14th European Signal Processing Conference (EUSIPCO 2006), Florence, Italy, September 4-8, 2006, copyright by EURASIP

A SPATIO-TEMPORAL COMPETING SCHEME FOR THE RATE-DISTORTION OPTIMIZED SELECTION AND CODING OF MOTION VECTORS

G. Laroche[1,2], *J. Jung*[1], *and B. Pesquet-Popescu*[2]

[1]France Telecom R&D
MAPS/MDG/SIA
38-40 rue du General Leclerc,
92794 Issy Les Moulineaux France
guillaume.laroche@francetelecom.com
joelb.jung@francetelecom.com

[2]ENST Paris
Signal and Image Proc. Department
46 rue Barrault,
75634 Paris, France
beatrice.pesquet@enst.fr

ABSTRACT

The recent H.264/MPEG4-AVC video coding standard has achieved a significant bitrate reduction compared to its predecessors. High performance texture coding tools and $\frac{1}{4}$-pel motion accuracy have however contributed to an increased proportion of bits dedicated to the motion information. The future ITU-T challenge, to provide a codec with 50% bitrate reduction compared to the current H.264, may result in even more accurate motion models. It is consequently of highest interest to reduce the motion information cost.

This paper proposes a competitive framework, with spatial and temporal predictors optimally selected by a rate-distortion criterion taking into account the cost of the motion vector and the predictor information. These new methods take benefit of temporal redundancies in the motion field, where the standard spatial median usually fails. Compared to an H.264/MPEG4-AVC standard codec, a systematic bitrate saving reaching up to 20% for complex sequences is reported.

1. INTRODUCTION

The recent ITU-T standard H.264 [1], also known as MPEG-4 AVC in ISO/IEC, achieves a significant compression gain compared to its predecessors H.263 and MPEG-4 part 2. This gain results from the improvement of existing tools and the inclusion of new ones. Efficient intra prediction, variable block sizes, $\frac{1}{4}$-pel motion estimation, in-loop deblocking filter, and context adaptive binary arithmetic coding (CABAC) are the most noticeable. In addition, a huge work has been accomplished on the reference software [2] that provides efficient non normative choices [3] based on rate-distortion schemes able to make optimal choices among the many new competing encoding modes. As a result of efficient texture coding tools, H.264 has allowed a reduction of the total bitrate, but the proportion of the motion information has increased. Indeed, at low bitrate, it can reach up to 40% of the total bitrate.

In the near future, the Video Coding Expert Group (VCEG/ITU-T SG16 Q6) will raise a new challenge to provide a 50% compression gain for an H.264 equivalent quality. The new standards may even increase this motion information, with more accurate motion models, and probably with a large bitrate reduction of luminance block residue. We have therefore focused our attention on the reduction of the motion information cost.

The cost reduction of the motion information has already been largely addressed in the literature. Methods based on lossy encoding [4] are not addressed here. Lossless methods are more widespread and some of them have already introduced temporal predictions. In [5], a temporal predictor is used to exploit temporal redundancies between motion vectors fields. This method yields good results in sequences with complex motion fields. However temporal prediction is not more efficient than spatial only prediction when a representative set of sequences is considered. In [6], temporal and spatial predictors are used. The choice between these two possible predictors is made depending on the value of the predictors, which does not ensure an optimal choice: a competitive framework is missing. In [7] a selection is made at the slice level between spatial or combined spatio-temporal correlation. Competitive schemes have also been proposed. They usually select the best predictor from a given set, and send the index of this predictor as side information. In [8], the set is composed of three neighboring motion vectors. In this competitive scheme, temporal predictors are missing.

This paper proposes two new techniques to improve these schemes: first, a competitive spatio-temporal scheme for the prediction of the motion vectors is introduced, including a modification of the RD criterion. Second, the occurrence of the SKIP mode is increased using a conditional spatio-temporal predictor. The modifications are implemented and tested into the latest H.264 reference software.

The remaining of this paper is organized as follows: a summary of H.264 motion vector coding is presented in Section 2. The proposed modifications are discussed in Section 3. Section 4 briefly comments the impact of the proposed method on the complexity and presents simulation results in which the average 4% compression gain (and up to 20% for complex motions), with equivalent quality to today's highest H.264 profile is highlighted.

