido depois do sucesso do DVD. Atualmente, novos formatos como *Blu-Ray* aparecem como promessa e candidato a sucessor do DVD.

Assim, vídeo digital é uma representação eletrônica de uma sequência de imagens. As imagens que formam o vídeo são chamadas de quadros ou *frames*. Ao número de *frames* apresentados por segundo numa apresentação de um vídeo damos o nome de taxa de quadro ou *frame rate*.

Uma cena de vídeo natural é espacial e temporalmente contínua [10]. Assim, a digitalização do vídeo requer uma amostragem tanto temporal quanto espacial dessa cena contínua. A amostragem temporal divide o vídeo em imagens estáticas, os quadros ou *frames*, conforme anteriormente explicado. A amostragem espacial divide cada quadro do vídeo em pontos, os *picture elements* ou *pixels* no formato aglutinado. A quantidade de *pixels* em cada quadro do vídeo define sua resolução espacial, enquanto a quantidade de quadros por segundo definidos para a apresentação do vídeo define sua resolução temporal, mais conhecida como taxa de quadro ou *frame rate*.

A amostragem espacial e temporal de uma sequência de vídeo é ilustrada na figura A.1.

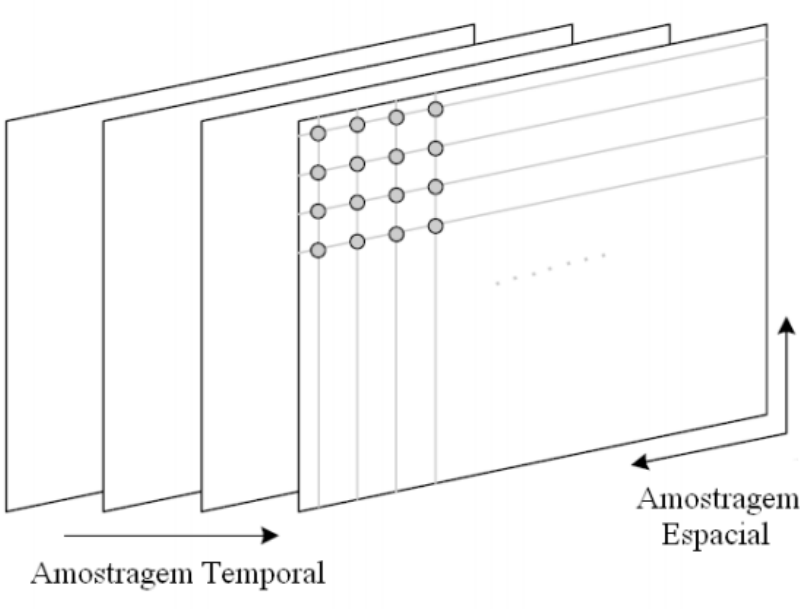


Figura A.1 – Ilustração dos conceitos de Amostragem Temporal e Amostragem Espacial numa sequência de vídeo
Fonte: Ref [10]

Quanto à noção de espaço ocupado por um vídeo, considere uma amostragem típica de um padrão num formato amplamente conhecido e utilizado, o DVD para o formato PAL (*Phase Alternating Line*) [9], cuja resolução espacial é de 720 x 576 *pixels* para cada *frame* e cuja resolução temporal ou *frame rate* é de 25 quadros por segundo. Neste cenário, cada quadro possui 414.720 *pixels* e cada segundo de vídeo 10.368.000 *pixels*. Se cada pixel for representado de forma

bruta por 24 bits, 8 bits para cada cor primária, um segundo de vídeo seria armazenado como 248.832.000 de bits ou 29,66 megabytes (MB). Um filme de noventa minutos teria 1.343.692.800.000 bits ou 156,43 gigabytes (GB), ou seja, precisaríamos de dez mídias de DVD dupla face e dupla camada para se armazenar um filme de vídeo digital bruto, já que cada uma armazena no máximo 17,08 GB.

Esses números tornariam inviável qualquer utilização prática de vídeo digital mesmo com as vantagens que já vimos em relação a sua robustez e manutenção da qualidade com o tempo. Assim, para se utilizar de forma prática o vídeo digital, foi necessário o desenvolvimento e utilização de técnicas de compressão de vídeo, o que era aplicável, já que estavam tratando de informações digitais, armazenando uma informação com a menor quantidade possível de bits.

Uma vez comprimido o vídeo, o processo de descompressão é inversível, dando origem a uma apresentação idêntica à original ou pelo menos aceitável em termos de qualidade objetiva e subjetiva.

Quando a apresentação após a descompressão é idêntica a original, significa que foram utilizadas técnicas de compressão sem perdas (*lossless*), onde a qualidade é maximizada, porém, a taxa de compressão não é muito alta. Quando a apresentação após a descompressão é aceitável do ponto de vista da qualidade subjetiva e objetiva, considerando as limitações do Sistema Visual Humano, significa que foram utilizadas técnicas de compressão com perdas (*lossy*). Devido a requisitos de banda e armazenamento, a compressão de vídeo é geralmente realizada com técnicas com perdas, alcançando taxas de compressão maiores, necessárias para a utilização prática desse tipo de informação.

A.2
Sistema Visual Humano

Muitas explicações e técnicas da codificação de vídeo digital foram desenvolvidas e aprimoradas a partir do estudo de vários aspectos e características de como capturamos, interpretamos e registramos tudo ao nosso redor do ponto de vista neuro-oftalmológico, o que ficou conhecido como Sistema Visual Humano ou simplesmente SVH.

O olho humano é uma verdadeira máquina biológica com a finalidade de capturar e processar imagens naturais coloridas, com muitos instrumentos, cada um com sua finalidade nesse processamento [11], conforme ilustra a figura A.2.

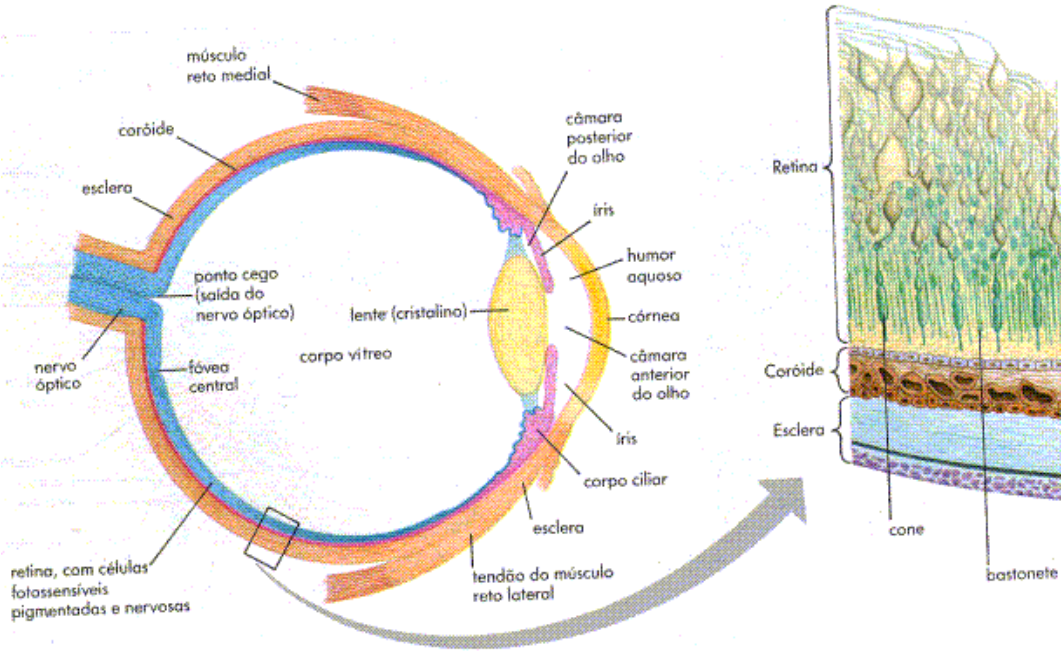


Figura A.2 – Sistema Visual Humano – SVH. Detalhes fisiológicos
Fonte: Ref [11]

Entre os componentes do SVH, a retina é uma rede de sensores óticos que transformam a informação luminosa em algo que possa ser transmitido para o cérebro através dos neurônios e possui dois tipos de sensores, os cones e os bastonetes. Os cones são os responsáveis pela captação das cores, funcionam com altos níveis de iluminação e são em número aproximado de 6,5 milhões em cada olho. Os bastonetes distinguem os detalhes em preto e branco, funcionam a partir de baixos níveis de iluminação e são em número aproximado de 100 milhões em cada olho. A córnea, a pupila, a íris e o cristalino são responsáveis por focalizar a imagem na retina.

Tanto o número de sensores na retina quanto o sistema de focalização do olho humano limitam a resolução do nosso sistema ótico e essas informações foram consideradas no desenvolvimento das técnicas de compressão.

A teoria tricromática da visão humana [12] afirma que as três classes de células cone do olho humano (foto-receptores de cor) têm sensibilidades diferentes aos comprimentos de onda do espectro visível, mas são principalmente sensíveis aos comprimentos de onda correspondentes às cores vermelho, verde e azul. Essa sensibilidade também é variável de acordo com o gráfico da figura A.3.

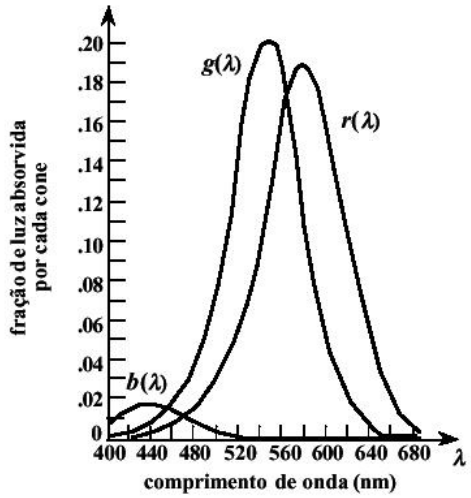


Figura A.3 – Teoria tricromática da visão humana – variação da absorção do comprimento de onda nas cores primárias
Fonte: Ref [12]

Assim, a partir do ponto de vista de interpretação do olho humano, qualquer representação pontual de uma figura ou quadro, ou seja, um *pixel*, pode ser definida por uma combinação dessas três cores primárias, dando origem ao conceito de espaço de cores.

A.3
Espaço de Cores e Subamostragem

Espaço, modelo ou sistema de cores é a definição da formação e distribuição de cores que compõe uma imagem e suas partes atômicas. No caso do vídeo a imagem pode ser definida como o quadro e as partes atômicas os *pixels*.

Devido aos motivos expostos na seção anterior desta dissertação, onde foram explicados alguns dos mecanismos do Sistema Visual Humano, o cérebro humano compõe quaisquer das milhares de cores que pode interpretar através da

combinação das três classes de células cones. Assim, a forma mais primária de simular as cores naturais e fazer sua captação digital é se baseando na teoria de visão tricromática de Young-Helmholtz, representando uma cor como uma combinação das três cores primárias vermelho, verde e azul, sendo este sistema chamado de modelo de cores RGB (do inglês *Red*, *Green* e *Blue*).

Para esse sistema, a tradução computacional é a representação de uma determinada cor através da definição dos níveis das três cores RGB numa escala de 0 a 255 para cada cor, sendo armazenado o nível de cada uma delas em 8 bits e definindo a cor representativa de um *pixel* como uma sequência de 24 bits. Isso gera um espaço de cores de 256 x 256 x 256, que resulta em mais de 16 milhões (16.777.216 ou 256^3) diferentes combinações de tons, saturação e brilho, o que dificilmente consegue ser distinguido pelo olho humano médio. Assim, nesse sistema, cada *pixel* será representado por 24 bits, o que do ponto de vista do tamanho de armazenamento parece inviável, conforme vimos nos cálculos que fizemos na seção A.1, quando mostramos a necessidade de se comprimir um vídeo digital bruto [13].

Alguns fatores convergiram para que esse sistema evoluísse para novos modelos de representação de cores mais adequados aos processos e técnicas de codificação.

Primeiramente, devido a limitações tecnológicas, os sistemas de televisão inicialmente eram preto e branco, representando apenas níveis de cinza. Após isso, houve o surgimento da tecnologia para televisão colorida, mas o sinal precisava manter a compatibilidade com a televisão preto e branco legada. A solução foi criar um sinal primário que continha somente a informação em preto e branco, que seria codificado tanto pela tv colorida quanto pela preto e branco e outros dois sinais com a informação de cores, que seriam descartados por esta última.

