# 2
# Conceitos iniciais

Neste capítulo apresentaremos alguns conceitos básicos sobre formatos e qualidade de vídeo, codificação de vídeo e medidas de energia de um bloco de pixels. Estes conceitos são estritamente necessários ao entendimento deste trabalho. Assuntos como formatos QCIF e CIF para sequências de vídeo, maneiras de medir a qualidade objetiva e subjetiva de vídeo, CODEC de vídeo, macrobloco e energia de um bloco serão apresentados nas próximas seções.



## 2.1.
## Tipos de formatos de seqüências de vídeo

O padrão H.264, que será apresentado no Capítulo 4, utiliza vários formatos de vídeo para compressão. O formato CIF (Common Intermediate Format) representa a base para outros formatos apresentados na Tabela 1. Ainda na Tabela 1 é apresentado as aplicações de cada formato. É apresentado na Figura 1 o componente de luminância de um quadro de vídeo amostrado nas resoluções: 4CIF, CIF, QCIF e SQCIF.

Um formato usado para codificação, indicando conversão para o formato digital sem compressão, de sinais de vídeo para televisão é o ITU-R Recommendation BT.601-5 [4]. A componente de luminância do sinal de vídeo é amostrado em 13,5 MHz e a de crominância em 6,75 MHz para produzir as componentes do sinal 4:2:2 Y:Cb:Cr. Os parâmetros do sinal digital amostrado dependem da taxa de quadros (30 Hz para um sinal NTSC e 25 Hz para um sinal PAL/SECAM) e são mostrados na Tabela 2. A taxa de bits total é a mesma em cada caso, pois a taxa de quadros de 30 Hz do NTSC é compensada por uma resolução espacial mais baixa. A área real mostrada no display, a área ativa, é menor do que o total, pois exclui intervalos de apagamento horizontais e verticais que existem fora da porção ativa do quadro.

| Formatos de vídeo | Resolução de luminância | Aplicação |
|---|---|---|
| **Sub-QCIF** | 128 x 96 | aplicações multimídia móveis |
| **Quarter CIF (QCIF)** | 176 x 144 | videoconferência e aplicações multimídia móveis |
| **CIF** | 352 x 288 | videoconferência |
| **4CIF** | 704 x 576 | padrão de televisão e DVD |

Tabela 1- Formatos de vídeo

| Parâmetro | Taxa de quadros: 30 Hz | Taxa de quadros: 25 Hz |
|---|---|---|
| **Campos por segundo** | 60 | 50 |
| **Linhas por quadro completo** | 525 | 625 |
| **Amostras de luminância por linha** | 858 | 864 |
| **Amostras de crominância por linha** | 429 | 432 |
| **Bits por amostra** | 8 | 8 |
| **Taxa de bits total** | 216 Mbps | 216 Mbps |
| **Linhas ativas por quadro** | 480 | 576 |
| **Amostras ativas por linha (Y)** | 720 | 720 |
| **Amostras ativas por linha (Cr, Cb)** | 360 | 360 |

Tabela 2- Parâmetros ITU-R BT.601-5 [1]

