Implementação modificada do codificador de vídeo H.264 com a transformada SHICT

Kleber Teraoka e Max H. M. Costa

***Resumo*— Este artigo apresenta uma implementação modificada do codificador de vídeo H.264 utilizando uma variação da transformada discreta do co-seno (DCT) denominada *Shifted Integer Cosine Transform - SHICT*. Esta transformada possui baixa complexidade e é implementada apenas com operações de adição e de deslocamentos binários. Assim, a transformada é implementada sem operações de multiplicação ou divisão. O trabalho mostra o desempenho da implementação em termos da complexidade computacional e das curvas de taxa x distorção resultantes. Concluindo, discutem-se possibilidade de aplicações.**

***Palavras-Chave*— H.264, DCT, ICT, SHICT, compressão de vídeo.**

***Abstract*— This article presents a modified H.264 codec implementation based on a variation of the discrete cosine transform (DCT) named Shifted Integer Cosine Transform (SHICT). The proposed low-complexity transform is implemented using only addition and binary shift operations, avoiding multiplications and divisions. The resulting performance is presented in terms of computational complexity and rate-distortion curves. In conclusion possible applications are discussed.**

***Keywords*— H.264, DCT, ICT, SHICT, video compression.**

I. Introdução

O padrão de compressão de vídeo H.264 [1] representa o estado da arte na codificação de vídeo digital. Este padrão apresenta um avanço em relação a padrões anteriores tais como o MPEG-2 e MPEG-4, com redução significativa da taxa de bits necessária para a representação de vídeo digital. Esta redução se deve a um conjunto de melhorias nas ferramentas de codificação, ao custo de um significativo aumento da complexidade computacional do sistema. Desta forma, é desejável que implementações do codificador H.264 sejam realizadas através de algoritmos simplificados e que mantenham um bom desempenho.

Neste artigo apresentamos a implementação de uma variação do codificador H.264 utilizando a transformada denominada Shifted Integer Cosine Transform (SHICT). A etapa de quantização/normalização inversa é realizada através de simples deslocamentos na representação binária dos coeficientes, enquanto que a etapa direta é realizada com mais precisão para compensar os erros introduzidos. Tal configuração simplifica os decodificadores e contribui para a redução do custo dos receptores de vídeo digital. Como a aproximação no cômputo da etapa inversa da transformada é compensada por cálculos de mais alta precisão na etapa direta, esta implementação pode ser benéfica em aplicações que envolvam muito mais receptores (decodificadores) que transmissores (codificadores), como em radiodifusão.

Departamento de Comunicações, Faculdade de Engenharia Elétrica e de Computação, Universidade Estadual de Campinas, Campinas, SP. klebert@decom.fee.unicamp.br, max@decom.fee.unicamp.br.

II. Shifted Integer Cosine Transform

A transformada SHICT se caracteriza por ser uma transformada inteira, cujos fatores de quantização/normalização são aproximados por potências de dois. Esta aproximação é intercambiável, ou seja, pode ser realizada no codificador (etapa direta) ou no decodificador (etapa inversa). A vantagem associada é a redução da complexidade computacional de uma das etapas do codec, que pode ser implementada sem operações de multiplicação ou de divisão, usando apenas operações de adição e de deslocamentos binários.

Os erros provenientes da aproximação podem ser parcialmente compensados na etapa complementar do codec (direta ou inversa), podendo ser realizada com mais precisão, utilizando ponto flutuante ou maior precisão inteira.

A transformada utilizada no padrão H.264 é a transformada inteira do coseno (ICT) de tamanho 4x4, a qual é uma variação da transformada discreta do co-seno (DCT).

Uma forma de se obter a matriz ICT [2] é através da multiplicação de uma matriz DCT $\mathbf{A}$ por um número α, seguido de arredondamento:

$$\mathbf{C} = \text{round}[\alpha \mathbf{A}], \tag{1}$$

onde round[.] é a função de arredondamento.