Fatores do SVH também contribuíram para a criação desse novo modelo. Analisando novamente o gráfico da sensibilidade do olho humano diante dos vários comprimentos de onda do espectro visível (figura A.3), observamos que o pico da sensibilidade para a cor verde é aproximadamente 5% maior que o pico para a cor vermelha e que estas duas sensibilidades são da ordem de dez vezes maiores do que a sensibilidade para a cor azul. Assim, os detalhes de alguns

comprimentos de onda são muito menos perceptíveis ao olho humano do que outros, descartando a necessidade de representá-los com a mesma quantidade de informação, fazendo com que o sistema de cores RGB seja desvantajoso, uma vez que representa as três componentes de cor com o mesmo número de bits.

Outra observação importante é que o olho humano é muito mais sensível à variações de brilho (*brightness*) que a de cores. Essa análise pode ser feita de forma simplista se considerarmos a quantidade de células bastonetes, na ordem de 100 milhões, responsáveis por identificar variações nos níveis de cinza ou brilho e a quantidade de células cones, na ordem de 6,5 milhões (bem menos que os bastonetes), responsáveis pela identificação das cores. Se quisermos ser mais precisos, podemos analisar o gráfico exato da distribuição da sensibilidade do olho humano em relação às componentes de luminância e crominância, conforme ilustrado na figura A.4, onde se mostra que o olho humano é muito mais sensível ao contraste da componente de luminância do que às duas componentes de crominância.

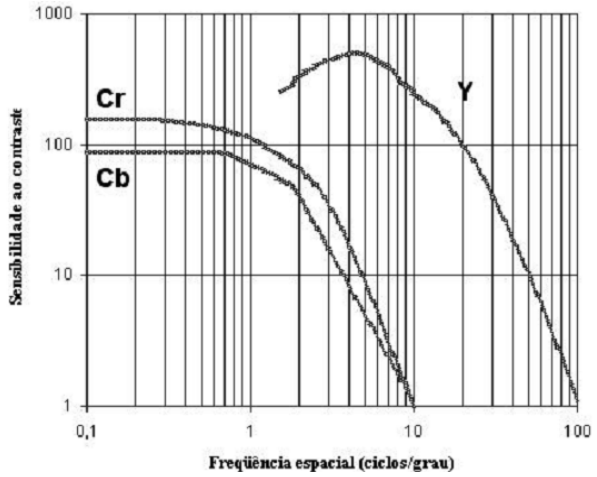


Figura A.4 – Sensibilidade do SVH ao contraste das componentes YC_BC_R
Fonte: Ref [13]

Esse novo sistema foi chamado de modelo YUV ou YC_BC_R, onde Y é a componente que representa os níveis de cinza ou luminância, que está relacionada à percepção de brilho da imagem e C_B e C_R ou U e V as duas componentes que estão associadas à percepção de saturação e matiz das cores. A composição RGB

de um determinado pixel pode ser convertida para o modelo YUV através da seguinte formulação matemática:

$$Y = 0.299R + 0.587G + 0.114B \quad (A.1)$$

$$U = 0.492\,(B - Y) \quad (A.2)$$

$$V = 0.877\,(R - Y) \quad (A.3)$$

Esse padrão e seus respectivos pesos são definidos na norma BT.601 da ITU-R [14] e resolve o problema da utilização de sinais comuns para a transmissão da televisão analógica colorida e preto e branco considerando o sistema PAL. Porém, esse modelo pode incorrer em valores negativos das componentes U e V e do ponto de vista computacional, principalmente se considerarmos o formato digital, pode ser mais interessante utilizarmos valores normalizados, somente com valores positivos. Esse modelo normalizado, mais utilizado para sistemas digitais, é conhecido como modelo YC_BC_R e tem a seguinte formulação matemática para quantização com 8 bits, ou seja, 256 níveis:

$$Y = +(77/256)R + (150/256)G + (29/256)B \quad (A.4)$$

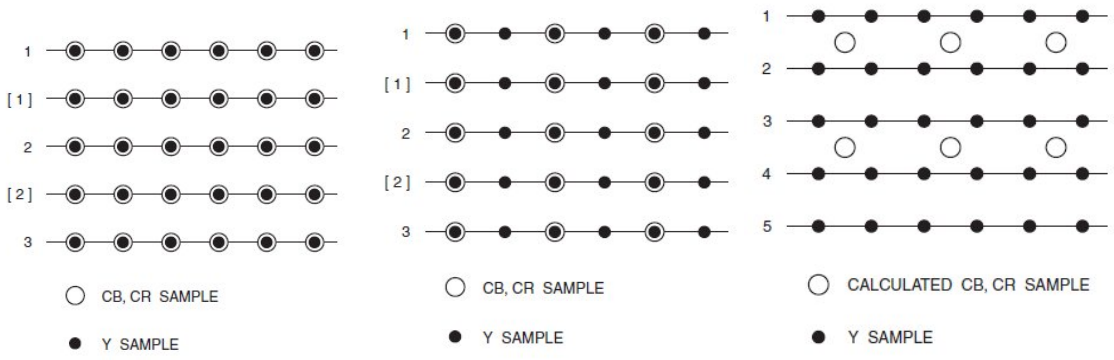
$$C_B = -(44/256)R - (87/256)G + (131/256)B + 128 \quad (A.5)$$

$$C_R = +(131/256)R - (110/256)G - (21/256)B + 128 \quad (A.6)$$

Dependendo da norma utilizada esses pesos podem ter pequenas alterações em seus valores, mas sempre seguindo o mesmo conceito.

Mesmo utilizando a transformação do modelo RGB para o modelo YC_BC_R, não teremos nenhuma vantagem do ponto de vista de compressão se essas três componentes forem quantizadas com 8 bits cada, utilizaríamos o mesmo espaço utilizado pelo modelo RGB. Assim, como o SVH tem uma sensibilidade muito menor às componentes de cores, como já explicamos anteriormente, podemos utilizar uma técnica chamada subamostragem para reduzirmos a quantidade de informação a ser codificada quanto às componentes de crominância C_B e C_R.

A possibilidade de amostragem das componentes de crominância com taxa inferior à utilizada para a componente de luminância gerou os três formatos de representação de vídeo mais utilizados, ilustrados na figura A.5.



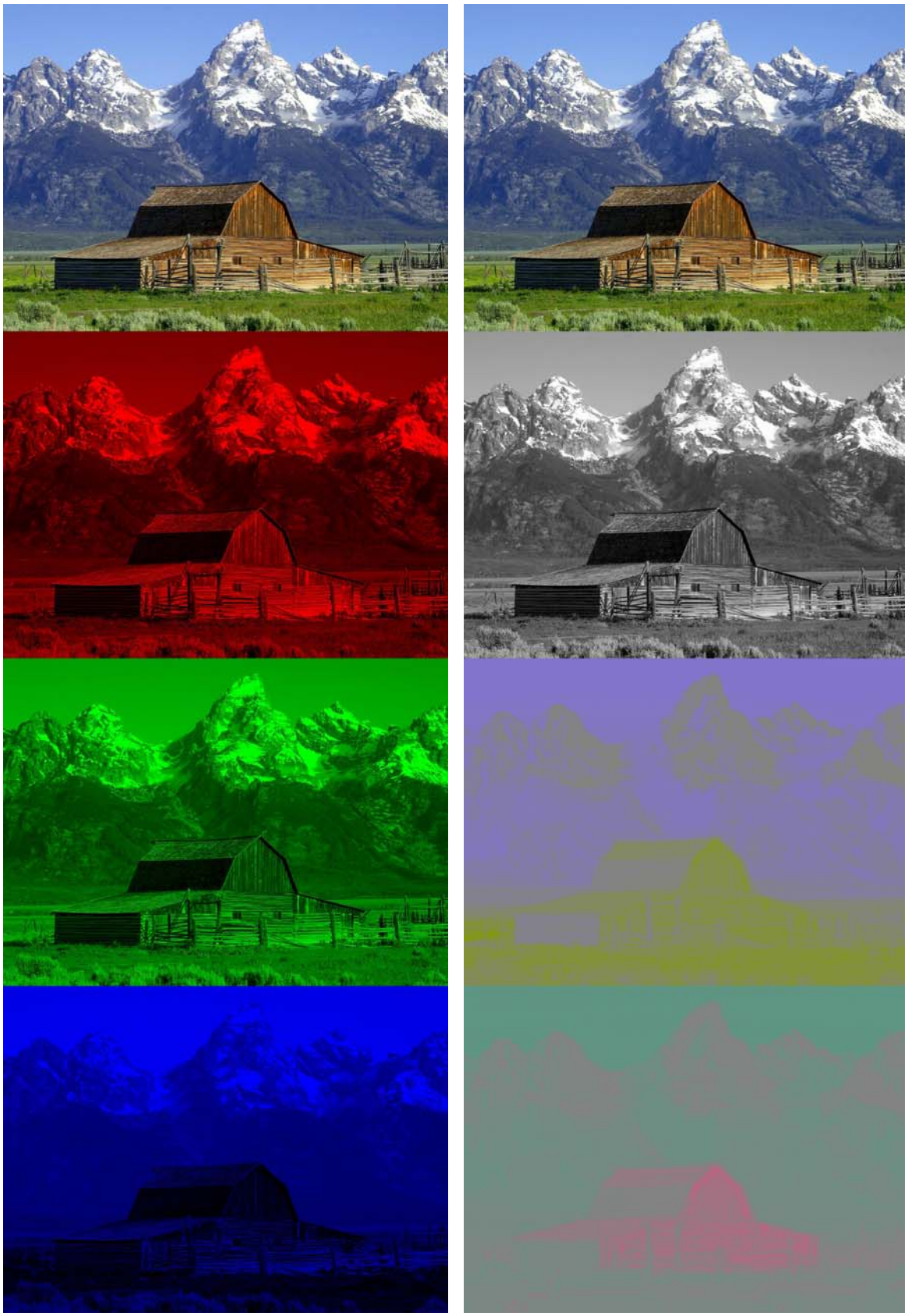
(a) YCBCR 4:4:4 (b) YCBCR 4:2:2 (c) YCBCR 4:2:0

Figura A.5 – Representação das componentes de luminância e crominância no modelo YC_BC_R. Em (a) temos o formato YC_BC_R 4:4:4, em (b) o YC_BC_R 4:2:2 e em (c) o YC_BC_R 4:2:0

Fonte: Ref [13]

No formato YC_BC_R 4:4:4, nenhuma informação é descartada, pois o número de amostras utilizado para as componentes de crominância é o mesmo utilizado para a luminância. Este formato é pouco utilizado porque gera uma representação com 24 bpp (bits por pixel), assim como o padrão RGB, não trazendo nenhuma vantagem de compressão. Já no formato YC_BC_R 4:2:2, as componentes de crominância são subamostradas por um fator de 2 na direção horizontal, gerando 16 bpp. Quando essa amostragem por um fator de 2 é também realizada na direção vertical, tem-se o formato YC_BC_R 4:2:0. Este é o formato utilizado na maioria dos padrões de codificação de vídeo, já que gera apenas 12 bpp, conseguindo a menor compressão comparado aos demais e uma boa qualidade dadas as limitações do SVH.

Um exercício muito interessante é a visualização em separado dessas componentes de luminância e crominância numa figura colorida e compará-las com a visualização individual dos componentes RGB, como ilustrado na figura A.6.



(a) Original colorida, canal R, G e B. (b) Original colorida, canal Y, C_B e C_R.

Figura A.6 – Ilustração com as figuras residuais a partir da original colorida, em relação às componentes individuais ou canais R, G e B em (a) e Y, C_B e C_R em (b).

A.4
Resolução e Formatos de Vídeo

Assim como numa imagem estática, a resolução espacial de um vídeo se dá pela quantidade de *pixels* num quadro. Considerando sua distribuição horizontal e vertical, ou seja, a distribuição desses *pixels* em linhas e colunas, tem-se o conceito de formato de vídeo.

Existem vários formatos de vídeo, alguns muito utilizados pela indústria de produção de monitores e televisões como os ilustrados na figura A.7.