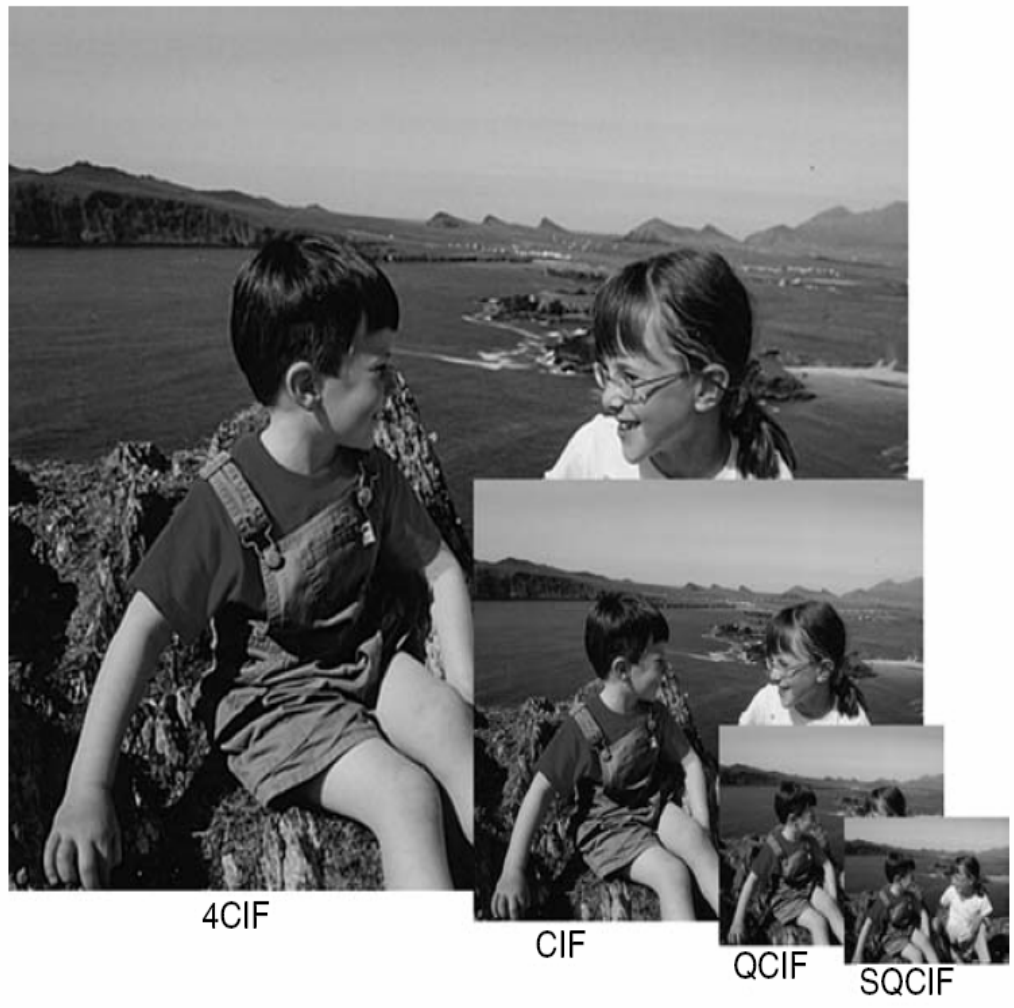


Figura 1-Quadros de vídeo amostrados nas resoluções: 4CIF, CIF, QCIF e SQCIF. Fonte: Ref [1]

## 2.2. Medida da Qualidade de vídeo

Determinar a qualidade das imagens de vídeo é importante para comparar e avaliar sistemas de vídeo. A medida da qualidade visual é essencialmente subjetiva sendo influenciada por muitos fatores. Por exemplo, um observador fornece uma opinião relativa à qualidade visual ao assistir passivamente um filme em DVD de forma diferente de ao participar ativamente em uma videoconferência.

Por outro lado, a medida da qualidade visual usando critérios objetivos gera resultados mais precisos, podendo ser repetidos.

### 2.2.1 Qualidade Subjetiva

Neste item apresentaremos os fatores que influenciam a qualidade subjetiva de uma sequência de vídeo e citaremos um procedimento de teste usado para avaliar esta qualidade.

A percepção humana da qualidade visual de uma cena de vídeo é afetada pela fidelidade espacial e temporal. A primeira refere-se a como nitidamente partes do cenário podem ser vistas e a segunda refere-se ao fato do movimento aparecer natural e suave.

Além dos fatores citados anteriormente, a opinião do observador acerca da qualidade é também afetada por outros fatores tais como o meio visual, o estado mental do observador e o nível de iteração do observador com o cenário visual. Por exemplo, um observador que se concentre em observar parte de uma cena de vídeo terá uma grande diferença para o conceito de "bom" em relação ao usuário que assiste passivamente ao filme. Experiências mostram que, se o ambiente visual é confortável, a opinião de um observador sobre a qualidade visual de uma cena de vídeo acaba sendo mais alta do que se o ambiente é desconfortável.

Outra importante influência na percepção da qualidade inclui atenção visual (um observador perceber uma cena fixando sua atenção em uma seqüência de pontos na imagem preferencialmente do que vendo todos os pontos simultaneamente) e ao chamado "efeito recente" (nossa opinião de uma seqüência visual é mais fortemente influenciada pela observação das cenas mais recentes do que das cenas mais antigas). Em função de todos estes fatores, medir a qualidade visual de forma precisa em termos quantitativos e qualitativos torna-se bastante difícil.

Diversos procedimentos de testes para avaliar a qualidade subjetiva são definidos na Recomendação ITU-R BT.500-11 [5]. Um procedimento comumente usado é o método Double Stimulus Continuous Quality Scale (DSCQS) [1].

### 2.2.2
### Qualidade Objetiva

Em função da relativa complexidade e do custo necessários para medir a qualidade subjetiva, procurou-se uma maneira de realizar esta medida através de uma expressão matemática. A medida mais usada é o PSNR (Peak Signal to Noise Ratio).