A matriz utilizada no padrão H.264 é obtida fazendo $\alpha = 2,5$ e a matriz $\mathbf{C}$ é dada por

$$\mathbf{C} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 2 & 1 & -1 & -2 \\ 1 & -1 & -1 & 1 \\ 1 & -2 & 2 & -1 \end{pmatrix}. \tag{2}$$

Nesta matriz, a soma dos valores absolutos em qualquer linha é igual a 6, de forma que o ganho de faixa dinâmica para uma transformada bidimensional é $log_2 6^2 = 5,17$, ou seja, o armazenamento dos coeficientes de freqüência da matriz $\mathbf{Y} = \mathbf{CX}$ requer 6 bits a mais do que os da matriz de entrada $\mathbf{X}$. Para uma imagem com 256 níveis de resolução (8 bits/*pixel*), o resíduo (diferença calculada entre os valores de pixel do quadro original e do predito) necessita de um bit adicional para indicação do sinal. Assim, são necessários $6 + 9 = 15$ bits e a ICT pode ser calculada usando aritmética de 16 bits.

Assim como a DCT, a ICT é implementada através de uma transformação linear. Os coeficientes freqüenciais da matriz transformada $\mathbf{Y}$ são obtidos aplicando-se a transformada ICT representada por $\mathbf{C}$ às colunas e às linhas da imagem $\mathbf{X}$, ou

seja, $\mathbf{Y} = \mathbf{CXC^T}$. Portanto, a implementação da transformada requer $N^2(N-1)$ adições e N^3 multiplicações para cada bloco NxN da imagem a ser transformada.

No entanto, é possível implementar esta transformada de modo a eliminar as operações de multiplicações [3] em uma das etapas da transformada.

Sendo a matriz ICT $\mathbf{C}$ ortogonal, temos que $\mathbf{CC}^T = \mathbf{D}$, onde $\mathbf{D}$ é uma matriz diagonal. Seja $\mathbf{\Delta}$ a inversa de $\mathbf{D}$. Podemos fatorar a matriz identidade $\mathbf{I}$ da forma $\mathbf{I} = \sqrt{\mathbf{\Delta}}\mathbf{C}\mathbf{C}^T\sqrt{\mathbf{\Delta}}$. A matriz $\mathbf{M} = \sqrt{\mathbf{\Delta}}\mathbf{C}$ é uma matriz ortonormal, com $\mathbf{M}^T = \mathbf{C}^T\sqrt{\mathbf{\Delta}} = \mathbf{M}^{-1}$ e representa a matriz ICT normalizada.

Notando que

$$\mathbf{C}^{-1} = \mathbf{C}^T\mathbf{\Delta}, \tag{3}$$

podemos recuperar $\mathbf{X}$ a partir de $\mathbf{Y}$, pois

$$\mathbf{X} = \mathbf{C}^{-1}\mathbf{Y} = \mathbf{C}^{-1}\mathbf{C}\mathbf{X} = (\mathbf{C}^T\sqrt{\mathbf{\Delta}})(\sqrt{\mathbf{\Delta}}\mathbf{C})\mathbf{X}. \tag{4}$$

A ICT bidimensional é separável, ou seja, pode ser calculada aplicando a ICT unidimensional nas colunas da imagem, seguido das linhas da matriz do resultado intermediário. Na forma matricial, isto é equivalente a pré-multiplicar o bloco de pixels pela matriz ICT direta e pós-multiplicá-lo pela matriz ICT transposta. Assim,

$$\mathbf{Y} = \sqrt{\mathbf{\Delta}}\mathbf{C}\mathbf{X}\mathbf{C}^T\sqrt{\mathbf{\Delta}} \tag{5}$$

e

$$\mathbf{X} = \mathbf{C}^T\sqrt{\mathbf{\Delta}}\mathbf{Y}\sqrt{\mathbf{\Delta}}\mathbf{C}. \tag{6}$$

Uma simplificação adicional pode ser feita notando que $\mathbf{CXC}^T$ e $\mathbf{Y}$ são pré-multiplicados e pós-multiplicados pela matriz diagonal $\sqrt{\mathbf{\Delta}}$.

