RATE-DISTORTION-COMPLEXITY OPTIMIZATION OF FAST MOTION ESTIMATION IN H.264/MPEG-4 AVC

Jesper Støttrup-Andersen

Milestone Systems A/S
Banemarksvej 50, DK-2605 Brøndby, Denmark

Søren Forchhammer and Shankar M. Aghito

Research Center COM, DTU
Bldg. 345v, DK-2800 Kgs. Lyngby, Denmark

ABSTRACT

This paper presents an operational method for optimizing integer motion estimation in real-time H.264 encoding with respect to the trade-off between rate-distortion and complexity. This three-parameter problem is converted into a more tractable two-parameter problem by a simple approximative elimination of either the rate or the distortion term by converting small differences in the term to be eliminated. This paper also presents an H.264-enhanced implementation of the fast EPZS motion estimation algorithm, which adds three early-stop criteria controlled by predefined thresholds allowing to stop after 16×16 and 8×8 block types and for each additional reference frame. Results for interlaced SDTV material are presented with indications of the applicability of the operational method when applied to the presented motion estimation procedure. For example, a speed-up by a factor of 4 compared to a basic H.264-adapted EPZS implementation is achieved at only 1% increase in rate.

1. INTRODUCTION

Currently, the MPEG-2 video coding standard enjoys the most widespread usage among compression schemes for broadcast applications. Developments in coding technology has lead to more efficient international standards, the most recent being the ITU H.264/ISO MPEG-4 AVC hybrid video coding standard [1, 2]. Its reported perfomance shows that a similar (subjective and objective) quality is obtainable at half the bit-rate of MPEG-2. This gain in coding efficiency comes at a price of significantly increased encoding complexity for H.264. The main part of the increased complexity stems from new features in the motion compensation stage. These include seven different block sizes and multiple reference-frame prediction that also allow B-frames as reference frames. The increased coding performance is achieved in the associated prediction part of H.264 together with e.g. the in-loop de-blocking filter, more intra-coding possibilities and better entropy coding.

For a real-time implementation of an H.264 encoder, e.g. using a programmable video DSP platform, Full-Search motion estimation is not feasible. A less complex solution must be applied to match the available computational resources. This will lead to reduced (rate-distortion) coding efficiency dictating a compromise. In this paper, a simple operational method is introduced, which permits optimization of rate-distortion vs. the complexity of the motion estimation (ME). The general optimization method is here applied to a fast ME method for a high-end H.264 encoder aimed at Standard Definition TV (SDTV) broadcast applications with a target bitrate around 1.5 Mbits/s. This ME method allows for control of the complexity through some early-stop criteria. The complexity of motion estimation is here measured by the number of *weighted search positions per second* (WSP/s) due to the multiple block sizes in H.264. This is found by applying a weight to the number of search positions for each block type according to the block size relative to that of a 4×4 block, i.e. a 4×4 block is weighted by 1, whereas e.g. an 8×8 block is weighted by 4, etc. For clarity, the complexity will be represented by MWSP/s, i.e. *millions* of WSP/s.

The applied motion estimation procedure with introduction of the enhancements is presented in Section 2. The issues of joint optimization of rate-distortion vs. complexity is introduced in Section 3 with results and discussion in Section 4.

2. FAST MOTION ESTIMATION FOR H.264

Full-Search (FS) motion estimation is widely used as reference within video coding. This method is also implemented in the H.264 reference software encoder (model version 6.1) with rate-distortion (RD) optimization of the motion vector (MV) and coding mode selections. Even though FS may be possible to use in e.g. MPEG-2, the inclusion of the tree-structured block division of macroblocks (MB) and the multiple reference frames in H.264 makes FS intractable. The focus here is on reducing the complexity of the *integer* part of the ME process, and therefore the default subpixel refinement procedure of the reference software is used to obtain quarter-pixel resolution of the MVs.

Recently, flexible ME algorithms with irregular search strategies based on gradient-descent search have become popular, primarily due to their large reduction in the number of visited search positions while maintaining a good coding performance. In these algorithms, it is implicitly assumed that the error-surface within a local search area is monotonically decreasing towards a single point where the error is minimal. This minimum is assumed to be the global minimum within the search area. An example of such a method is the *Enhanced Predictive Zonal Search* (EPZS) algorithm [3] that employ pattern search around a local search center, which is determined by initially evalutating a number of spatial and temporal prediction vectors. A number of criteria stops the algorithm prematurely, if a given MV is found to be adequate.

In our work, the EPZS algorithm has been employed with some required modifications allowing it to be used for H.264. The algorithm has been implemented with basis in the EPZS description [3] originally targeted at previous MPEG standards with some inspiration from [4]. The modifications include that for every MB the MVs related to the best block type are saved across all block types. The predictors are then taken from this set of best chosen vectors disregarding the actual block type. This requires that the

This work was supported in part by a grant from the Danish Research Agency. The first author was with COM, DTU during the project. Author contact: sf@com.dtu.dk

0-7803-8554-3/04/$20.00 ©2004 IEEE.

coding cost for each MV is stored such that this cost is distributed over the 4×4 blocks within each block type. Furthermore, the H.264-defined MV-predictor is checked first, then the usual median predictor is checked, and the spatial prediction vector placed top-left relative to the current block is included as well. This implementation shall be referred to as the *H.264-adapted EPZS* algorithm. The best MV is found in a RD-optimized sense by minimizing the *Lagrange* cost function

$$J(\mathbf{m}, r, \lambda_{\text{MV}}) = \text{SAD}(s, c(r, \mathbf{m})) + \lambda_{\text{MV}}(R(\mathbf{m}-\mathbf{p}) + R(r)) \,, \quad (1)$$

where $\mathbf{m}$ is the current MV in the reference frame denoted by r relative to the prediction vector $\mathbf{p}$. The reference frame signal is c, and the original video signal is s. The Lagrange multiplier is λ_{MV}, $R(\mathbf{m}–\mathbf{p})$ is the bits required to code the MV information, and $R(r)$ is the bits to code the reference frame index. SAD is the usual Sum of Absolute Differences measure calculated for the given video signals at the current block size. The best coding mode (block type) is selected after applying full entropy coding using a similar RD-optimization with a different Lagrange multiplier, λ_{MODE}.