PSNR é medido em uma escala logarítmica e depende do Erro Médio Quadrático (MSE) entre as imagens ou quadros de vídeo originais e com distúrbios e do fator $(2^n - 1)^2$ (quadrado do mais alto valor de sinal possível na imagem, onde n é o número de bits por amostra da imagem). A medida de PSNR é dado pela eq. (2.1).

$$PSNR_{dB} = 10\log_{10}\frac{(2^n - 1)^2}{MSE} \qquad (2.1)$$

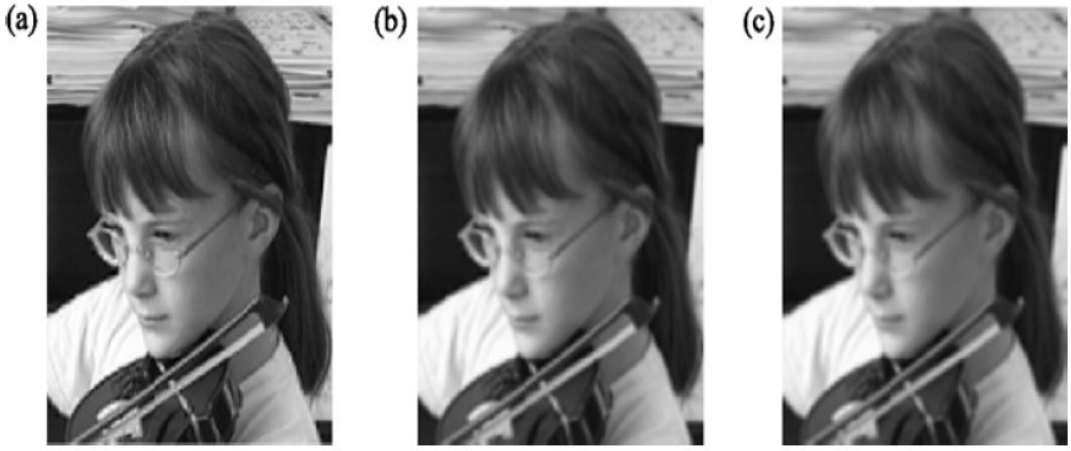


Figura 2- Exemplos de PSNR: (a) original; (b) 30.6 dB; (c) 28.3 dB
Fonte: Ref [1]

Figura 3-Imagem com fundo borrado (PSNR = 27.7 dB)
Fonte: Ref [1]



A Figura 2 mostra três imagens: a primeira imagem (a) é a original, (b) e (c) são versões degradadas da imagem original. Imagem (b) tem uma medida de PSNR de 30,6 dB enquanto que a imagem (c) tem um PSNR de 28,3 dB, mostrando uma medida da qualidade menor da imagem.

A medida de PSNR apresenta algumas limitações. A primeira limitação é que PSNR exige uma imagem original sem distorções para comparação no cálculo de MSE. No entanto, esta imagem pode não estar disponível. A segunda limitação é que PSNR não necessariamente se correlaciona com a qualidade subjetiva verdadeira. Por exemplo, a Figura 3 mostra uma versão distorcida da imagem original da Figura 2 no qual somente o fundo da imagem foi borrado. Essa imagem apresenta um PSNR de 27.7 dB em relação ao original. A maioria dos observadores classificaria esta imagem como significativamente melhor do que a imagem (c) na Figura 2, pois a face (parte da imagem que não inclui o fundo) é mais clara, contradizendo o valor de PSNR. Isto acontece pois um observador humano é particularmente sensível à distorção nesta área.

## 2.3
## Codificação de vídeo

Compressão ou codificação de vídeo é o processo de compactar uma seqüência de vídeo digital em uma quantidade menor de bits. O vídeo digital não codificado geralmente necessita de uma enorme taxa de bits (aproximadamente

216 Mbits/segundo). A taxa de bits representa quantos bits são necessários para reproduzir um tempo de um segundo de uma seqüência de vídeo. Portanto, a compressão faz-se necessária para reduzir esta taxa de bits e possibilitar assim o armazenamento e a transmissão de vídeo digital.

Compressão apresenta um compactador (codificador) e um descompactador (decodificador). O codificador converte a fonte de dados em uma forma com um número de bits menor, antes de armazenar ou transmitir. Já o decodificador realiza o processo inverso, convertendo os dados comprimidos em uma representação dos dados do vídeo original. O par codificador/decodificador, em função da tradução para a língua inglesa ser enCOder/DECoder, é freqüentemente denominado CODEC.

Muitos tipos de dados contêm redundâncias estatísticas e podem ser comprimidos através de uma compressão sem perdas, de tal forma que os dados reconstruídos na saída do decodificador são uma cópia idêntica dos dados da fonte. No entanto, este tipo de codificação apresenta uma taxa de compressão moderada. Para realizar uma maior compressão, com perda da qualidade visual, é preciso uma codificação com perdas. Nesse tipo de compressão, os dados descompactados não são idênticos aos dados da fonte. Sistemas de compressão de vídeo com perdas baseiam-se na eliminação de redundância subjetiva, ou seja, retirada de elementos da imagem ou seqüência de vídeo sem efeitos significativos na percepção da qualidade visual do observador.