Dado duas matrizes diagonais, $\mathbf{D_1}$ e $\mathbf{D_2}$ e uma matriz qualquer $\mathbf{B}$, o produto $\mathbf{D_1BD_2}$ pode ser calculado de forma mais eficiente multiplicando-se termo a termo todos os elementos de $\mathbf{B}$ pelos elementos do produto $\mathbf{D_1}\mathbf{1}\mathbf{D_2}$, onde $\mathbf{1}$ é uma matriz do mesmo tamanho de B, composta exclusivamente por coeficientes iguais a "1". Este procedimento reduz o número de multiplicações pela metade. Assim, temos

$$\mathbf{Y} = \mathbf{CXC}^T\#\mathbf{N} \tag{7}$$

e

$$\mathbf{X} = \mathbf{C}^T(\mathbf{Y}\#\mathbf{N})\mathbf{C}, \tag{8}$$

onde o símbolo "#" denota o produto termo a termo e $\mathbf{N}$ é a matriz de normalização definida como

$$\mathbf{N} = \sqrt{\mathbf{\Delta}}1\sqrt{\mathbf{\Delta}}. \tag{9}$$

Ou seja, a transfomada DCT que utiliza a matriz $\mathbf{A}$ de elementos não-inteiros

$$\mathbf{Y} = \mathbf{AXA}^T = \tag{10}$$

$$\begin{pmatrix} a & a & a & a \\ b & c & -c & -b \\ a & -a & -a & a \\ c & -b & b & -c \end{pmatrix} X \begin{pmatrix} a & a & a & a \\ b & c & -c & -b \\ a & -a & -a & a \\ c & -b & b & -c \end{pmatrix}, \tag{11}$$

onde $a = 0,5$, $b = 0,6533$ e $c = 0,2706$ pode ser calculada alternativamente utilizando a matriz de inteiros $\mathbf{C}$ da seguinte forma

$$\mathbf{Y} = \mathbf{CXC}^T\#\mathbf{N} = \tag{12}$$

$$\mathbf{CXC}^T\# \begin{pmatrix} a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \\ a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \end{pmatrix}. \tag{13}$$

Os coeficientes de $\mathbf{Y}$ são quantizados através da divisão termo a termo de seus elementos por uma matriz de quantização $\mathbf{H}$. Assim, temos a matriz de transformação quantizada $\mathbf{Y}^*$, expressa por

$$\mathbf{Y}^* = \mathbf{Y}(\#)\mathbf{H} = (\mathbf{CXC}^T)\#\mathbf{N}(\#)\mathbf{H}, \tag{14}$$

onde o símbolo (#) denota a operação de arredondamento para o inteiro mais próximo.

Podemos combinar a operação de normalização e quantização em uma única matriz $\mathbf{Q}$, de forma que

$$\mathbf{Y}^* = (\mathbf{CXC}^T)(\#)\mathbf{Q}, \tag{15}$$

onde $\mathbf{Y}^*$ representa a matriz dos coeficientes quantizados e $\mathbf{Q} = \mathbf{N}\#\mathbf{H}$.

A imagem reconstruída $\mathbf{X}^*$ pode ser obtida por

$$\mathbf{X}^* = \mathbf{C}^T(\mathbf{Y}^*\#\mathbf{Q}^*)\mathbf{C}, \tag{16}$$

onde $\mathbf{Q}^*$ e $\mathbf{Q}$ satisfazem

$$\mathbf{Q}^*\#\mathbf{Q} = \mathbf{\Delta}\mathbf{1}\mathbf{\Delta}. \tag{17}$$

A Eq.(17) indica uma relação entre codificador e decodificador, relacionando as matrizes de quantização/normalização direta e inversa. Deste modo, é possível direcionar a complexidade para uma das partes (alterando os valores de uma das matrizes), sem prejuízos significativos de precisão, desde que a Eq.(17) seja satisfeita.

A transformada SHICT consiste de uma variação da ICT, sendo que os elementos das matrizes de quantização/normalização utilizadas são aproximados pelas potências de dois mais próximas, privilegiando-se os elementos de menor magnitude em caso de proximidade idêntica (para uma quantização mais suave).