2. STATE OF THE ART OF MV CODING

We describe here the non-normative process implemented in the H.264 reference software to select the motion vectors. H.264 applies predictive motion vectors coding. The motion vector residual ε_{mv} is given by:

$$\varepsilon_{mv} = mv - p \qquad (1)$$

where *mv* is the motion vector and *p* is a median of the three neighboring motion vectors depicted in Fig. 1 (mv_a, mv_b,

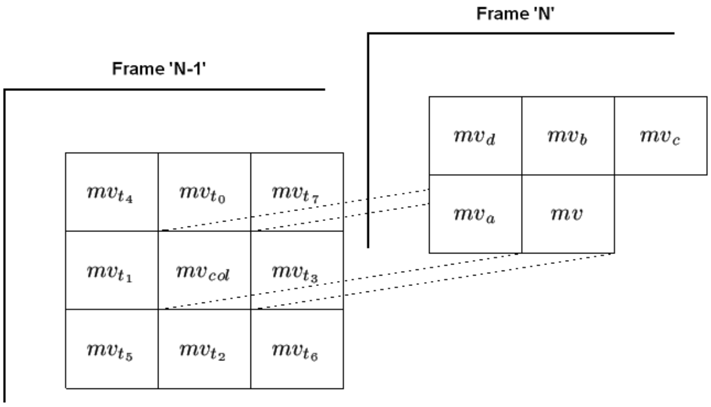


Figure 1: Designation and location of spatial and temporal vectors used for the prediction.

mv_c). However if mv_b is not available, mv_b value is equal to mv_d. If one or more neighboring motion vectors are not available, p value is equal, according to the neighboring motion vectors availabilities, to mv_a or mv_b or mv_c or 0.

The best trade-off between quality and bitrate is obtained by minimizing the rate-distortion criterion:

$$J = D + \lambda R \tag{2}$$

where D is the distortion computed in the spatial or transformed domain and λ is the Lagrange multiplier. R is the rate which introduces all bitrate components [9]:

$$R = R_r + \lambda_m R_m + \lambda_o R_o + \lambda_{mv} R_{mv} \tag{3}$$

where R_r is the rate for block residue (luma+chroma), R_m the rate of macroblock mode, R_{mv} the rate of motion vector residue and R_o the rate of others components (header, coded block pattern CBP, stuffing bits, delta quantization). λ_m, λ_o and λ_{mv} are weighting factors depending on the quantization step. The estimators of the distortion and rates are described in [3]. This selection is computationally intensive, yet it is optimal in a RD sense. Note however that this process yields a non natural motion field, and that this selection process is made among all block partitions (16x16, ..., 4x4), all reference frames, at each sub-pixel accuracy.

A particular case is that of the SKIP mode. A skipped macroblock has neither block residue, nor motion vector or reference index parameter to transmit except the mode itself. The motion vector predictor for the SKIP mode is almost computed as the motion vector predictor for an Inter 16x16, except that if mv_a or mv_b are not available the predictor is equal to 0.

Fig. 2 shows the relative bitrates proportions of the components in Eq. 3, depending on the quantization parameter and for a high profile H.264. At low bitrate (high QP), the motion information R_{mv} is the major part of the total bitstream, and can reach up to 38%.

These observations have motivated our research based on reduction of entropy of motion vector residue with a joint selection and coding of the motion vectors in an RD optimal framework, with adapted temporal and spatial redundancies exploitation in a competition method.

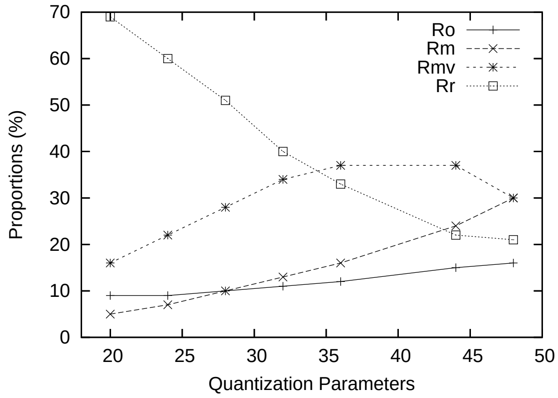


Figure 2: Bitrate proportion depending on the quantization parameter (QP).

3. COMPETITION-BASED MV CODING

This section details the two modifications made on the selection of the motion vector predictor, in comparison with the H.264 standard scheme described in section 2.