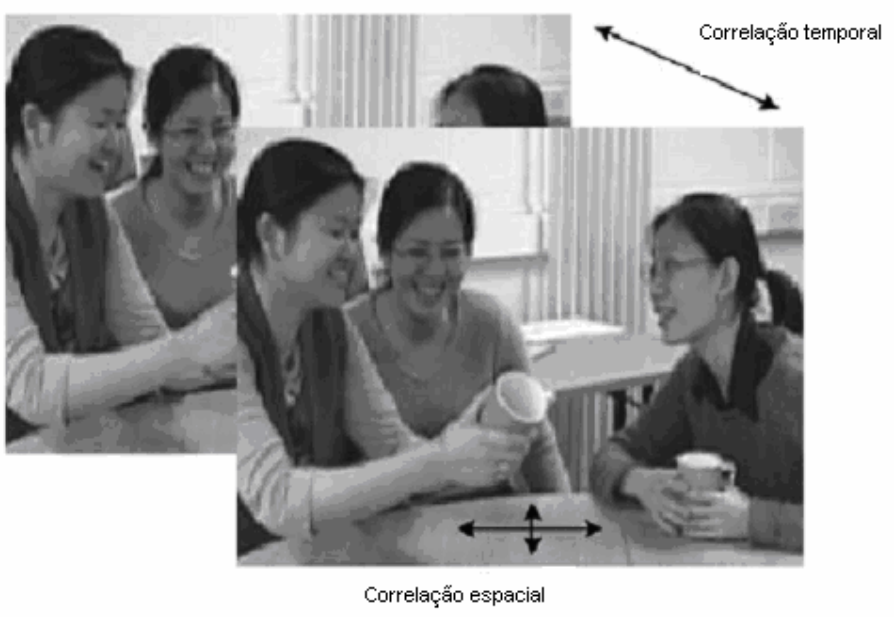


Figura 4- Correlação Espacial e temporal numa seqüência de vídeo
Fonte: Ref [1]

A maioria dos métodos de codificação de vídeo exploram tanto a redundância temporal quanto a espacial. No domínio temporal, quadros de vídeo que foram capturados em tempos próximos geralmente apresentam uma alta correlação. No domínio espacial, pixels que estão próximos uns dos outros geralmente também apresentam uma alta correlação, isto é, os valores das amostras vizinhas são geralmente muito similares, como mostra a Figura 4.

## 2.4
## Codificação e Decodificação de vídeo

Um codificador de vídeo é formado por três principais unidades funcionais: um modelo temporal (que realiza os processos de compensação e estimação de movimento), um modelo espacial (que realiza os processos de transformação e quantização) e um codificador entrópico. Estes modelos são mostrados na Figura 5.

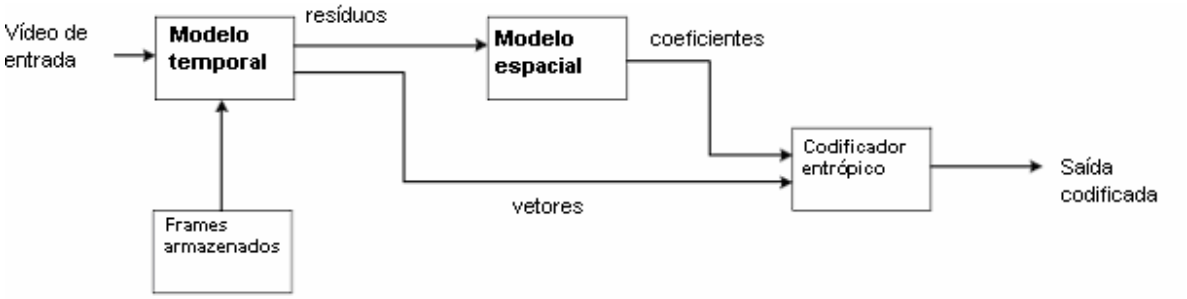


Figura 5- Diagrama de blocos de um vídeo codificado.

Fonte: Ref [1]

Uma seqüência de vídeo não comprimida é a entrada para o modelo temporal. Este modelo diminui a redundância temporal explorando as similaridades entre *quadros* de vídeos vizinhos, geralmente pela predição do quadro de vídeo atual. No padrão H.264 a predição é realizada a partir de um ou mais *quadros* anteriores ou futuros e seguida por uma compensação das diferenças entre os quadros. A subtração entre o quadro de referência e o atual gera um quadro residual que, juntamente com um conjunto vetores de movimento e de parâmetros, representam a saída do modelo temporal .

O quadro residual é a entrada para o modelo espacial que diminui a redundância espacial utilizando as similaridades entre pixels vizinhos neste quadro. No padrão H.264 isso é realizado por uma transformada (variante da Transformada Cosseno Discreta [1]) das amostras residuais e uma quantização dos resultados.