III. A TRANSFORMADA ICT NO CODIFICADOR H.264

Na transformada direta, um total de 52 passos de quantização (Q_{step}) são definidos e indexados por um parâmetro QP (Tabela I). O valor de Q_{step} dobra de tamanho a cada incremento de 6 em QP, sendo que o passo aumenta aproximadamente em 12,5% para cada incremento unitário em QP. Esta variação nos passos de quantização possibilita um controle mais preciso e flexível do compromisso existente entre a taxa de bits e a qualidade da reprodução.

TABELA I

PASSOS DE QUANTIZAÇÃO DO CODIFICADOR H.264.

QP	0	1	2	3	...	51
Q_{step}	0,625	0,6875	0,8125	0,875	...	224

O cálculo da transformada direta é implementado a partir da Eq.(12). A etapa de quantização é realizada em conjunto com a normalização segundo a Eq.(15) com

$$\mathbf{Q} = \mathbf{N}\#\mathbf{H} = \frac{1}{Q_{step}} \begin{pmatrix} a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \\ a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \end{pmatrix}. \tag{18}$$

A fim de se evitar instruções de divisão, o *software* de referência implementa a quantização/normalização utilizando uma matriz de fatores multiplicativos $\mathbf{MF}$, onde

$$\mathbf{Q} = \mathbf{N}\#\mathbf{H} = \frac{1}{2^{qbits}}\mathbf{MF}, \tag{19}$$

ou seja,

$$\mathbf{MF} = \frac{2^{qbits}}{Q_{step}} \begin{pmatrix} a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \\ a^2 & ab/2 & a^2 & ab/2 \\ ab/2 & b^2/4 & ab/2 & b^2/4 \end{pmatrix}, \tag{20}$$

com $qbits = 15 + \lfloor QP/6 \rfloor$

A matriz $\mathbf{MF}$ depende somente do índice QP desejado, e, portanto, varia de acordo com o passo de quantização. No entanto, para $QP > 5$, os elementos de $\mathbf{MF}$ se repetem ciclicamente, pois o divisor $Qstep$ aumenta por um fator de 2 a cada incremento de 6 em QP, assim como 2^{qbits}.

Dessa forma, a matriz $\mathbf{MF}$ é fixa e não necessita ser recalculada para todo valor de QP (Tabela II). Dependendo da implementação, $\mathbf{MF}$ pode ser armazenada em memória não volátil.

TABELA II

MATRIZ DE FATORES MULTIPLICATIVOS $\mathbf{MF}$.

QP	Pos. (0,0),(2,0),(2,2),(0,2)	Pos. (1,1),(1,3),(3,1),(3,3)	Outras
0	13107	5243	8066
1	11916	4660	7490
2	10082	4194	6554
3	9362	3647	5825
4	8192	3355	5243
5	7282	2893	4559

O processo de decodificação visa recuperar a imagem original a partir dos coeficientes freqüenciais quantizados e codificados. Após o processo de decodificação de entropia, os coeficientes são recuperados a partir da equação

$$\mathbf{Y}^{'} = \mathbf{Y}^{*}\#\mathbf{Q}^{*} = 64Q_{step}\mathbf{Y}^{*}\#\mathbf{N_i}. \tag{21}$$

A equação acima é multiplicada por um fator constante igual a 64 para evitar erros de arredondamento.

Por fim, a imagem original pode ser reconstruída a partir de

$$\mathbf{X}^{'} = \text{round}\left[\frac{1}{64}\mathbf{C}^{T}\mathbf{Y}^{'}\mathbf{C}\right]. \tag{22}$$

A implementação de referência não especifica os valores de Q_{step} e $\mathbf{N_i}$ diretamente. Alternativamente, pode-se utilizar a matriz $\mathbf{V}$, definida para $0 \leq QP \leq 5$, sendo que

$$\mathbf{V} = 64Q_{step}\mathbf{N_i}. \tag{23}$$

Portanto,

$$\mathbf{Y}^{'} = 2^{\lfloor QP/6 \rfloor}\mathbf{Y}^{*}\#\mathbf{V}. \tag{24}$$

TABELA III

MATRIZ DE FATORES MULTIPLICATIVOS $\mathbf{V}$.