3.1 Competition and RD selection of the best predictor

When dealing with lossless coding of motion vectors, the efficiency is closely related to the predictor performance. A competitive scheme allows selecting the best predictor from a given set, and implies several possible predictors for one motion vector. We have defined a set $\mathscr{P}$ including spatial, temporal and spatio-temporal predictors as depicted in Fig. 1. The available spatial predictors are the neighboring motion vectors mv_a, mv_b, mv_c, mv_d, and the H.264 median predictor $mv_{H.264}$. The temporal predictors are the collocated motion vector mv_{col} (motion vector at the same position in the previous frame) and 2 temporal median predictors mv_{tm5} and mv_{tm9} defined as:

$$mv_{tm5} = median\{mv_{col}, mv_{t_j}\}, \forall j < 4 \tag{4}$$

$$mv_{tm9} = median\{mv_{col}, mv_{t_j}\}, \forall j < 8 \tag{5}$$

Spatio-temporal predictors are combinations of spatial and temporal ones. In particular, mv_{spt} is defined by:

$$mv_{spt} = median\{mv_{col}, mv_{col}, mv_a, mv_b, mv_c\} \tag{6}$$

A mode i and a residual ε_{mv_i} is now associated to each predictor $p_i \in \mathscr{P}$:

$$\varepsilon_{mv_i} = mv - p_i, \forall i \in [1,n] \tag{7}$$

where n is the number of predictors of $\mathscr{P}$.

The mode needs to be transmitted in the bitstream as well as the residual. The weight of this new information is however significant (in average 3.5% of the bitrate, and 12.5% of the motion information). The efficiency of the competitive scheme is therefore related to the trade-off between this additional cost and the gain obtained by a more accurate prediction.

For the selection of the motion vector Eq. 3 is replaced by Eq. 8: the rate of the motion vector residue R_{mv} is replaced by $R_{mv/mm}$ that contains the cost of the predictor and the cost of the mode information.

$$R = R_r + \lambda_m R_m + \lambda_o R_o + \lambda_{mv/mm} R_{mv/mm} \tag{8}$$

$R_{mv/mm}$ is given by:

$$R_{mv/mm} = \min\{\varsigma(\varepsilon_{mv_i}) + \varsigma(i)\}_{i=1..n} \tag{9}$$

where $\varsigma(x)$ is the computed cost of the data x in the bitstream.

The prediction modes have been encoded using CABAC. A key element in our method is that the decoder is sometimes able to “guess” which predictor has been used. The encoder can simulate the decoding process to see if the predictor can be guessed or not and accordingly it encodes or not the mode. This guess relies on the information available at the decoder: the value of the residual ε_{mv_i}, the knowledge of the set $\mathscr{P}$, the coding modes for neighboring blocks (spatially and temporally). In practice the mode can be guessed in 27% of the cases.

3.2 Modification of the motion vector for the SKIP mode

As explained in Section 2 the SKIP mode is a powerful mode. Its selection basically means that it is more interesting in a RD sense to send nothing instead of the residual and the vector. This mode is largely used, especially on sequences with static backgrounds. Our objective is consequently to increase as much as possible its occurrence. To increase the number of SKIP mode without modification on the RD criterion, the only solution is to change the motion vector for the SKIP mode by a vector which gives a lower distortion. The standard SKIP is strictly spatial and jumps from the $median(mv_a, mv_b, mv_c)$ to mv_a or mv_b or mv_c or 0 value, if data for computing the median are not available. We modify this process by defining a privileged ordering of spatial and temporal predictors, and the jump from one to another depends on the availability of the data needed to compute the prediction. This availability depends on the position of the block, and the encoding modes (Intra/Inter). The decoder is able to reproduce the same behavior, so no additional information is transmitted.

In practice, several ordering have been tested. The most efficient for a general test set is given in Table 1. From this table we see that: 1- the spatial median is used if mv_a, mv_b, mv_c are available, 2,3- otherwise the temporal median with 9 then with 5 components is used if all its components are available, otherwise 4- the collocated vector is used, then 5- mv_a, 6- mv_b, 7- mv_c, and finally 8- the value ’0’.