A transformada é realizada para cada um dos blocos separadamente e tem como saída uma matriz de coeficientes de mesma dimensão do bloco de entrada. O elemento (0,0) é o valor médio do bloco. Os outros elementos informam quanta potência espectral está presente em cada freqüência espacial. Teoricamente, uma transformação deste tipo não tem perdas. Na prática, o uso de número de ponto flutuante e funções transcendentais sempre acarreta alguns erros de arredondamento, que resultam em uma pequena perda de informação. Normalmente, estes elementos desaparecem com rapidez com o aumento da distancia em relação à origem (0,0), como sugere a Figura 6.

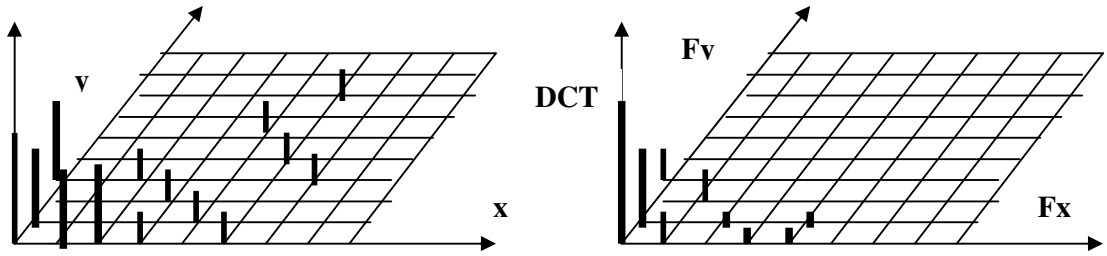


Figura 6 – (a) Bloco 8x8 da matriz de luminância Y (b) coeficientes transformados.

Fonte: Ref [6]

Na quantização, os coeficientes menos importantes são descartados. Essa operação (com perdas) é realizada através da divisão de cada coeficiente da matriz por um peso obtido em uma tabela. Se todos os pesos forem iguais a 1, a transformação nada fará. No entanto, se os pesos aumentarem muito a partir da origem, freqüência espaciais mais altas serão abandonadas rapidamente. Um exemplo dessa etapa é mostrado na Figura 7. Nela, vemos a matriz de coeficientes transformados de um bloco 8x8 (a), a tabela de quantização (b) e o resultado obtido ao se dividir cada elemento de (a) por (b). Os valores na tabela de quantização não fazem parte do padrão H.264 e cada aplicação deve fornecer seus próprios valores, garantindo assim o controle da negociação entre perdas e compactação. Um conjunto de coeficientes transformados e quantizados é a saída do modelo espacial.



| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 150 | 80 | 40 | 14 | 4 | 2 | 1 | 0 |
| 92 | 75 | 36 | 10 | 6 | 1 | 0 | 0 |
| 52 | 38 | 26 | 8 | 7 | 4 | 0 | 0 |
| 12 | 8 | 8 | 4 | 2 | 1 | 0 | 0 |
| 4 | 3 | 2 | 0 | 0 | 0 | 0 | 0 |
| 2 | 2 | 1 | 1 | 0 | 0 | 0 | 0 |
| 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |

Coeficientes Transformados

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 1 | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| 1 | 1 | 2 | 4 | 8 | 16 | 32 | 64 |
| 2 | 2 | 2 | 4 | 8 | 16 | 32 | 64 |
| 4 | 4 | 4 | 4 | 8 | 16 | 32 | 64 |
| 8 | 8 | 8 | 8 | 8 | 16 | 32 | 64 |
| 16 | 16 | 16 | 16 | 16 | 16 | 32 | 64 |
| 32 | 32 | 32 | 32 | 32 | 32 | 32 | 64 |
| 64 | 64 | 64 | 64 | 64 | 64 | 64 | 64 |

Tabela de Quantização

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 150 | 80 | 20 | 4 | 1 | 0 | 0 | 0 |
| 92 | 75 | 18 | 3 | 1 | 0 | 0 | 0 |
| 26 | 19 | 13 | 2 | 1 | 0 | 0 | 0 |
| 3 | 2 | 2 | 1 | 0 | 0 | 0 | 0 |
| 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 |

Coeficientes quantizados

Figura 7 – Cálculo dos coeficientes transformados e quantizados.

Fonte: Ref [6]

Os parâmetros do modelo temporal (tipicamente vetores de movimentos) e do modelo espacial (coeficientes) são comprimidos por um codificador entrópico. Isso retira redundâncias estatísticas nos dados (por exemplo, utilização de códigos binários mais curtos para representação de vetores e coeficientes que ocorrem com maior freqüência) e produz um bitstream comprimido que pode ser transmitido e/ou armazenado. Uma seqüência comprimida consiste de vetores de movimento codificados, coeficientes residuais codificados e cabeçalho de informações.