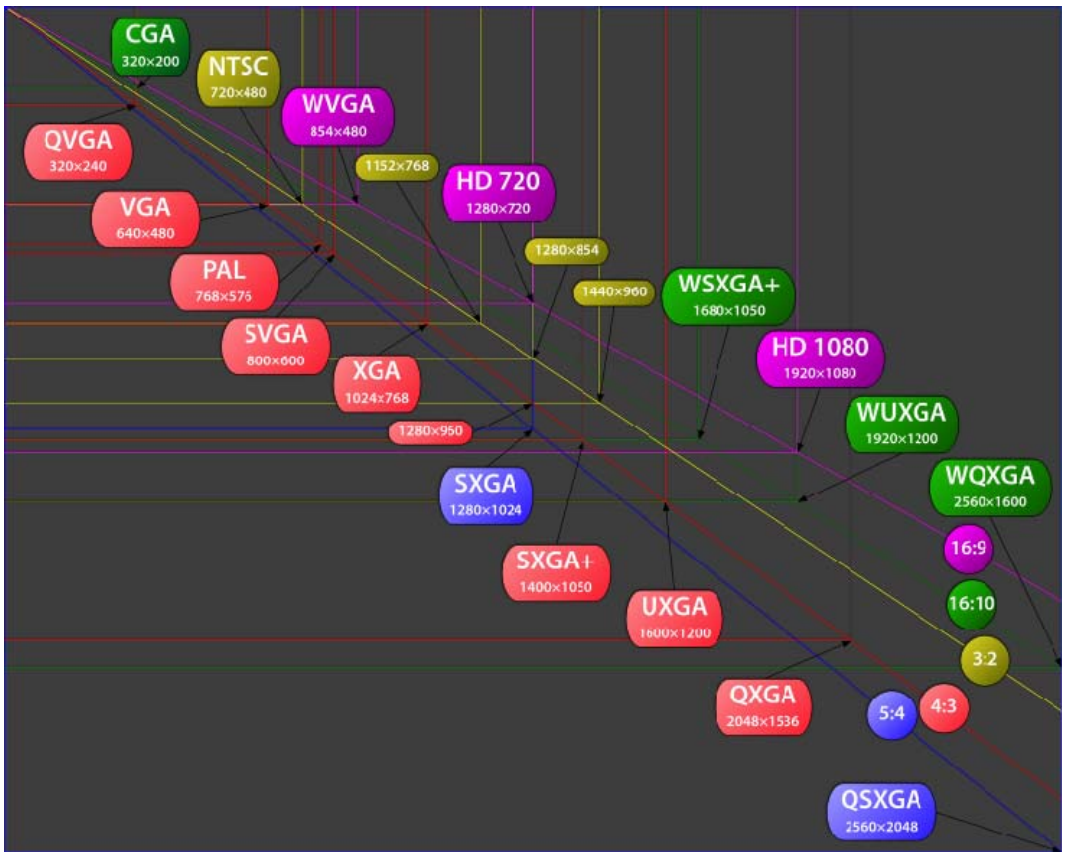


Figura A.7 – Padrões de formatos e dimensões de vídeo em produtos comerciais (números apenas informativos, não estão em escala real)

Além dessa distribuição dos *pixels* em linhas e colunas, temos também o conceito de *Aspect Ratio*, como podemos ver na mesma figura, que é a proporção entre a altura e a largura dos pixels que compõem uma imagem digital. A proporção mais antiga e conhecida é a famosa 4:3, das televisões CRT, que aos

poucos vem sendo substituída pelo novo formato 16:9 dos novos televisores de plasma e LCD, que valoriza mais o cenário de um vídeo num formato conhecido genericamente como *widescreen*.

Existem outros formatos que são utilizados pelos codificadores de vídeo, sendo que cada padrão oficial, por exemplo, os da ISO ou ITU, define um ou vários desses formatos para seu funcionamento. Uma das famílias de formatos mais utilizadas em codificadores são os múltiplos do formato CIF (*Common Intermediate Format*) [10], que têm suas configurações definidas conforme a Tabela A.1 para amostragem dos componentes de luminância e crominância seguindo o padrão YC_BC_R 4:2:0 e que são ilustrados na figura A.8.

Formato	**Resolução (*em pixels*)**	**Bits por quadro**	**Aplicação**
Sub-QCIF	128 x 96	147.456	Aplicações multimídia móveis
QCIF	176 x 144	304.128	Videoconferência e aplicações multimídia móveis
CIF	352 x 288	1.216.512	Videoconferência
4CIF	704 x 576	4.866.048	Padrões de televisão e DVD

Tabela A.1 – Formatos de vídeo da família CIF e informações comparativas

A.5
Modelo Geral de Compressão e Codificação

Com as ferramentas, recursos e padrões apresentados até agora, já se consegue reduzir consideravelmente o volume de armazenamento e a banda de transmissão, como será mostrado nos cálculos a seguir.

Considerando-se um vídeo no formato padrão de DVD com 720 colunas por 480 linhas, ou seja, 345.600 pixels, com subamostragem YC_BC_R 4:2:0, ou seja, 12 bpp com taxa de quadro de 30 quadros por segundo e com noventa minutos de duração. Um segundo desse vídeo ocupará aproximadamente 119 Mb (megabits) e o vídeo completo em torno de 78 GB (gigabytes), necessitando de cinco mídias de DVD dupla face e dupla camada para se armazenar esse filme com essas configurações, já que cada uma armazena no máximo 17,08 GB. Reduziu-se consideravelmente o tamanho do vídeo, mas ainda é inviável do ponto de vista de

armazenamento do vídeo e também se for analisada a ocupação de banda passante para a transmissão desse vídeo.

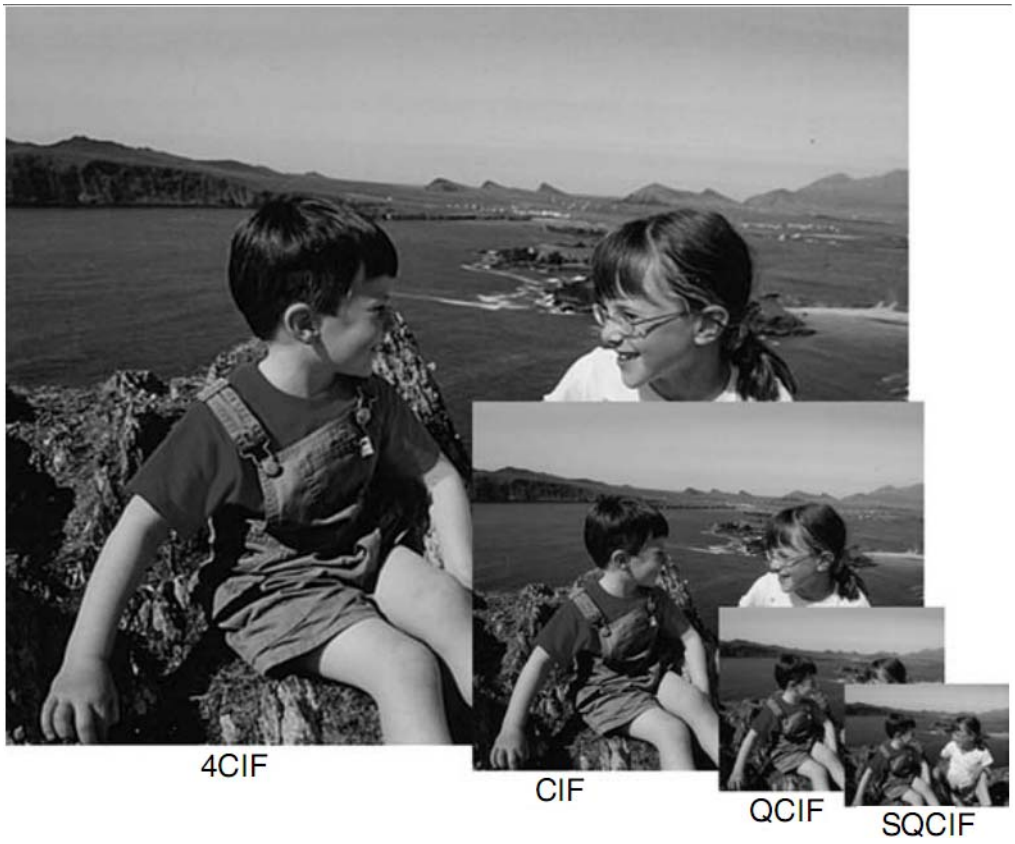


Figura A.8 – Formatos de vídeo da família CIF
Fonte: Ref [10]

Portanto, é preciso melhorar o mecanismo de compressão. O que foi apresentado até agora, foram transformações e conversões matemáticas que rearranjam o vídeo capturado para um formato minimamente comprimido, mas somente após essa conversão de RGB para YC_BC_R e subamostragem, no caso para o formato 4:2:0, é que realmente começa o processo de compressão para se retirar as redundâncias espaciais e temporais do vídeo.

A seguir, serão explicadas as técnicas utilizadas pelos codificadores de vídeo, em inglês *encoders*, para realizar tal redução [8].

Considerando uma figura simples conforme a figura A.9, que representa um círculo num determinado tom de cinza com um fundo branco de tamanho 256 por 256 pixels, para se representar essa figura de maneira bruta são necessários 256 x 256 x 8 = 524.288 bits. Porém, outra forma de representar essa mesma figura de

forma descritiva sem perder nenhuma informação seria: um círculo com raio de 60 pixels, nível de cinza 120, fundo branco e centro em x = 127 e y = 127. O raio é armazenado com 8 bits, o nível de cinza do círculo com 8 bits, a cor do fundo com 8 bits, a posição do centro em x com 8 bits e em y com mais 8 bits, ou seja, no total gastou-se 40 bits para se fazer a mesma representação, conseguindo uma taxa de compressão de mais de 13.000 para 1. Isso seria equivalente ao que o *encoder* vai fazer, pois apesar das imagens num vídeo não serem tão simples como o nosso exemplo, as imagens naturais possuem estatisticamente uma quantidade muito grande de redundâncias.

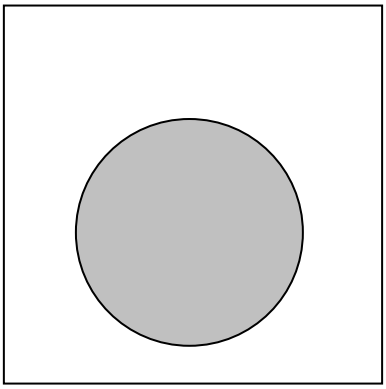


Figura A.9 – Círculo em tom de cinza com fundo branco

Assim, o objetivo do codificador será realizar uma compressão para reduzir a redundância de uma sequência de imagens e da própria imagem estática, armazenando-as e transmitindo-as num formato com a menor quantidade possível de bits e recuperando-as com a melhor qualidade possível após passar pelo canal de comunicação, pois a banda disponível para esse canal sempre será pequena se compararmos ao tamanho do vídeo bruto diante da necessidade da taxa de transmissão deste.

Analisando alguns canais como, por exemplo, o de uma WLAN, onde temos uma taxa disponível de 10 Mbps aproximadamente, o ATM, onde a taxa varia de 4 a 8 Mbps, o ISDN, onde a taxa é um múltiplo de 64 Kbps, conclui-se que nunca a taxa será suficiente, visto que é preciso aproximadamente 120 Mbits para um segundo de vídeo bruto, como foi calculado acima. Seria como "passar um camelo pelo buraco de uma agulha", no caso, o camelo é o vídeo, a agulha tem na entrada do seu buraco o codificador que comprime o camelo e faz com que ele passe no

restante do buraco, que é o canal, que no fim tem outro mecanismo que faz com que o camelo comprimido se transforme novamente no camelo original, ou algo bem parecido. Esse mecanismo se chama decodificador ou *decoder* em inglês. Quando são desenvolvidos o codificador e o decodificador, ou seja, o *encoder* e o *decoder*, trata-se do chamado *codec*, aglutinação de *encoder* e *decoder*.

Basicamente, o mecanismo de compressão utilizado pelos codificadores é composto por três processos conforme ilustrado na figura A.10, onde primeiramente a imagem original passa por uma transformação, depois a saída dessa transformação passa por um processo de quantização e por último por um processo de codificação antes de ir para o canal, com uma quantidade de informação bem menor do que a original. Do outro lado do canal, o decodificador recebe o fluxo de vídeo (*video stream*) e faz todo o processo inverso executado no codificador.