Level	Predictor	Availability of data
1	$median(mv_a, mv_b, mv_c)$	mv_a, mv_b, mv_c
2	mv_{tm9}	$mv_{t_j}, \forall j < 8$, mv_{col}
3	mv_{tm5}	$mv_{t_j}, \forall j < 4$, mv_{col}
4	mv_{col}	mv_{col}
5	mv_a	mv_a
6	mv_b	mv_b
7	mv_c	mv_c
8	0	

Table 1: Predictor ordering for SKIP motion vector.

Motion vector coding and modification of the SKIP mode allowed us to obtain some compression gains for equivalent quality that are described in following section.

4. EXPERIMENTAL RESULTS AND COMPLEXITY

Simulations were performed on the JM10.0 H.264 reference software [2]. The high profile (Fr-Ext) is selected with 32x32 search range, RD-optimization enabled (RdOpt=1), and CABAC entropy coding. It corresponds to the highest possible quality using the H.264 normative tools and efficient non-normative encoding decisions implemented in this recent JM, except B-frames and multiple reference frames, for which the modification are not yet implemented. We plan to do it in the near future and expect improved results (B-Frames tend to increase the proportion of motion information). The test set is composed of 10 QCIF, 10 CIF, and 5 SD sequences of 100 frames each, with various representative contents and motions. Quantization parameters are equal to 30, 36, and 42. As the obtained gain is equivalent at all resolutions, only CIF results are detailed below.

4.1 Impact on complexity

The objective is not to give a complexity analysis but to highlight the impacted modules of the algorithm. Indeed, for each candidate vector of the search range, the predictors p_i, the residual ε_{mv_i} and $R_{mv/mm}$ are computed, see Eq. 7 and Eq. 9. The main complexity increase comes from $\varsigma(\varepsilon_{mv_i})$. The computational impact of the modification of the SKIP mode is negligible. In terms of memory management, temporal predictions require the storage of the motion vectors and coding modes of the previous frame. As an indication, simulations performed on the JM10.0 reflect an average increase of 3% of the computational time for a non-optimized C code.

4.2 Number of predictors

The results of the proposed competition scheme depend on the number and type of predictors. Our experiments made with either 1 ($mv_{H.264}$), 2 ($mv_{H.264} + mv_{col}$) or 4 ($mv_{H.264} + mv_a + mv_{col} + mv_{tm9}$) predictors and the SKIP modification, show that the best compromise is obtained with 2 predictors. The results of this experiments are illustrated in Fig. 3. The average bitrate saving for these 3 configurations is respectively 1.8%, 4.2%, and 3.9%. It can be noticed that except for one sequence (Carphone), the bitrate reduction for the configuration with 2 predictors is higher. Obviously the re-

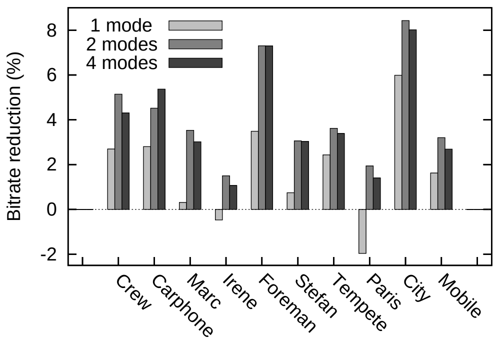


Figure 3: Bitrate reduction for 1, 2 and 4 modes configuration.

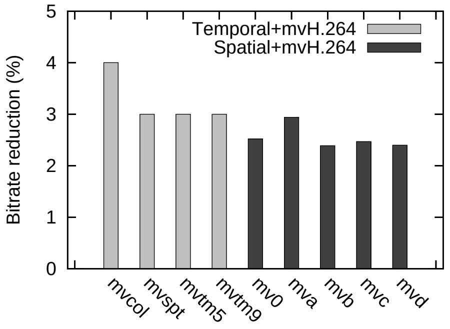


Figure 4: Bitrate reduction for each combination of median and another predictor.

duction of the motion vector bitrate (R_{mv}) is increased when using 4 modes, but the compromise with the mode coding ($R_{mv/mm}$) leads to slightly worse results. An adaptive set of predictors according to statistical and local characteristics is expected to increase the gain.

In Fig. 4, the median $mv_{H.264}$ is associated to each possible predictor previously described, to select the best “2 predictors” configuration. It can be noticed that a combination of the $mv_{H.264}$ with a temporal predictor gives better results than $mv_{H.264}$ with ony other spatial predictor. This is explained by the fact that the $mv_{H.264}$ and a spatial predictor have similar values and the two motion vector residues have the same cost in number of bits.