QP	Pos. (0,0),(2,0),(2,2),(0,2)	Pos. (1,1),(1,3),(3,1),(3,3)	Outras
0	10	16	13
1	11	18	14
2	13	20	16
3	14	23	18
4	16	25	20
5	18	29	23

Do mesmo modo que a matriz de fatores multiplicativos $\mathbf{MF}$, a matriz $\mathbf{V}$ do decodificador também é fixa (Tabela III) e o termo $2^{\lfloor QP/6 \rfloor}$ na Eq.(24) multiplica a saída por um fator de 2 a cada incremento de 6 em QP.

IV. A TRANSFORMADA SHICT NO CODIFICADOR H.264

A implementação da transformada SHICT no *software* de referência do padrão H.264 (JM v.9.3) consistiu em aproximar os elementos da matriz de fatores multiplicativos $\mathbf{V}$ por potências de dois, de forma que o decodificador pudesse realizar a etapa de quantização/normalização de forma mais simples e rápida. Deste modo, podemos obter a matriz de quantização modificada $\mathbf{V_{shifts}}$ dada pela Tabela IV.

TABELA IV

MATRIZ DE FATORES MULTIPLICATIVOS $\mathbf{V_{shifts}}$.

QP	Pos. (0,0),(2,0),(2,2),(0,2)	Pos. (1,1),(1,3),(3,1),(3,3)	Outras
0	$>> 3$	$>> 2$	$>> 2$
1	$>> 3$	$>> 2$	$>> 2$
2	$>> 2$	$>> 2$	$>> 2$
3	$>> 2$	$>> 2$	$>> 2$
4	$>> 2$	$>> 2$	$>> 1$
5	$>> 2$	$>> 2$	$>> 1$

Em contrapartida, o codificador deve implementar a etapa de quantização/normalização direta de forma que a Eq.(17) seja satisfeita. Assim, a matriz de fatores multiplicativos $\mathbf{MF}$ deve ser alterada, resultando na matriz $\mathbf{MF_{shifts}}$ mostrada na Tabela V. Para maior eficiência computacional, a implementação da multiplicação dos coeficientes por $\mathbf{MF_{shifts}}$ é realizada através de uma multiplicação inteira (pelo número resultante de $2^{15} * \mathbf{MF_{shifts}}(\mathbf{i}, \mathbf{j})$, seguida de deslocamento binário ($>> 15$),

TABELA V
MATRIZ DE FATORES MULTIPLICATIVOS $\mathbf{MF_{shifts}}$.

QP	Pos. (0,0),(2,0),(2,2),(0,2)	Pos. (1,1),(1,3),(3,1),(3,3)	Outras
0	0,5	0,2	0,16
1	0,5	0,2	0,16
2	0,25	0,2	0,16
3	0,25	0,2	0,16
4	0,25	0,2	0,08
5	0,25	0,2	0,08

O *software* de referência de implementação do padrão H.264 foi modificado de forma que os coeficientes de luminância foram transformados com a SHICT em substituição à tradicional ICT. Ensaios foram realizados com a seqüência *Foreman*, QCIF, no formato YUV 4:2:0 para o perfil de codificação *baseline*, utilizando um GOP do tipo IPPP.

V. RESULTADOS

Como era de se esperar, o codificador modificado apresentou desempenho inferior em relação à implementação de referência, conforme se pode observar na Fig. 1. No entanto, para baixas taxas o codificador com SHICT tem desempenho comparável.

Com relação aos diversos passos de quantização, o codificador mantém uma relação sinal/ruído praticamente equivalente à de referência, variando de forma aproximadamente linear (em dB). No entanto, para uma dada relação sinal-ruído, a taxa de bits requerida é maior, como visto na Fig. 1.