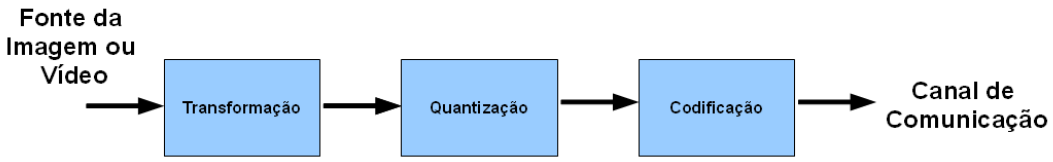


Figura A.10 – Blocos básicos envolvidos nos processos de compressão

Resumidamente, a transformação reduz a imagem a um conjunto de coeficientes menos redundantes e mais agrupados, matematicamente é como uma rotação algébrica. A quantização mapeia os coeficientes em um conjunto finito de símbolos e a codificação mapeia os símbolos em bits a serem enviados no fluxo de vídeo.

Cada modelo ou padrão de *codec* tem em sua especificação a definição das técnicas utilizadas em cada um desses processos da compressão. Nessa dissertação, foi definida uma proposta de implementação de um *codec* para vídeo distribuído.

A.6
Conceitos utilizados na Codificação de Vídeo

Por ocasião da codificação de vídeo digital alguns conceitos comuns são empregados durante o ciclo da codificação, independente do codec utilizado. Entre os principais conceitos, podemos destacar como básicos os definidos a seguir [52], sendo também utilizados na codificação do Open DVC:

- **Pixel**: um pixel é o menor componente de uma imagem. Os pixels podem variar em tamanho e forma e são compostos de cor e intensidade. O número de pixels por unidade de área é chamada de resolução. Quanto mais pixels por unidade de área maior é o nível de detalhes na imagem;

- **Blocos**: os blocos são partes de uma imagem dentro de um quadro de vídeo, normalmente definido por um número de pixels horizontais e verticais. Para os sistemas MPEG, por exemplo, cada bloco é composto por uma matriz de 8 por 8 pixels, sendo cada bloco processado separadamente.

- **Macroblocos**: um macrobloco é uma região de uma imagem em uma sequência de imagem digital (imagens em movimento) que podem ser utilizados para determinar a compensação de movimento a partir de um quadro de referência em uma sequência de imagens. Tipicamente, um quadro é dividido macroblocos de tamanho 16 por 16 pixels, sendo este o agrupamento de quatro blocos de tamanho 8 por 8 pixels;

- **Quadro**: um quadro é uma única imagem dentro da sequência de imagens que compõem o vídeo. Em um sistema de varredura entrelaçada de vídeo, um quadro é composto por dois campos. Cada campo contém metade das linhas de varredura que compõem a imagem, o primeiro campo normalmente contendo as linhas de varredura ímpares e o segundo campo tipicamente contendo as linhas de varredura pares. Para comprimir os sinais de vídeo, os sistemas de codificação categorizam as imagens de vídeo (quadros) em formatos diferentes. Estes formatos variam de tipos que só utilizam a compressão espacial (independentemente comprimidos) até quadros que utilizam a compressão espacial e temporal (quadros P). Os tipos de quadros dos sistemas de codificação incluem quadros de referência independentes (quadros I), quadros preditos que são baseados em quadros de referência anterior (quadros P), quadros bi-

direcionais preditos utilizando quadros anteriores e quadros posteriores (quadros B) e quadros DC (níveis de referência básica do bloco).

- **Quadros intra**: também chamados de **quadros I**, são imagens completas (fotos) dentro de uma sequência de imagens (como, por exemplo, em uma sequência de vídeo). Os quadros I são utilizados como referência para outros quadros de imagem comprimidos e são completamente independentes de outros quadros, inclusive na sua codificação. Assim, a única redundância que só pode ser explorada nos quadros I é a redundância espacial. Como consequência, os quadros I ocupam maior espaço ao serem comprimidos.

- **Quadros preditos**: também chamados de **quadros P**, são imagens (fotos) dentro de uma sequência de imagens (como em uma sequência de vídeo) que são criados a partir de informações de outras imagens, como os quadros I, por exemplo. Como os componentes de imagem são muitas vezes repetidos dentro de uma sequência de imagens (redundância temporal), o uso de quadros P proporciona uma redução substancial do número de bits que são utilizados para representar uma sequência de vídeo digital (compressão de dados temporal).

- **Quadros bi-direcionais**: também chamados de **quadros B**, são imagens (fotos) dentro de uma sequência de imagens (como em uma sequência de vídeo) que são criados usando informações de imagens anteriores e imagens posteriores (como os quadros I e quadros P, respectivamente). Como os quadros B são criados usando as duas imagens anteriores e imagens posteriores, eles oferecem uma capacidade de compressão de dados ainda maior do que os quadros P. Por causa da comparação que deve ser feita com os quadros anteriores e posteriores, a quantidade de processamento de imagem e conseqüentemente a carga de esforço computacional é tipicamente maior do que a aplicada em outros tipos de quadros.

- **Grupo de imagens**: os quadros podem ser agrupados em seqüências chamadas de grupo de imagens (**GOP**). Um GOP é uma seqüência de quadros que contêm todas as informações para que esses quadros sejam codificados e depois decodificados. Para todos os quadros dentro de um GOP que são construídos a partir de quadros de referência (quadros B e quadros P), os quadros referenciados e que servem de base para essa construção também devem estar contidos nesse mesmo GOP (quadros P e quadros I). Os tipos de quadros e sua localização dentro de um GOP podem ser definidos no tempo da seqüência. A distância temporal das

imagens é o tempo ou o número de imagens entre tipos específicos de imagens em um vídeo digital. M é a distância entre sucessivos quadros P e N é a distância entre sucessivos quadros I. Valores típicos para o GOP do sistema de codificação MPEG, por exemplo, são M igual a 3 e N igual a 12. A figura A.11 ilustra como os diferentes tipos de quadros podem compor um grupo de imagens (GOP). Como os quadros P e B são criados a partir de outros quadros, quando erros ocorrem em quadros anteriores, o erro pode se propagar através de quadros adicionais (retenção de erro). Para superar o problema da propagação de erros, quadros I são enviados periodicamente para atualizar as imagens e remover os erros existentes nos blocos. A figura A.12 ilustra como os erros que ocorrem em uma imagem podem ser mantidos nos quadros que se seguem. Este exemplo ilustra como os erros em um quadro B são transferidos para os quadros que se seguem e como são corrigidos com a atualização a partir de um quadro I.

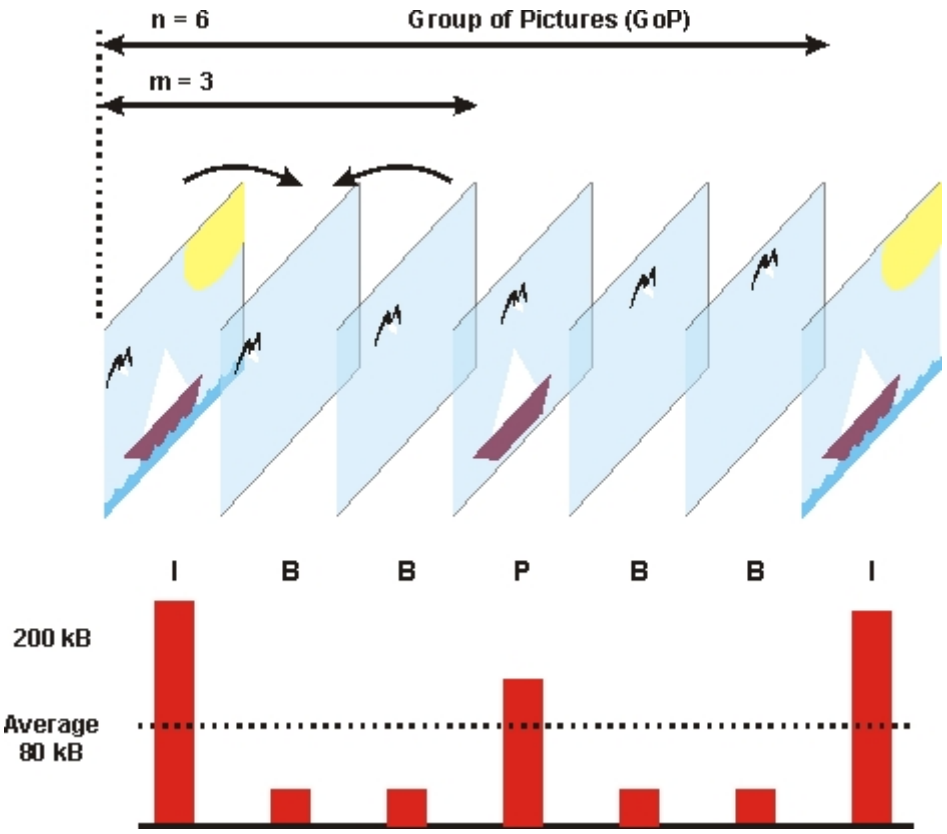


Figura A.11 – Grupo de imagens (GOP), exemplo do padrão MPEG
Fonte: Ref [52]

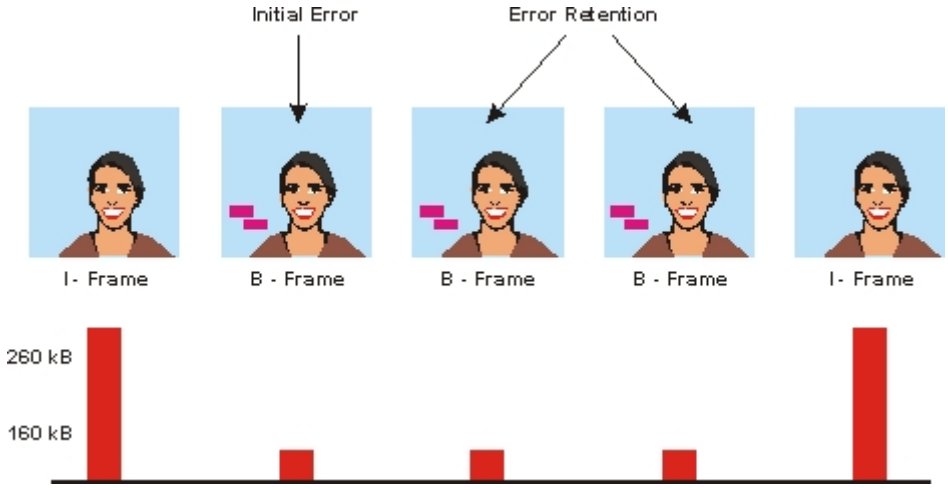


Figura A.12 – Retenção de erro e correção
Fonte: Ref [52]

- **Estimativa de movimento**: é o processo de busca de uma região fixa de um quadro anterior de vídeo para encontrar um bloco de pixels correspondente ao mesmo tamanho em relação ao quadro atual. O processo envolve uma busca exaustiva a vários blocos em torno do bloco atual a partir do quadro anterior. A estimativa de movimento é um processo computacional custoso que é utilizado para atingir altas taxas de compressão. A figura A.13 ilustra como um sistema de vídeo digital pode usar a estimativa de movimento para identificar objetos e como eles mudam de posição em uma série de imagens. Este diagrama mostra um pássaro em uma imagem voando em várias posições. Em cada quadro de imagem, o sistema de estimativa de movimento procura por blocos que se aproximam de outros blocos em imagens anteriores. Ao longo do tempo, o sistema de vídeo digital de estimativa de movimento encontra correspondências e determina os caminhos (vetores de movimento) que esses objetos tomar.