4.3 Two predictors configuration

We have selected as the best method the combination of the spatial median predictor of the standard H.264 and the temporal predictor given by the collocated vector. All the results below are given with these two predictors combined.

4.3.1 Bitrate reduction on the motion information

Fig. 5(a) compares the cost of the motion information of a standard H.264 with the new cost ($R_{mv/mm}$). The average reduction is close to 10%. The increase at low bitrate and for complex motions is higher, because the number of bits for the motion information represents a larger proportion of the global bitrate as illustrated in Fig. 2.

QP	Spatial predictor $mv_{H.264}$	Temporal predictor mv_{col}	$mv_{H.264} = mv_{col}$
30	62%	38%	16%
36	56%	44%	15%
42	46%	54%	17%
Average	55%	45%	16%

Table 2: Distribution of the predictors selection.

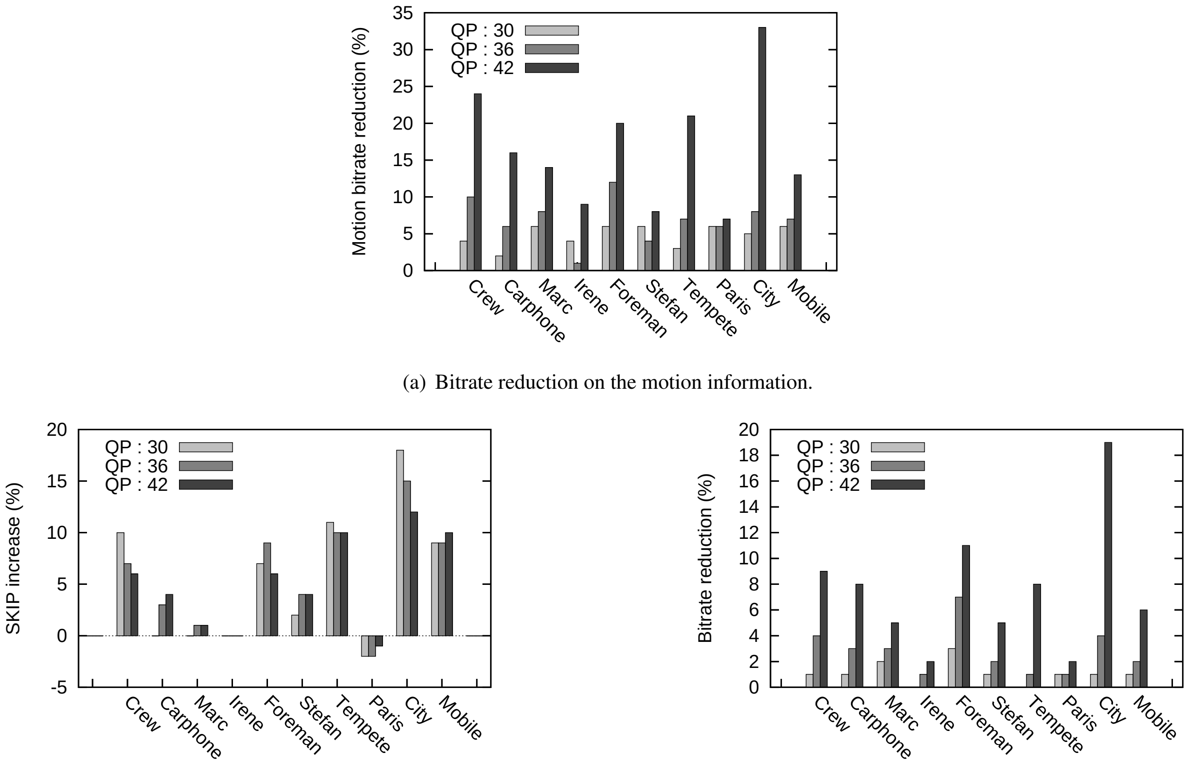
4.3.2 Relative increase of the SKIP mode

Fig. 5(b) shows the percentage of increase of the number of macroblocks encoded with the SKIP mode. The average is 6%. The increase is not correlated with the QP but with the sequence type. It is to notice that only sequences with static background (Marc, Irene, Paris) do not benefit from the modification of the SKIP. For these sequences the order given in section 3.2 is not optimal: the value ’0’ should have higher priority. It is also interesting to notice from Fig. 3 that the average bitrate gain resulting from the SKIP modification only (without motion vector competition) is 1.8%. This average increase is small, however it reaches 2.8% on non static background sequences, which is significant in contrast with the quite null additional complexity. Again, some adaptivity to sequence context may provide improved results.