O decodificador reconstrói um quadro de vídeo a partir do bitstream comprimido. Os coeficientes e vetores de movimento são decodificados por um decodificador entrópico. O modelo espacial decodifica gerando um quadro residual. O decodificador usa os parâmetros do vetor de movimento, junto com um ou mais quadros decodificados previamente, para realizar uma predição do quadro atual, e assim o quadro é reconstruído pela adição do quadro residual ao seu preditor.

No próximo capítulo, será abordado com mais detalhes o modelo temporal (que engloba as fases de estimação e compensação de movimento), pois relaciona-se diretamente com este trabalho. Detalhes sobre as demais fases da codificação podem ser obtidos em [1].

## 2.5
## Energia de um bloco residual

O resíduo ou bloco residual é a diferença entre o bloco atual e a área de referência. A compensação de movimento tem o objetivo de minimizar a energia dos coeficientes residuais transformados. A energia em um bloco transformado depende da energia em um bloco residual (nome dado ao bloco antes da transformação). Estimação de movimento, conseqüentemente, tem por objetivo encontrar um casamento para o bloco atual que minimize a energia no resíduo com movimento compensado. Isso envolve a avaliação da energia residual em um número de diferentes possibilidades. A escolha da medida desta energia afeta a complexidade computacional e a precisão do processo de estimação de movimento. Eq. (2.1), eq. (2.2) e eq. (2.3) são mostradas abaixo e descrevem três medidas da energia: MSE (Erro Médio Quadrático ou Mean Square Error), MAE (Erro Médio Absoluto ou Mean Absolute Error) e SAE (Soma dos Erros

Absolutos ou Sum of Absolute Errors). O SAE é algumas vezes chamado de SAD (Soma das Diferenças Absolutas ou Sum of Absolute Differences). O tamanho do bloco de compensação de movimento é de NxN pixel; $C_{ij}$ e $R_{ij}$ são valores de amostras da área atual e de referência, respectivamente.

$$MSE = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left(C_{ij} - R_{ij}\right)^2 \qquad (2.1)$$

$$MAE = \frac{1}{N^2}\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left|C_{ij} - R_{ij}\right| \qquad (2.2)$$

$$SAE = \sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\left|C_{ij} - R_{ij}\right| \qquad (2.3)$$

SAE (algumas vezes chamado de SAD, Soma das Diferenças Absolutas ou Sum of Absolute Differences) é provavelmente a forma de medida da energia residual mais amplamente usada por razões de simplicidade computacional. O software de referência do padrão H.264[7] usa SA(T)D. O SAD é usado quando se calcula a estimação de movimento de pixel inteiro, que será apresentada no item 5.2.1, enquanto SATD é usado para estimação de movimento de subpixel, que será apresentado no item 5.2.2. SA(T)D refere-se tanto ao SAD quanto ao SATD. SATD representa a soma das diferenças absolutas dos dados residuais transformados como a forma de medida da energia. Realizar a transformada no resíduo de cada localização de pesquisa aumenta o esforço computacional, mas melhora a precisão da medida de energia. Uma simples transformada é usada de tal forma que o custo computacional extra não é excessivo. Abaixo apresentaremos um exemplo, extraído de [1], que mostra a escolha do melhor vetor de movimento.

***Exemplo:***

Avaliando MSE para todas as possíveis posições na região de pesquisa da Figura 8 tem-se um mapa de MSE (Figura 9). Esse gráfico apresenta um mínimo em (+2,0) o que significa que o melhor casamento é obtido pela seleção de uma região de tamanho 32x32 amostras (pixel) em uma posição de 2 pixels à direita da posição do bloco no quadro atual. MAE e SAE são mais fáceis de calcular do que MSE. Seus "mapas" são mostrados nas Figuras 10 e 11. Apesar do gradiente

mostrado no mapa ser diferente para o caso de MSE, ambas as medidas apresentam um ponto de mínimo localizado em (+2,0).

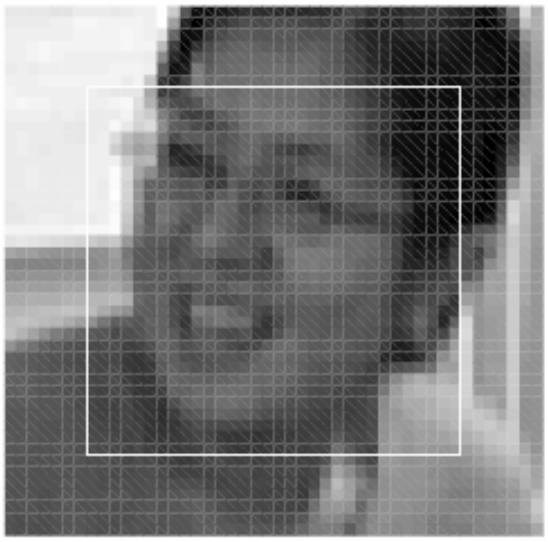

Figura 8- Região de pesquisa no quadro anterior (referência)