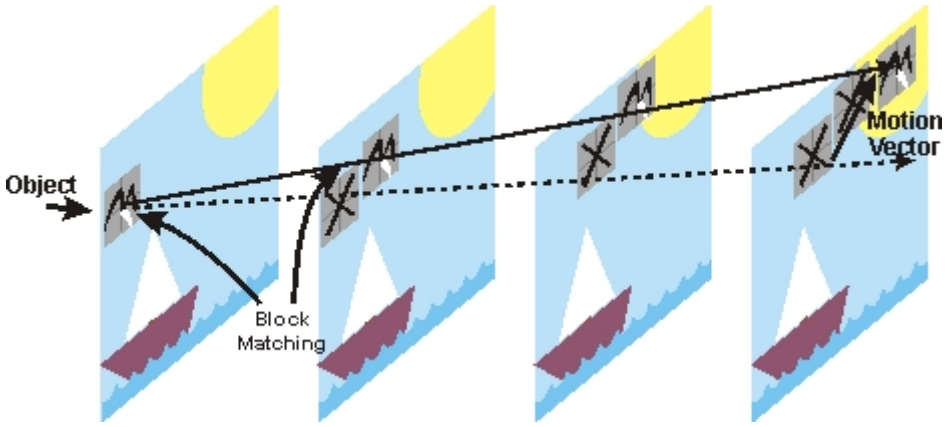


Figura A.13 – Estimativa de movimento
Fonte: Ref [52]

A.7
Técnicas de Codificação utilizadas na Arquitetura do Open DVC

Nas próximas subseções, iremos detalhar as técnicas utilizadas na arquitetura do *codec* Open DVC, tanto no seu aspecto matemático quanto no seu aspecto operacional.

A.7.1
Transformadas DCT e IDCT

A transformada DCT aplicada aos blocos de um quadro Wyner-Ziv, tem a finalidade de compactar a energia do bloco em uma quantidade pequena de coeficientes, explorando a redundância espacial do vídeo.

A transformada de Karhunen-Loève [34], conhecida como KLT, é a transformada ótima em termos de capacidade de compactação de energia. Porém, a KLT tem alguns problemas de ordem prática, por exemplo, a transformada KLT é dependente do sinal, ou seja, as funções-base da KLT são dependentes do sinal a ser transformado. Para a maioria das aplicações, a Transformada Discreta Coseno, cujo acrônimo DCT é originado do nome da operação em inglês *Discrete Cosine Transform*, é uma aproximação muito utilizada da transformada KLT [35], pois independe do sinal. Muitos padrões de codificação de vídeo, como o H.264/AVC [22], utilizam a DCT como solução para reduzir a redundância espacial.

No caso da codificação de vídeo, é utilizada a transformada DCT bidimensional aplicada a uma matriz de amostras *n x n*. No codificador Wyner-Ziv, a transformada é aplicada a blocos 4 x 4 originados a partir dos quadros Wyner-Ziv, da esquerda para a direita e de cima para baixo, podendo ser expandida para blocos 8 x 8 ou 16 x 16, se necessário.

A matriz de transformação H é definida como:

$$H = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{bmatrix} \tag{A.7}$$

A matriz de transformação inversa $\widetilde{H}_{inv}$ é definida como:

$$\widetilde{H}_{inv} = \begin{bmatrix} 1 & 1 & 1 & 1/2 \\ 1 & 1/2 & -1 & -1 \\ 1 & -1/2 & -1 & 1 \\ 1 & -1 & 1 & -1/2 \end{bmatrix} \tag{A.8}$$

Depois de aplicada a transformada DCT a um bloco de amostras de tamanho 4 x 4, as 16 amostras correlatadas dentro do bloco 4 x 4 são convertidas em 16 coeficientes independentes no domínio da frequência espacial. Estes coeficientes DCT são arranjados em um bloco 4 x 4 chamado de bloco de coeficientes DCT. A figura A.14 ilustra um bloco 4 x 4 de coeficientes DCT.

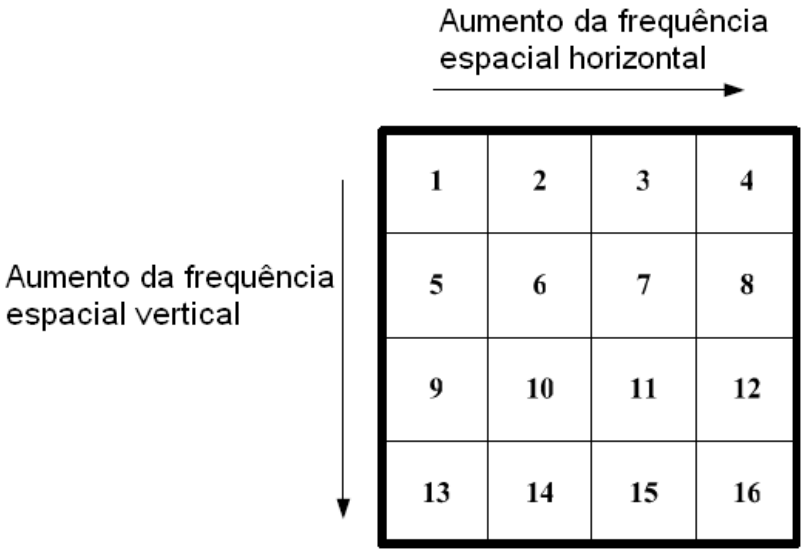


Figura A.14 – Bloco 4 x 4 com a posição dos coeficientes DCT

O coeficiente no canto superior esquerdo é o chamado coeficiente DC, corresponde à frequência zero e é o coeficiente que carrega a maior parte da informação. Os quinze coeficientes restantes são os coeficientes AC e correspondem a frequências espaciais diferentes de zero, sendo o coeficiente do lado inferior direito, na posição 16 da figura, correspondente à frequência espacial mais alta.

Todos os blocos 4 x 4 dos quadros X_{2i} (quadros Wyner-Ziv, quadros pares da sequência de vídeo) são submetidos à transformada DCT, sendo os coeficientes DCT agrupados de acordo com a posição ocupada dentro do bloco transformado, conforme a ilustração anterior, dando origem às Bandas de Coeficientes DCT.

A Banda b_k de Coeficientes DCT é um conjunto de todos os coeficientes DCT que ocupam a k-ésima posição em cada bloco 4 x 4 de coeficientes DCT. A primeira banda de coeficientes DCT b_1 corresponde à banda de coeficientes DC de cada bloco 4 x 4 (posição 1 da figura A.14) enquanto a banda de coeficientes DCT b_{16} corresponde à banda de coeficientes AC de maior frequência (posição 16 da figura A.14).

A.7.2
Quantização e Reescala

Depois da transformada DCT, cada Banda b_k de Coeficientes DCT é independentemente codificada. O primeiro passo dessa codificação é a quantização através de um quantizador uniforme [34].

A banda de coeficientes DC é caracterizada por valores positivos de alta amplitude, já que cada coeficiente DC da transformada expressa a energia média do bloco 4 x 4. A figura A.15 ilustra o quantizador escalar uniforme utilizado no procedimento de quantização dos coeficientes DC, onde v representa o eixo dos coeficientes DC, os números 0, 1, 2, ... acima do eixo v simbolizam os índices dos intervalos de quantização e W a largura do intervalo de quantização.

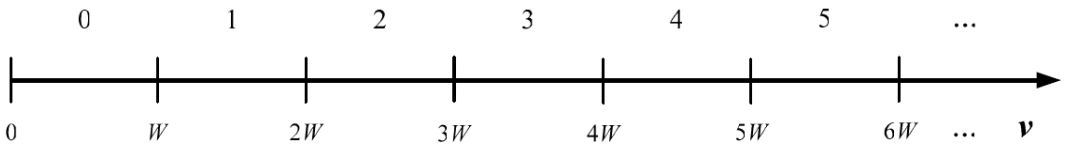


Figura A.15 – Quantizador escalar uniforme com largura W e índices positivos para o coeficiente DC

Para as bandas AC, a entrada do quantizador assume tanto valores positivos quanto negativos, pois as funções-base associadas aos coeficientes AC apresentam seu valor médio zero.

A figura A.16 ilustra o quantizador utilizado no procedimento de quantização dos coeficientes AC, onde valores positivos de coeficientes DCT são mapeados em intervalos de quantização com índices maiores ou iguais a zero e valores negativos são mapeados para intervalos de quantização com índices negativos.

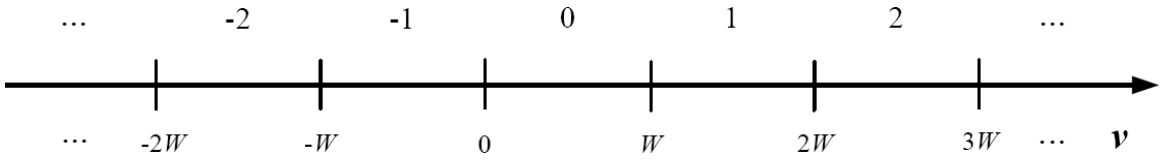


Figura A.16 – Quantizador escalar uniforme com largura *W* e índices positivos e negativos para os coeficientes AC

Conforme já vimos, Y_{2i}^{DCT}, que são os quadros formados a partir da informação lateral, são estimativas de X_{2i}^{DCT} (quadros Wyner-Ziv originais depois da DCT) calculadas a partir de quadros de origens diferentes, podendo existir erros associados a esta diferença de geração entre os dois quadros. Caso esses erros não sejam corrigidos pelo código corretor de erro, artefatos de bloco serão verificados no quadro decodificado X'_{2i} (quadro Wyner-Ziv reconstruído). Os erros entre os coeficientes Y_{2i}^{DCT} e X_{2i}^{DCT} são mais acentuados para os coeficientes DCT em torno da amplitude zero.

Nesta região, pode acontecer que um dado coeficiente Y_{2i}^{DCT} e o correspondente em X_{2i}^{DCT} tenham diferentes sinais, conforme ilustrado na figura A.17 que exemplifica este cenário, onde o coeficiente Y_{2i}^{DCT} é mapeado para um intervalo de quantização -1 e o coeficiente correspondente em X_{2i}^{DCT} é mapeado para o intervalo de quantização 0. Se depois da decodificação do canal, o intervalo de quantização decodificado for -1, ou seja, indicação de que o erro não foi corrigido, o efeito dos artefatos de bloco será visto no quadro decodificado.

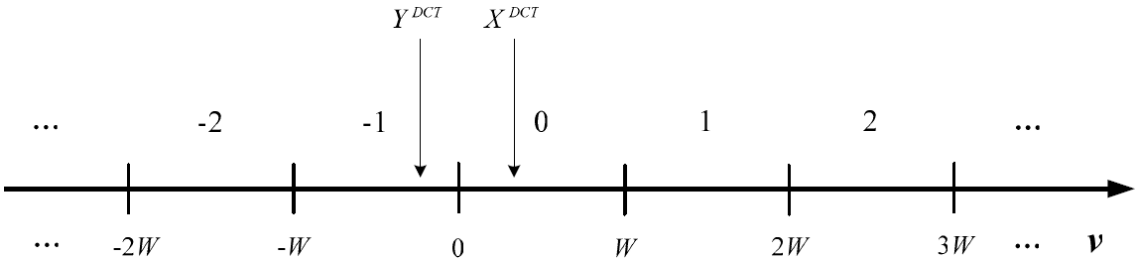


Figura A.17 – Erro na quantização dos coeficientes AC devido à simetria do eixo *v*

Para reduzirmos esses efeitos podemos utilizar um quantizador com um intervalo de quantização simétrico em torno do zero, como ilustrado na figura A.18. Os valores dos coeficientes DCT em torno de zero são deste modo,

quantizados sob o mesmo índice de intervalo de quantização independentemente do seu sinal, evitando erros entre os símbolos quantizados de Y_{2i}^{DCT} e os correspondentes em X_{2i}^{DCT}, reduzindo assim o efeito dos artefatos de bloco.

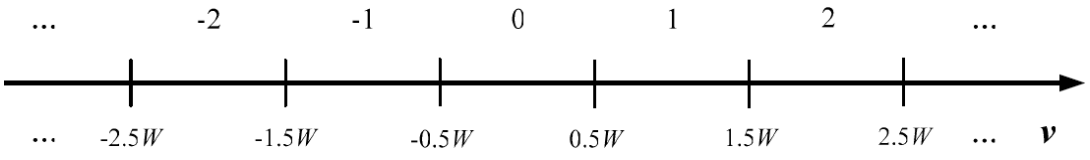


Figura A.18 – Quantizador escalar uniforme com um intervalo de quantização simétrico em torno da amplitude zero

Os coeficientes AC, dentro de um bloco de coeficientes DCT 4 x 4, normalmente têm uma amplitude maior quanto mais próximos ao coeficiente DC e uma amplitude menor nos coeficientes localizados em posições de frequências espaciais maiores. Em termos de bandas, isto significa que as bandas AC designadas com índices mais próximos de 1 (índice da banda DC) têm amplitudes mais altas comparadas às bandas AC designadas com índices distantes da banda DC.