4.3.3 Global bitrate reduction

The analysis of the selection of the spatial or temporal prediction mode has been performed and is depicted in Table 2. It shows that on average on the test set, the temporal predictor mv_{col} is selected 45% of the time. Given that the selection results from a RD choice, this average result confirms that the temporal predictors are useful. Note that these values exclude the cases where both predictors provide the same value, which represents in average 16%. As an interesting feature, the percentage of selection of the temporal predictor increases when the QP increases. Sequence by sequence this percentage evolves between 24% and 62% depending on the type of the sequence.

Fig. 5(c) represents the percentage of bitrate saved for each sequence and each QP value. It can be noticed that the proposed method offers a compression gain for all sequences of the test set even those with simple motion, with an average decrease of the PSNR of 0.04dB (maximum loss: 0.12dB), compared to the standard H.264. The visual quality is equivalent. The average bitrate gain is 4%, and can reach 20%. Obviously, the increase is lower on sequences with simple or no motion (e.g., videoconferencing sequences), given that the SKIP mode is already widely used. For sequences with fast or complex motions, the compression gain is higher. Finally, sequences exhibiting global and constant motion, combined with a high level of spatial details (such as City) take full advantage of the temporal prediction, whereas the classical spatial median usually fails. The bitrate reduction is also closely related to the QP value. At low bitrates, the motion information tends to become a significant part of the total bitstream, so its reduction leads to the highest improvements.



(a) Bitrate reduction on the motion information.

(b) Relative increase of the SKIP mode.

(c) Global bitrate reduction.

Figure 5: Experimental results for 2 predictors competitive scheme.

5. CONCLUSION

In this paper, a competitive method for the prediction of the motion vectors is proposed. Both spatial and temporal predictors are used and optimally selected via a rate-distortion criterion that considers the cost of the residual and the mode for the prediction. In addition, a modification is proposed to increase the amount of skipped macroblocks. These two combined techniques, implemented in JM10.0 H.264 reference software, provide a systematic bitrate reduction (4% in average and up to 20% with negligible PSNR degradation) with slight computational increase. It is planned to implement the modifications for B-frames and multiple reference frames to increase the gain.

REFERENCES

[1] ITU-T Recommendation H.264 and ISO/IEC 14496-10 AVC, "Advanced video coding for generic audiovisual services," version 3: 2005.

[2] H.264/AVC software coordination, *K. Suehring*, http://iphome.hhi.de/suehring/tml/.

[3] K.P. Lim, G. Sullivan, and T. Wiegand, *Text Description of JM Reference Encoding Methods and Decoding Concealment Methods*, JVT-N046 contribution, Hong-Kong, Jan. 2005.

[4] L.A. Da Silva Cruz and J.W. Woods, "Adaptive motion vector quantization for video coding," in *IEEE ICIP*, Oct. 2000, vol. 2, pp. 867–870.

[5] J. Yeh, M. Vetterli, and M. Khansari, "Motion compensation of motion vectors," in *IEEE ICIP*, Oct. 1995, vol. 1, pp. 574–577.

[6] M.C. Chen and A.N. Willson, "A spatial and temporal motion vector coding algorithm for low-bit-rate video coding," in *IEEE ICIP*, Oct. 1997, vol. 2, pp. 791–794.

[7] A.M. Tourapis, F. Wu, and S. Li, "Direct mode for bipredictive slices in the H.264 standard," *IEEE Trans. on CSVT*, vol. 15, no. 1, Jan. 2005.

[8] S. Deuk Kim and J. Beom Ra, "An efficient motion vector coding scheme based on minimum bitrate prediction," *IEEE Trans. on Image. Proc*, vol. 8, no. 8, pp. 1117–1120, Aug. 1999.

[9] G.J. Sullivan and T. Wiegand, "Rate-distortion optimization for video compression," *IEEE Signal Proc. Mag.*, pp. 74–90, 1998.