Fonte: Ref [1]

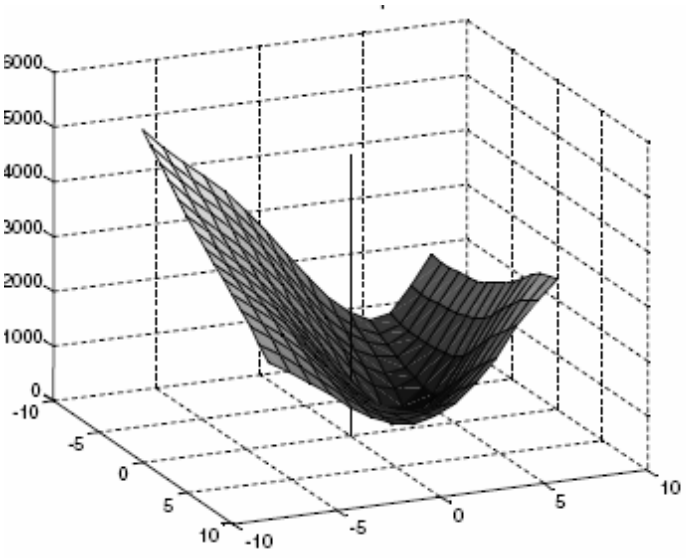


Figura 9- Mapa MSE

Fonte: Ref [1]

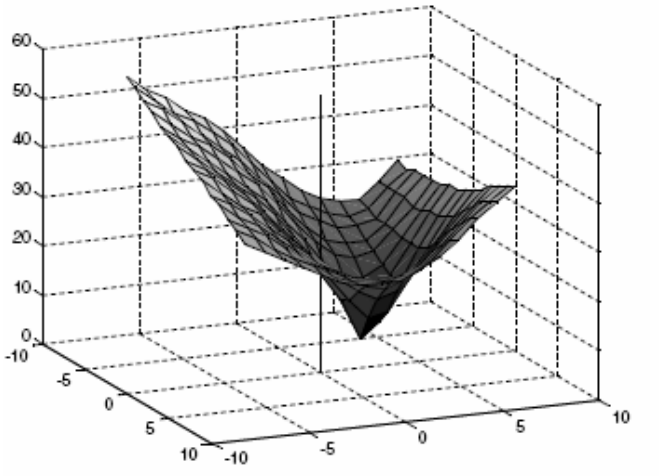


Figura 10- Mapa MAE
Fonte: Ref [1]



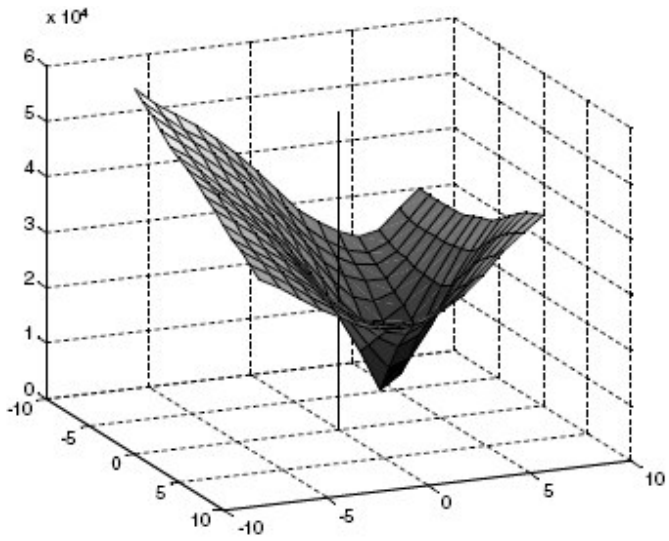


Figura 11- Mapa SAE (SAD)
Fonte: Ref [1]

Os resultados do exemplo acima indicam que a melhor escolha do vetor de movimento é (+2,0). O mínimo do mapa MSE ou SAE indicam o ponto que produz a menor energia residual e é provável que apresente a menor energia dos coeficientes transformados e quantizados. O vetor de movimento deve ser transmitido para o decodificador. Contudo, como vetores de grande magnitude são codificados usando mais bits do que vetores de pequena magnitude deve-se

procurar vetores que sejam próximos de (0,0). Isso pode ser realizado simplesmente pela subtração de uma constante a partir de MSE ou SAE que leve para posição (0,0).

Uma técnica mais sofisticada é tratar a escolha do vetor de uma forma otimizada. O codificador do modelo de referência H.264 adiciona um parâmetro para cada elemento codificado (MVD, modo de predição, etc) antes de escolher o menor custo total para a predição de movimento. Assim, não seria necessário calcular SAE (ou MAE ou MSE) completamente em cada localização. Um atalho usado é terminar o cálculo no momento em que se verifica que o SAE excede o mínimo SAE obtido anteriormente.