De fato, dentro de um bloco de coeficientes DCT 4 x 4, as frequências espaciais menores possuem informação mais relevante sobre o bloco do que as altas frequências, as quais geralmente correspondem a ruído ou a detalhes menos importantes na interpretação feita pelo sistema visual humano (SVH).

Como o olho humano é mais sensível à baixas frequências espaciais, os coeficientes DCT que representam essas frequências são quantizados utilizando tamanhos menores de passo de quantização, ou seja, com um maior número de níveis de quantização. As frequências espaciais mais altas são quantizadas de forma menos precisa, ou seja, com menos níveis de quantização, sem diminuir significantemente a qualidade visual subjetiva da imagem decodificada.

A escolha do número de níveis de quantização associado a cada banda de coeficientes DCT é uma forma importante de explorar a sensibilidade visual humana para frequências mais baixas quando comparada a frequências mais altas, conseguindo assim uma forma de reduzir a quantidade de informação a ser enviada sem comprometer a qualidade subjetiva do vídeo.

Para cada banda b_k, calcula-se a faixa dinâmica, isto é, aquela na qual os coeficientes DCT variam, além do número de níveis de quantização associado àquela banda. O tamanho do passo de quantização, necessário para definir os limites dos intervalos de quantização, pode ser calculado a partir desses dois parâmetros. No decodificador DVC, a faixa dinâmica é conhecida para cada banda de coeficientes DCT.

A vantagem de se utilizar uma faixa dinâmica ao invés de usar um valor fixo é permitir larguras de intervalos de quantização ajustadas ao tamanho de cada banda. Com a faixa dinâmica menor do que uma faixa fixa utilizada (mais ajustada) o mesmo número de níveis de quantização é distribuído sobre um intervalo mais curto, tendo como consequência, passos de quantização mais curtos e menor distorção em relação ao erro inserido pela quantização. Na arquitetura do Open DVC, a faixa dinâmica para cada banda DCT é transmitida quadro a quadro para o decodificador.

O passo de quantização para cada banda DCT b_k é calculado pela divisão da faixa dinâmica da banda b_k pelo número de níveis de quantização 2^{M_k}. O limite superior do valor do coeficiente DC é definido por:

$$\sqrt{n^2}\, I_{máx} \tag{A.9}$$

Onde n^2 é o número de *pixels* em um bloco n x n e $I_{máx}$ é a intensidade máxima do *pixel*. Como o tamanho do bloco utilizado é de 4 x 4 *pixels* com precisão da informação de 8 bits, o limite superior da banda DC, que nesse caso coincide com sua faixa dinâmica, é fixo e igual a 1024, ou seja, 4×2^8.

Para o cálculo do passo de quantização das bandas AC b_k, k variando de 2 a 16, precisamos determinar o maior valor absoluto dentro de cada banda, que corresponderá à metade da faixa dinâmica total da banda b_k. Assim, o tamanho do passo de quantização W é definido da seguinte forma:

$$W = \frac{2\left|V_k\right|_{máx}}{2^{M_K} - 1} \tag{A.10}$$

Onde $|V_k|_{máx}$ é o maior valor absoluto dentro da banda b_k e 2^{M_k} é o número de níveis de quantização.

O índice do intervalo de quantização q, também conhecido como símbolo quantizado, é definido como:

$$q = \frac{V_k}{W} \tag{A.11}$$

Onde V_k é o valor do coeficiente DCT da banda b_k e W é o tamanho do passo de quantização associado a esta banda.

E assim, unindo todas as técnicas e cálculos explicados, o processo de quantização do *codec* Open DVC pode ser definido da seguinte forma: cada banda b_k de coeficientes DCT é quantizada utilizando um quantizador escalar uniforme com 2^{M_k} níveis. O parâmetro M_k corresponde ao número de bits necessários para mapear cada coeficiente DCT da banda b_k para um dos 2^{M_k} níveis de quantização associados a esta banda.

Vale observar que cada valor M_k tem uma taxa-distorção associado. Assim, diferentes índices de qualidade podem ser alcançados variando o valor de M_k para a banda b_k. Na arquitetura proposta, consideraremos dezoito níveis de taxa-distorção, associados aos Vetores de Níveis de Quantização da Matriz de Quantização da figura A.19 e que variam o nível de qualidade desejado na decodificação do vídeo. O vetor escolhido influenciará não só no nível de qualidade, mas também na taxa de transmissão da informação. Com a utilização do vetor (a) atingimos a taxa de bits mais baixa e consequentemente a maior distorção, por outro lado, com a utilização do vetor (r) temos um cenário com taxa de bits mais alta e, conseqüentemente, a menor distorção. Nos Vetores de Quantização 4 x 4 da figura A.19, o valor na posição k da figura, k variando de 1 a 16 conforme foi definido na figura A.14, indica o número de níveis de quantização associado à banda b_k de coeficientes DCT. Os vetores de quantização são usados para determinar o desempenho da taxa-distorção do *codec* proposto e são assumidos como conhecidos tanto pelo codificador quanto pelo decodificador.

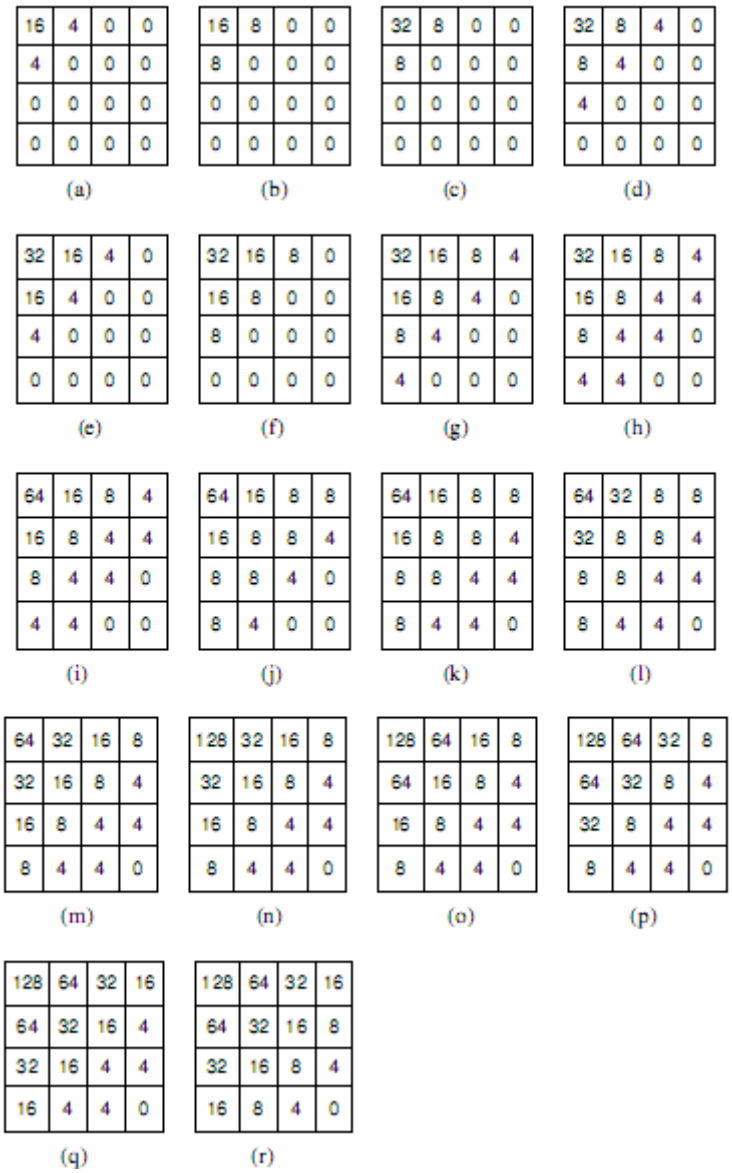


Figura A.19 – Os 18 Vetores de Níveis de Quantização da Matriz de Quantização definida para o Open DVC

O valor zero numa das posições das matrizes de quantização, significa que nenhum bit será transmitido para o decodificador nas bandas correspondentes. Essas bandas serão decodificadas apenas pelos coeficientes DCT da informação lateral, uma vez que são bandas que carregam uma parte menos significativa da informação.

Após uma banda b_k ter sido quantizada, seus símbolos quantizados (valores inteiros) são convertidos para uma sequência binária. Os bits são então agrupados formando os Planos de Bits. Os Planos de Bits são independentemente codificados

com o codificador LDPC utilizado. A codificação LDPC da banda b_k começa com o plano de bits correspondente aos bits mais significativos dos símbolos quantizados nesta banda. Após a codificação, a síndrome gerada é transmitida pelo Canal de Comunicações ao receptor, que executa os processos relativos à decodificação DVC, com o que seriam as operações inversas da codificação, iniciando, portanto, pela decodificação LDPC, tema que será detalhado adiante.

A.7.3
Codificador de Canal

No *codec* Open DVC, o codificador de canal, também chamado numa arquitetura DVC de codificador Slepian-Wolf, utiliza um código LDPC irregular acumulado, baseado nos estudos do LDPC da Universidade de Stanford, cujos autores são Bernd Girod, David Varodayan e Anne Aaron [36].

Os códigos LDPC são os que obtêm melhor desempenho se comparados, por exemplo, com códigos turbo, que são outra opção utilizada em DVC, chegando próximos à qualidade de *codecs* híbridos de alto desempenho.

O codificador LDPC utilizado consiste de um código LDPC formador de síndrome, concatenado a um acumulador, também conhecido como codificador LDPCA. Para cada plano de bits, bits de síndrome são criados utilizando o código LDPC e após isso, um acumulador módulo-2 é utilizado, produzindo o que chamamos de síndrome acumulada.

A.8
Qualidade de Vídeo

Como foi visto anteriormente, uma característica muito desejável para os codificadores de vídeo é a sua capacidade de compressão. Altas taxas de compressão só conseguem ser atingidas através de técnicas de compressão com perdas, ou seja, a informação após ser decodificada não é exatamente a mesma informação que foi codificada. Quem define qual a taxa de compressão desejável ou mínima para uma determinada técnica é justamente a aplicação que será

utilizada junto com os requisitos de banda disponível na rede de transmissão utilizada.

Supondo que dois *codecs* consigam a mesma taxa de compressão, para sabermos qual *codec* é mais eficiente, precisamos aplicar conceitos de qualidade de vídeo, onde se define uma forma para medir o quão fiel foi a informação que chegou do outro lado (no receptor) em relação à informação original ou ainda, se foi atingido o padrão de qualidade exigido no requisito de determinada aplicação, já que se pode abrir mão de qualidade até um determinado limite em casos extremos, onde a infra-estrutura de comunicação faz com que não se tenha banda o suficiente para se transmitir um vídeo com alta qualidade.

Assim, medir essa qualidade é necessário, mas não é simples. Vários fatores podem influenciar a percepção humana na medida da qualidade visual, o que pode ser um problema, já que muitas vezes essa percepção é muito mais subjetiva do que objetiva. Portanto, apesar de parecer razoável que os métodos se atenham a medidas matemáticas para se definir qual a melhor técnica para codificação de um vídeo, não se pode abrir mão das técnicas subjetivas, já que, por definição, a medida da qualidade visual é um aspecto essencialmente subjetivo, sendo influenciada por muitos aspectos não científicos como aspectos comportamentais, econômicos, biológicos e outros, além de aspectos visuais. Por exemplo, um telespectador pode estar assistindo a um vídeo de excelente qualidade, mas se por questão pessoal ele não gostar do assunto do vídeo, isso certamente influenciará na percepção particular de qualidade deste vídeo, do ponto de vista deste telespectador.

Tecnicamente falando, a percepção humana da qualidade visual de uma cena de vídeo considera aspectos da fidelidade temporal e espacial. Espacialmente falando, se percebe a nitidez em relação ao cenário de uma cena. Em relação ao aspecto temporal, somos sensíveis a movimentos bruscos, qualificando melhor movimentos mais suaves e naturais.

Além dos aspectos técnicos subjetivos, outros fatores irão interferir na opinião de um observador em relação à qualidade do vídeo, como o estado físico e mental do observador, a interatividade com o meio visual e o cenário visual. Assim, se o ambiente visual é confortável a opinião do observador em relação à

qualidade visual da cena que está sendo assistida terá uma nota melhor do que se o ambiente for desconfortável, independentemente da qualidade real do vídeo.