Com relação aos tempos de decodificação, podemos observar que, de modo geral, a implementação da SHICT mostrou-se mais rápida quando se considera apenas o cálculo da transformada. (Fig. 3). Já com relação ao tempo total de decodificação (Fig. 4), que inclui as operações de estimação e compensação de movimento, temos que para os primeiros 17 passos de quantização, a SHICT se mostrou como uma implementação atraente, visto que os tempos para decodificar a seqüência resultaram inferiores à implementação de referência. Para os demais passos, os tempos de implementação da SHICT e da ICT se mostraram comparáveis.

Os resultados mostram que a implementação do decodificador por intermédio da SHICT exige do codificador uma taxa de bits mais alta para a mesma qualidade de reprodução (mesma relação sinal/ruído). Este pode um preço alto para as vantagens de redução de complexidade computacional que se observam em reproduções de mais alta fidelidade (maior relação sinal/ruído). Nos operações de baixa taxa e baixa relação sinal/ruído, as implementações baseadas na ICT e na SHICT apresentaram resultados essencialmente equivalentes. Visto que a maior parte da carga no processamento de um sistema de codificação de vídeo é ocupado pela estimação/compensação de movimento, a redução de tempo que se obtém com o uso transformada proposta é relativamente pequena em relação ao tempo total para codificação de uma seqüência de vídeo. Portanto, no contexto das comparações realizadas, a implementação do decodificador com a SHICT não apresentou vantagens significativas.

Uma aplicação que pode ser beneficiada pela eficiência computacional da SHICT consiste de um *codec* de vídeo baseado em transformadas de blocos 3-D (com duas dimensões espaciais e uma dimensão temporal). Neste sistema as operações de estimação e compensação de movimento, que representam aproximadamente 70% do custo computacional dos *codecs* atuais, são abolidas. Portanto, o esforço computacional deste tipo de *codec* se concentra nas transformadas 3-D, e a simplificação de uma das etapas desta transformada (no decodificador ou no codificador) pode representar uma vantagem relevante nas aplicações de baixas taxas, compatíveis com receptores portáteis e de limitados recursos computacionais.

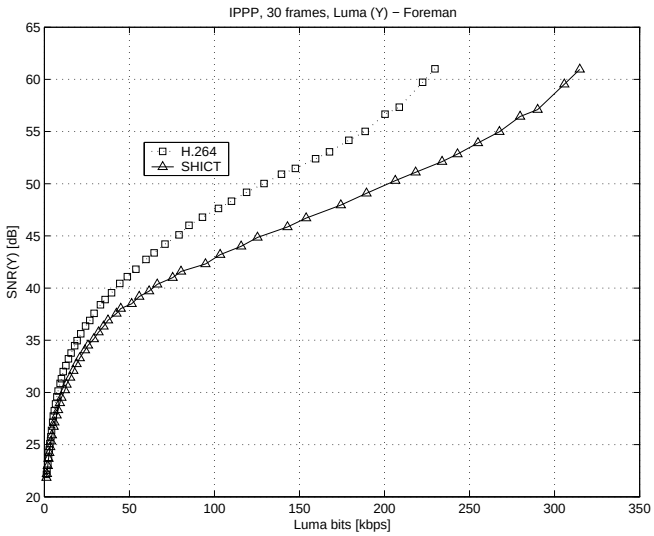

Fig. 1. Relação sinal/ruído (SNR) versus taxa de bits

AGRADECIMENTOS

Os autores agradecem ao grupo JVT pela disponibilização do *software* de referência utilizado para a codificação de seqüências de vídeo. Mais informações podem ser encontradas em *http://iphome.hhi.de/suehring/tml/index.htm*.

REFERÊNCIAS

[1] R. Schäfer, T. Wiegand and H. Schwarz, "The emerging H.264/AVC standard". *European Broadcasting Union Technical Review*, jan. 2003.

[2] H. Malvar, A. Hallapuro, M. Karczewicz e L. Kerofsky, "Low-complexity transform and quantization with 16-bit arithmetic for H.26L". *IEEE International Conference on Image Processing*, set. 2002.

[3] M. H. M. Costa e K. Tong, "A Simplified Integer Cosine Transform and Its Application in Image Compression". *The Telecommunications and Data Acquisition Progress Report 42-119, Jet Propulsion Laboratory, NASA, Pasadena, CA, EUA*, pp. 129-139, nov. 1994.