Por exemplo, após calcular cada soma da eq. (2.3), $\left(\sum_{J=0}^{N-1}\left|C_{ij}-R_{ij}\right|\right)$, o codificador compara o total SAE com o mínimo obtido anteriormente. Se o total exceder o mínimo anterior, o cálculo é terminado.

## 2.6
## Conceito de Macrobloco

O macrobloco, correspondendo a uma região de um quadro de tamanho 16x16 pixel, é a unidade básica para estimação envolvendo compensação de movimento em um número grande de padrões de codificação visuais incluindo MPEG-1, MPEG-2, MPEG-4 Visual, H.261, H.263 e H.264 [1].

Para fonte de vídeo no formato 4:2:0, um macrobloco é organizado como mostra a Figura 12. Uma região de tamanho 16x16 pixels do quadro de origem é representada por 256 amostras de luminância (arranjadas em 4 blocos de amostras 8x8), 64 amostras de crominância azul (um bloco 8x8) e 64 amostras de crominância vermelha (um bloco 8x8), dando um total de seis blocos 8x8. Um codificador MPEG Visual ou H.264 processa cada quadro de vídeo em unidades de um macrobloco.

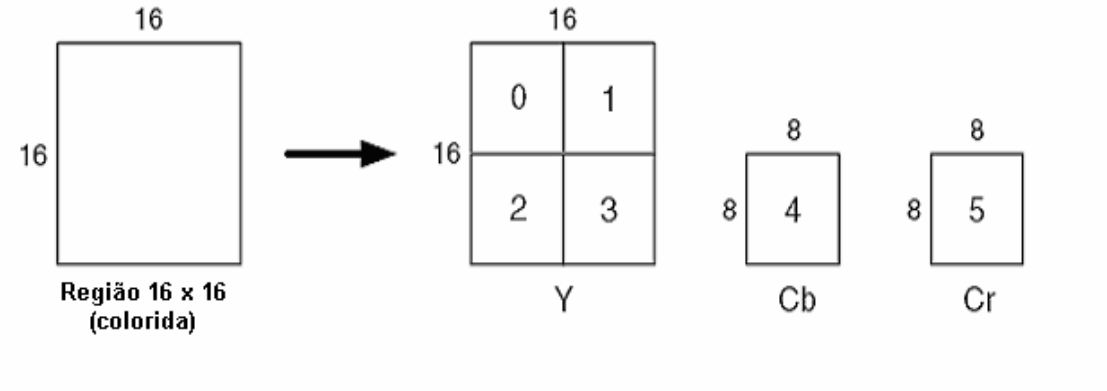


Figura 12- Macrobloco (4:2:0)
Fonte: Ref [1]



## 2.7
## Resumo e conclusão do capítulo

Neste capítulo foram apresentados alguns conceitos básicos sobre formatos e qualidade de vídeo, codificação de vídeo e medidas de energia de um bloco de pixels.

Quanto aos formatos de vídeo, foram abordadas a resolução de luminância de quadro e a aplicação dos formatos 4CIF, CIF, QCIF e SQCIF. Particularmente o formato QCIF pode ser usado em videoconferência e aplicações multimídia móveis, enquanto o formato CIF pode ser usado em videoconferência.

A medida da qualidade de vídeo pode ser realizada de forma subjetiva ou objetiva. A qualidade subjetiva refere-se a opinião de um observador ao assistir um vídeo, sendo influenciada por muitos fatores. A qualidade objetiva pode ser medida através de uma expressão matemática, sendo a mais usada o PSNR (Peak Signal to Noise Ratio) que é dada pela eq. (2.1).

Compressão ou codificação de vídeo é o processo de compactar uma seqüência de vídeo digital em uma quantidade menor de bits, em função da existência de redundâncias estatísticas. Esta codificação pode ser com perdas ou sem perdas.

O resíduo ou bloco residual é a diferença entre o bloco atual e a área de referência. Estimação de movimento tem por objetivo encontrar um casamento para o bloco atual que minimize a energia no resíduo. Isso envolve a avaliação da

energia residual em um número de diferentes possibilidades. A energia de um bloco residual é comumente medida pela eq. (2.3) e é denominada SAE (Soma dos Erros Absolutos ou Sum of Absolute Errors) ou SAD (Soma das Diferenças Absolutas ou Sum of Absolute Differences).

O macrobloco, correspondendo a uma região de um quadro de tamanho 16x16 pixel, é a unidade básica para estimação envolvendo compensação de movimento em um número grande de padrões de codificação.

Os conceitos apresentados neste capítulo foram particularmente importantes para o entendimento deste trabalho.