Outro importante fator que influencia na percepção da qualidade é a atenção visual ou a capacidade do observador se concentrar em um único ponto da cena ou dispersar sua atenção para todos os outros pontos simultaneamente. Caso esse observador se concentre num pedaço do *frame* que sofra alguma imperfeição na codificação ou decodificação isso será percebido, porém, caso ele divida sua atenção ao *frame* todo ele pode não perceber tais falhas. Um outro fenômeno chamado "efeito recente" mostra que nossa opinião é influenciada positivamente por cenas mais recentes do que por cenas antigas.

Enfim, em função de todos os fatores apresentados, percebe-se que medir a qualidade subjetiva de forma precisa é bastante difícil.

Para tentar ajudar nessa avaliação da qualidade subjetiva, alguns procedimentos de testes foram descritos sendo o mais conhecido o definido na recomendação BT.500-11 da ITU-R [15], chamado *Double Stimulus Continuous Quality Scale* (DSCQS) [10], no qual um observador é apresentado a um par de sequências de vídeo, uma de cada vez e depois emite uma nota de qualidade para cada vídeo num intervalo de 1 a 5, onde 1 é a nota ruim e 5 é excelente.

Normalmente, um observador é submetido ao teste de vários pares de sequências sendo uma das sequências o vídeo original e a outra sequência o mesmo vídeo codificado e depois decodificado antes de ser apresentado. Entre os pares a ordem entre o vídeo original e o decodificado é alterada aleatoriamente para não causar um vício visual, pois ao longo de vários pares se a ordem fosse mantida haveria chance do observador perceber a ordem do vídeo que é sempre o original e a do que é sempre o decodificado.

O teste DSCQS é bem aceito cientificamente como uma boa medida de qualidade visual subjetiva de vídeo, porém sofre de problemas práticos, pois o resultado pode variar dependendo do público observador e das sequências de vídeo sob teste. Esta variação pode ser compensada pela repetição do teste com uma quantidade significativa de pares de vídeos e com um universo numeroso e heterogêneo de observadores, indo desde especialistas em vídeo até leigos. Assim, isso faz do DSCQS um teste caro e que consome muito tempo se considerarmos as condições ideais para que seus resultados sejam cientificamente válidos.

Devido ao custo e complexidade envolvidos na medida subjetiva, a forma mais utilizada para medição da qualidade de vídeo é a medida de qualidade objetiva. Nesta, necessita-se unicamente de manipulações e cálculos matemáticos para obtê-la, os quais normalmente são realizados por meios computacionais.

A medida objetiva mais utilizada em vídeo é o PSNR [10, 16], acrônimo de *Peak Signal to Noise Ratio* em inglês. O PSNR é medido em escala logarítmica e tem relação com o cálculo do Erro Médio Quadrático, cujo acrônimo é MSE, originado a partir do nome da medida em inglês, *Mean Squared Error*. A medida do PSNR é dada pela equação A.12:

$$PSNR_{dB} = 10 \log_{10} \frac{(2^n - 1)^2}{MSE} \qquad (A.12)$$

A figura A.20 mostra três imagens, onde a imagem (a) é o quadro original e as imagens (b) e (c) versões com menor qualidade originadas a partir da codificação e decodificação da primeira. A imagem (b) tem um PSNR de 30,6 dB enquanto a imagem (c) tem um PSNR de 28,3 dB, sendo, portanto uma imagem de pior qualidade em relação às anteriores.

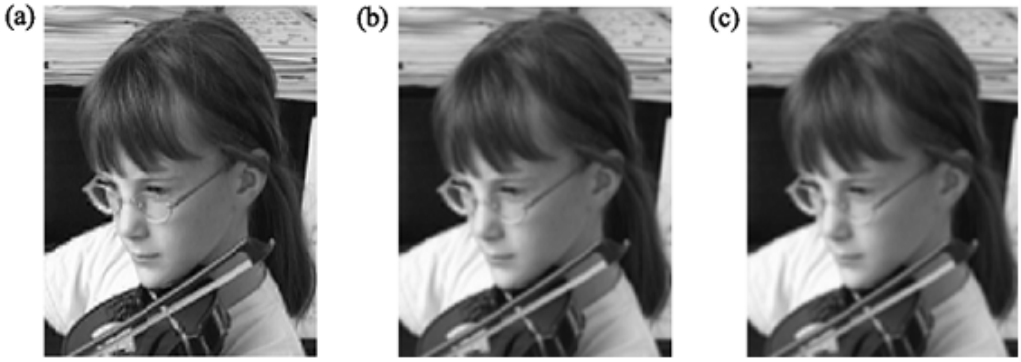


Figura A.20 – Comparações de PSNR: (a) foto original; (b) 30.6 dB; (c) 28.3 dB
Fonte: Ref [10]

A principal limitação em relação ao PSNR é que necessitamos da imagem original para fazermos o cálculo em relação à degradação de um vídeo bruto, mas não podemos garantir que ela estará sempre disponível. Outro detalhe, é que devido às características de interpretação neurológica do SVH, muitas vezes uma imagem com melhor qualidade objetiva pode ser classificada como uma imagem de pior qualidade na avaliação subjetiva, dependendo da área do quadro onde

estiver inserida a maior parte dos erros ou distorções. Se for numa área com mais detalhes e mudanças de níveis de cores, a interpretação humana colocará esta imagem como mais pobre devido ao posicionamento do erro chamar mais atenção do olho humano.

Assim, conclui-se que é importante do ponto de vista científico correlacionarmos análises e medições objetivas com as subjetivas para se garantir uma maior confiabilidade dos resultados e medidas do vídeo em questão e do *codec* que está sendo analisado.

APÊNDICE B
Teoria da Informação e Codificação de Fonte Distribuída

B.1
Fundamentos de Teoria da Informação

A teoria da informação estuda duas questões fundamentais e correlacionadas da compressão e codificação de dados [17]:

1. Qual a máxima compressão possível num sistema de codificação?
2. Qual a melhor taxa de transmissão num sistema comunicações?

Essa teoria tem seus estudos baseados no campo estatístico das variáveis aleatórias e processos estocásticos [18]. Uma variável aleatória é uma função mensurável que atribui valores numéricos únicos aos possíveis resultados de um experimento aleatório sob determinadas condições. Processos estocásticos permitem expressar matematicamente as relações entre suas variáveis aleatórias. Considera-se conhecido o embasamento matemático sobre variáveis aleatórias e processos estocásticos necessários para o entendimento da teoria da informação utilizada como fundamento deste trabalho.

Um conceito básico para o entendimento da teoria da informação é o conceito de fonte. O termo fonte é usado para indicar um processo que gera mensagens de informação sucessivas dentre um dado conjunto de mensagens possíveis. Uma fonte pode ser modelada como uma variável aleatória X que emite símbolos de um alfabeto χ e com função massa de probabilidade $p(x)$. Associada a cada fonte há uma entropia $H(X)$, que é uma medida da incerteza de uma variável aleatória. Em termos de teoria da informação, a entropia indica a média da informação que uma fonte possui, em bits por símbolo.

Para uma variável aleatória discreta X, com um alfabeto χ e função massa de probabilidade $p(x)$, $x \in \chi$, a entropia $H(X)$ é definida como:

$$H(X) = -\sum_{x \in \chi} p(x) \log_2 p(x) \tag{B.1}$$

A partir da definição de entropia é possível definir entropia conjunta e entropia condicional de duas variáveis aleatórias.

Sendo *Y* uma variável aleatória discreta com alfabeto γ, a entropia conjunta *H(X, Y)* de duas variáveis discretas *(X, Y)* com função massa de probabilidade conjunta *p(x, y)* é definida como:

$$H(X,Y) = -\sum_{y\in\gamma}\sum_{x\in\chi} p(x,y)\log_2\, p(x,y) \tag{B.2}$$

E a entropia condicional de *Y* dado *X* é expressa por:

$$H(Y \mid X) = -\sum_{x\in\chi}\sum_{y\in\gamma} p(x,y)\log_2\, p(x \mid y) \tag{B.3}$$

Além disso, com algumas manipulações algébricas, pode ser mostrado que a entropia conjunta de duas variáveis aleatórias é igual à entropia de uma delas mais a entropia condicional da outra, conforme mostrado abaixo:

$$H(X,Y) = H(X) + H(Y \mid X) \tag{B.4}$$

B.2
Teorema da Codificação de Fontes e Teorema da Taxa-Distorção

Nos estudos da Teoria da Informação desenvolvidos por Shannon em 1948 [1] existem três resultados fundamentais:

- Teorema da codificação de fontes;
- Teorema da taxa-distorção;
- Teorema da codificação do canal.

O teorema da codificação de fontes consiste na explicação e na comprovação da compressão sem perdas de uma fonte discreta, onde define também que fontes contínuas não podem ser reproduzidas sem perdas.

Neste teorema, foi provado que uma fonte discreta *X* pode ser reconstruída perfeitamente se e somente se foi transmitida com uma taxa R_X não menor do que a entropia *H(X)*, conforme mostrado na fórmula abaixo:

$$R_X \geq H(X) \tag{B.5}$$

A figura B.1 descreve um sistema de codificação, transmissão e decodificação de uma fonte sem perda de informação.

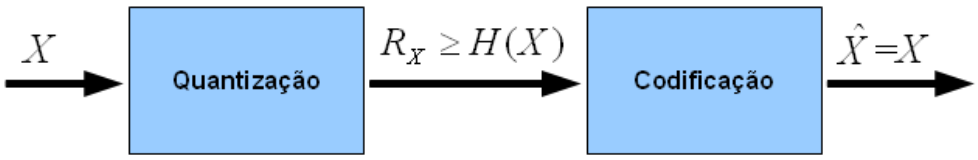


Figura B.1 – Compressão de fonte sem perdas. A fonte *X* deve ser transmitida com pelo menos *H(X)* bits.

A compressão com perdas é tratada no teorema da taxa-distorção. Seja a reconstrução do sinal *X* denotada por *X'* ou $\hat{X}$, então define-se a distorção *D* como $D = d(X; \hat{X})$, onde *d* é uma medida de distorção. Uma das medidas mais utilizadas é o Erro Quadrático Médio ou *Mean Square Error* em inglês, cujo acrônimo é MSE, definido como $E[(X - \hat{X})^2]$.

O teorema da taxa-distorção [17] prova que dada uma distorção *D* aceitável, existe uma taxa mínima R_X associada a esta distorção. A taxa é determinada pela função de taxa-distorção $R_X(D)$. A função $R_X(D)$ é convexa e retorna a mínima taxa para reconstruir *X* com uma distorção máxima *D*.

B.3
Teorema da Codificação do Canal

Um codificador de canal transforma uma entrada binária $\vec{i}$ em um código $\vec{c}$. A taxa do código é definida como $R_C = k/n \leq 1$, o qual especifica que uma entrada de tamanho *k* gera um código de tamanho *n*. Os códigos corretores de erro [19] ajudam a inferir a informação original $\vec{i}$ mesmo se o código $\vec{c}$ estiver corrompido. O teorema da codificação do canal [17] indica que para um valor real $\varepsilon > 0$ e uma taxa de codificação $R_C < C$, onde C é a capacidade do canal, existe um código *C* tal que a probabilidade de erro depois da decodificação é menor que ε. Se *X* é a entrada de um código discreto de canal sem memória que gera a saída *C*, a capacidade do canal é definida como:

$$C = max_{p(x)} I(X;C) \tag{B.6}$$

A capacidade do canal indica quanta informação pode ser transmitida por um canal com uma probabilidade de erro próxima a zero. Uma extensão dos teoremas da codificação de fontes e da codificação do canal é o teorema da codificação da fonte-canal [17], que estabelece a existência de um *codec* fonte-canal, o qual permite codificar uma fonte com entropia *H(X)* de maneira confiável num determinado canal se e somente se $H(X) < C$. No caso de codificação com perdas, sendo *D* a distorção permitida, pode-se obter um código com taxa $R(D) < C$, pois que $H(X) = R(0) \geq R(D)$.

Um código de canal sistemático é formado pelo vetor original de entrada mais uma informação extra chamada paridade. A paridade ajuda a corrigir o vetor original no processo de decodificação caso um erro de transmissão tenha ocorrido. Em um codificador sistemático $C = [X|P_d]$, P_d é a paridade e o operador "|" representa a concatenação de vetores.