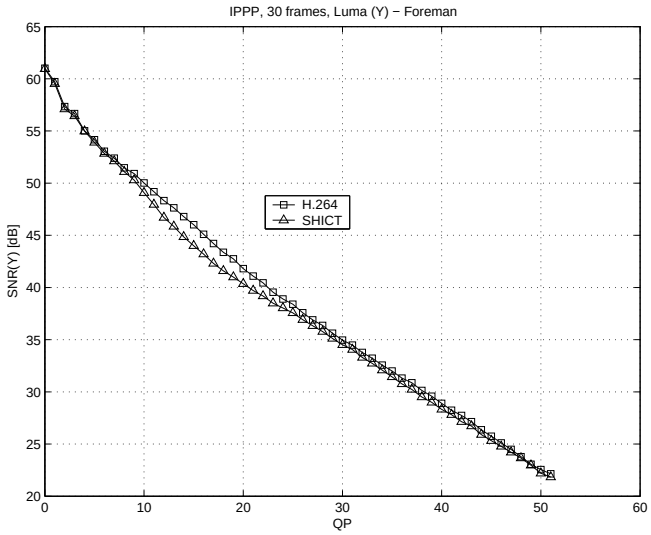

Fig. 2. Relação sinal/ruído (SNR) versus passos de quantização (QP)

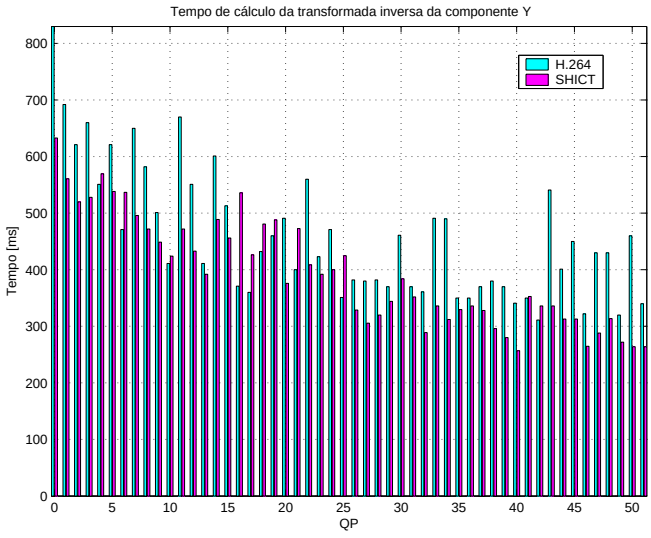

Fig. 3. Comparação entre os tempos de cálculo da transformada inversa padrão e proposta (H.264)

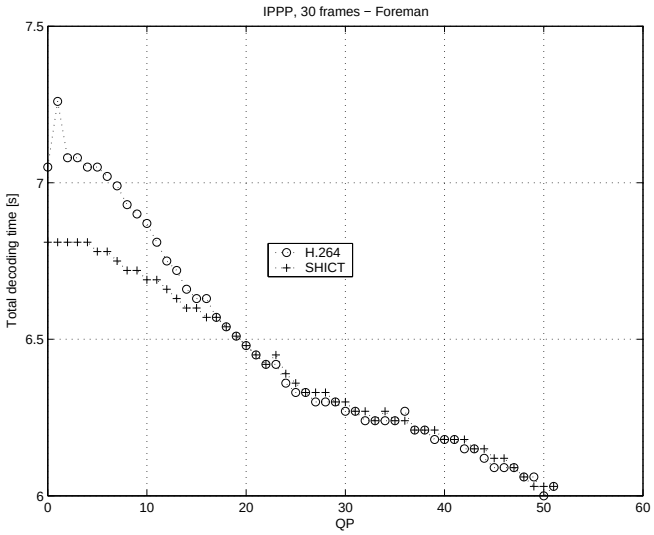

Fig. 4. Tempos de decodificação com estimação e compensação de movimento