B.4
Codificação de Fonte com Informação Lateral

Para o entendimento dos fundamentos e aplicações desse novo ramo da teoria da informação, a Codificação Distribuída de Fonte, do inglês *Distributed Source Coding*, cujo acrônimo é DSC, e sua aplicação na área de vídeo digital distribuído, conhecida como Codificação Distribuída de Vídeo, do inglês *Distributed Vídeo Coding*, cujo acrônimo é DVC, e que é alvo de estudo e implementação desta dissertação, é necessário entender o conceito de codificação de fonte com informação lateral, que é baseado nos resultados do Teorema de Slepian-Wolf [3] para codificação sem perdas e do Teorema de Wyner-Ziv [4] para a codificação com perdas. Essas teorias serão explicadas nas seções a seguir.

B.4.1
Teorema de Slepian-Wolf

O teorema da codificação de fontes de Shannon, que foi definido na seção B.2, pode ser expandido para a codificação conjunta de duas fontes *(X, Y)* com entropia conjunta *H(X, Y)*, se for tratada como a codificação de uma única fonte *Z* com entropia *H(Z) = H(X, Y)*. Assim, para se obter uma reconstrução sem perdas, tem que ser utilizada uma taxa $R_Z \geq H(Z)$. Esse mesmo problema pode ser interpretado de uma outra forma. Pode se transmitir a fonte *X* a uma taxa $R_X \geq H(X)$ e transmitir *Y* com uma taxa $R_Y \geq H(Y|X)$, baseado no conhecimento prévio de *X*. Como já fora explicado anteriormente, $R_X + R_Y \geq H(X, Y)$. O diagrama de codificação conjunta de duas fontes é ilustrado na figura B.2.

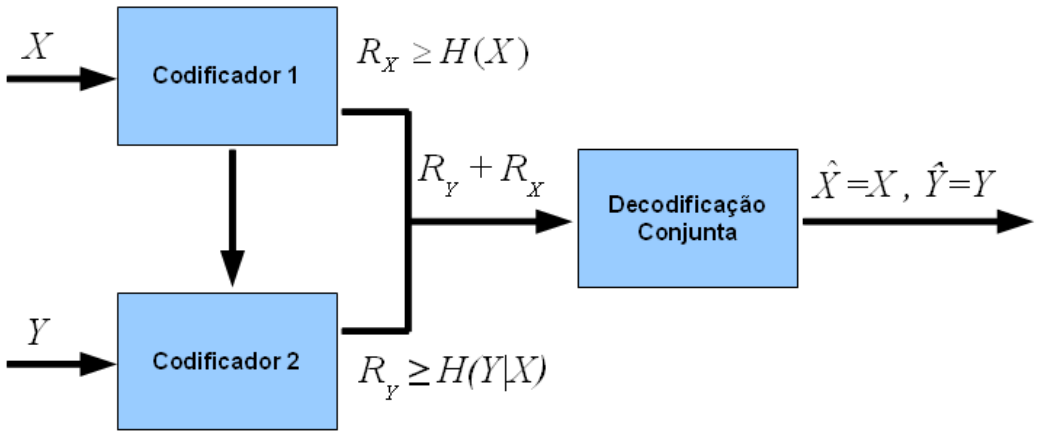


Figura B.2 – Codificação conjunta de duas fontes *X* e *Y*. Uma maior ou igual a *H(X, Y)* é suficiente para a reconstrução sem perdas

O Teorema de Slepian-Wolf de 1973 [3], conseguiu expandir a teoria de Shannon para a codificação separada de duas fontes correlatadas. De acordo com o teorema de Slepian-Wolf, duas fontes podem ser codificadas separadamente e reconstruídas sem perdas se as seguintes condições forem satisfeitas:

- A correlação estatística entre as fontes deve ser conhecida no decodificador;
- $R_X \geq H(X|Y)$;
- $R_Y \geq H(Y|X)$;
- $R_X + R_Y \geq H(Y, X)$.

Essas condições são ilustradas na figura B.3.

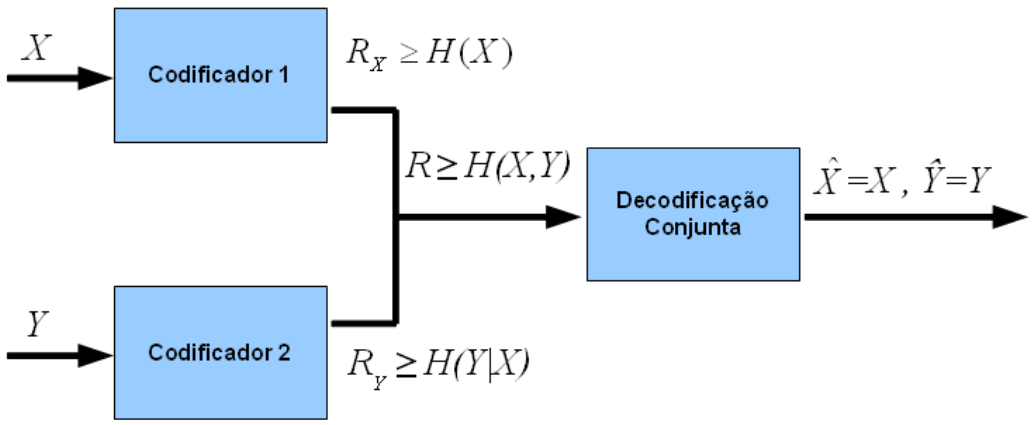


Figura B.3 – Codificação separada de duas fontes *X* e *Y* com decodificação conjunta. H(X,Y) continua sendo suficiente para reconstrução sem perdas

Este teorema aumenta a região de codificação com possível reconstrução sem perdas para duas fontes correlatas. A figura B.4 mostra a região de possível decodificação perfeita, ressaltando a região triangular entre os pontos A e B, que representa a expansão definida no teorema de Slepian-Wolf.

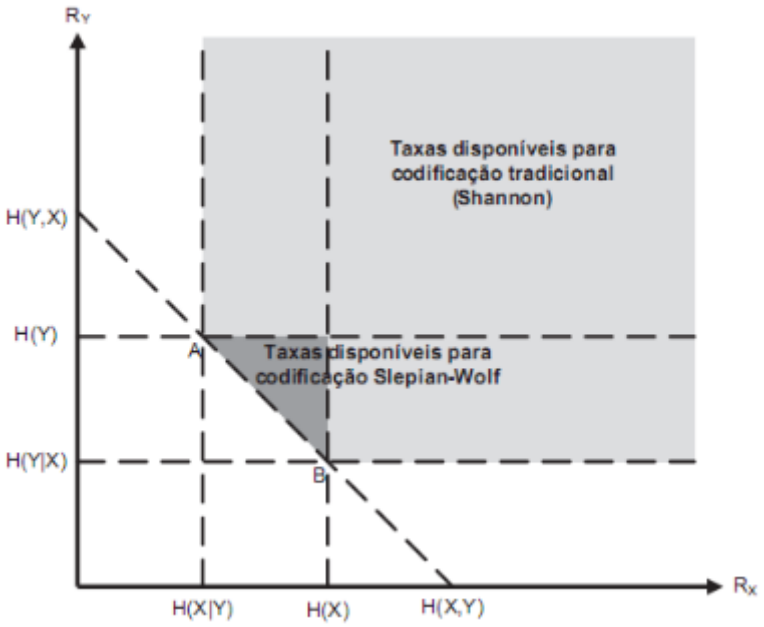


Figura B.4 – Codificação separada de duas fontes *X* e *Y* com decodificação

A codificação de fonte com informação lateral sem perda de informação é um caso particular da codificação de Slepian-Wolf. Pelo teorema de Slepian-Wolf, se a fonte *Y* existe somente no decodificador ou foi transmitida a uma taxa não menor que *H(Y)*, é possível codificar a fonte correlata *X* a uma taxa não menor que *H(X|Y)* para obter uma reconstrução perfeita $\hat{X} = X$. A fonte *Y* recebe o nome de informação lateral.

Na codificação com informação lateral $R_X = H(X|Y)$, o sistema está operando no ponto A da figura B.4.

O diagrama da codificação de fonte sem perdas com informação lateral é ilustrado na figura B.5.

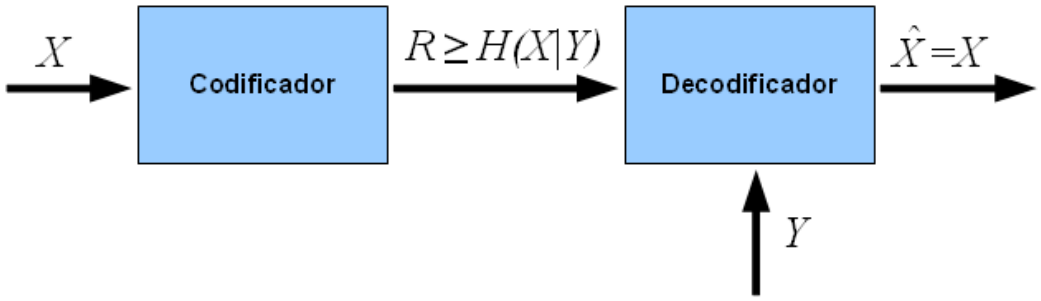


Figura B.5 – Codificação de fonte com informação lateral sem perdas

B.4.2
Teorema de Wyner-Ziv

Em 1976, Wyner e Ziv expandiram os resultados do Teorema de Slepian-Wolf para a codificação com perdas com informação lateral [4].

Assim como o teorema da codificação de fontes pode ser visto como um caso especial do teorema da taxa-distorção para $D = 0$, a codificação de fonte com informação lateral de Slepian-Wolf pode ser vista como o caso de distorção zero para a codificação de Wyner-Ziv.

Seja $R_{WZ}(D)$ a função taxa-distorção de Wyner-Ziv. Se $R_X(D)$ é a função taxa-distorção para codificar e decodificar a fonte X com um valor esperado de distorção D e $R_{X|Y}(D)$ é a função associada à codificação de X, dada a informação perfeita de Y, com distorção aceitável D, o teorema de Wyner-Ziv prova que $R_{WZ}(D) \geq R_{X|Y}(D)$ e $R_{WZ}(D) \leq R_X(D)$.

O caso importante para o DSC, e consequentemente para o DVC, é quando a informação lateral Y está presente somente no decodificador. Os estudos de Wyner-Ziv provam que nesse caso, quando é utilizada uma codificação com informação lateral, existe um cenário onde $R_{WZ}(D) = R_{X|Y}(D)$, com MSE como medida de distorção. Isto significa que a codificação com informação lateral iguala o desempenho do codificador com conhecimento perfeito de Y. Para isso, basta X ser uma fonte discreta ou contínua com distribuição de massa ou densidade, respectivamente, gaussiana e a informação lateral deve ser igual a $Y = X + Z$, onde Z é gaussiana e independente de X.

Mais tarde foi provado que é necessário somente que Z seja gaussiana [20]. Z é conhecido como ruído de correlação de Wyner-Ziv. O cenário da informação

lateral somente no decodificar é ilustrado na figura B.6 e é o caso estudado na implementação dos *codecs* DVC.

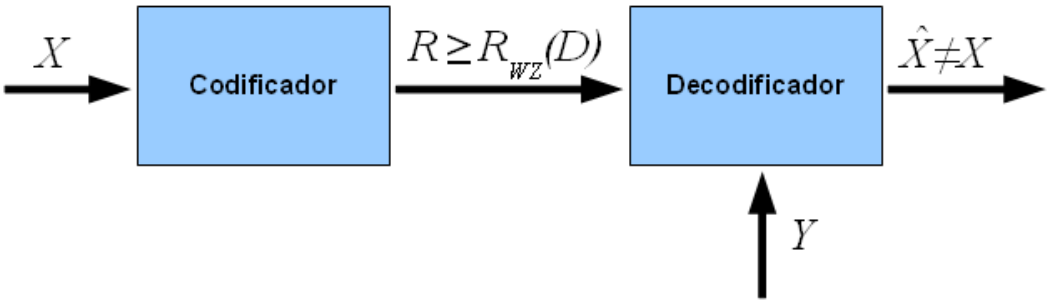


Figura B.6 – Codificação de fonte com informação lateral com perdas