The adapted EPZS algorithm has been extended with additional early-stop criteria of two different types. This shall be referred to as the *H.264-enhanced EPZS* algorithm. The first criterion checks the cost function in (1) for the best MV after completing the search in the first reference frame for the current block type and determines if the cost has reached an acceptable value. If the cost is worse, the next reference frame is searched and the criterion is reevaluated; otherwise the search is stopped and continues at the next sub-block. This method is partly inspired from motion statistics observations, which show that the coding penalty is often negligible when using fewer than five reference frames, as also reported in [5]. A related suggestion has been given in [6].

The second added criterion consists of two sub-criteria for early-stopping the ME process after searching the 16×16 and the 8×8 block modes, respectively. The Lagrange cost is evaluated after initially searching the 16×16 block type. If the cost is accepted by the criterion, all remaining block types within the current MB are skipped. If the cost is worse, the 16×8, 8×16 and 8×8 partitions are searched. A similar evaluation is carried out after the 8×8 type to determine if the smallest three block types should also be searched. The three stop-criteria are in effect in parallel (with the reference frame criterion checked first) and are controlled by simple comparisons with three pre-defined and fixed threshold values.

3. RATE-DISTORTION-COMPLEXITY OPTIMIZATION

In a practical implementation of the H.264-enhanced EPZS algorithm for real-time applications, the fixed resources available dictates both a reduction of the complexity and some control of the complexity in order to achieve the most efficient coding. The complexity may be reduced and partly controlled by adjusting the three early-stopping criteria in the ME algorithm. Therefore, it is necessary to balance and optimize the selection of the constant thresholds with respect to both rate-distortion *and* complexity.

An operational method for selecting the thresholds optimally has been developed. The basic idea is to transform the three-dimensional problem of concurrently optimizing rate-distortion and complexity into a more tractable two-dimensional problem. The origin of the threshold-search (referred to as the *point of origin*) is taken with basis in the result from H.264-adapted EPZS, i.e. equal to H.264-enh. EPZS with no early-stopping (all thresholds zero). The optimal thresholds with lower complexity are then searched.

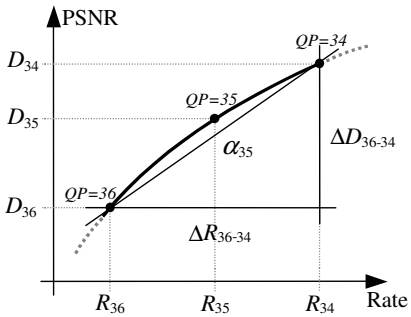

Fig. 1: Illustration of how the slope α_{QP} is found when optimizing for the fixed QP value 35.

In broadcast applications, rate control would be applied to obtain a given target rate. This is initially disabled, and fixed quantization is used in the experiments instead. Having a fixed quantization parameter (QP) for all picture types results in a more or less constant distortion for a given sequence. Therefore, the minor changes in distortion relative to the point of origin can be converted into small changes in rate by using a local slope for the searched parameter setting. Thereby, the small differences in distortion are included as part of the rate term, and this total term is referred to as the *modified rate* (*rate**). This effectively eliminates the distortion parameter, and the optimization problem has thus been reduced to a problem between two parameters, rate* and complexity.

The local slope for each searched setting is approximated as the linear relation between the rate and the distortion changes between results for those fixed QP values that surround the given QP value in the optimization search. A further approximation has been employed, which replaces the local slope value for each setting with a single global slope found in the point of origin. Investigations using slightly different global slope values found for other settings have confirmed the reported findings indicating that the suggested approximation is reasonable. The process for finding the slope α_{QP} is illustrated in Fig. 1 and given by

$$\alpha_{\text{QP}} = \left(D_{\text{QP+1}} - D_{\text{QP-1}}\right) / \left(R_{\text{QP+1}} - R_{\text{QP-1}}\right) , \quad (2)$$

where D_{QP} is the distortion as PSNR for the given QP value, and R_{QP} is the associated rate.

The optimization search has been carried out manually testing different interesting settings with intermediate results used as guide for subsequent searching. Included in the searching are *fixed* ME settings that enforce restrictions on the number of reference frames and/or the block types allowed. An extreme example of such a fixed setting is the *MPEG-2-like* setting, where only a single reference frame (and the additional one for B-frames) and only 16×16 blocks are allowed mimicking the features of MPEG-2. This can also be achieved by having all thresholds equal to infinity.

4. EXPERIMENTAL RESULTS AND DISCUSSION

Results from experiments with the above rate-distortion vs. complexity optimization method are presented in this section using the H.264-enhanced ME algorithm (Section 2). The well-known interlaced PAL (SDTV) video sequences *Cycling*, *MobCal* and *Barcelona* constitute the material for the initial test set. Field-coding has been employed with up to five reference frames (i.e. ten reference *fields*) plus the additional reference frame for B-frames. The search range has been set to ±39; RD-optimization, Hadamard,

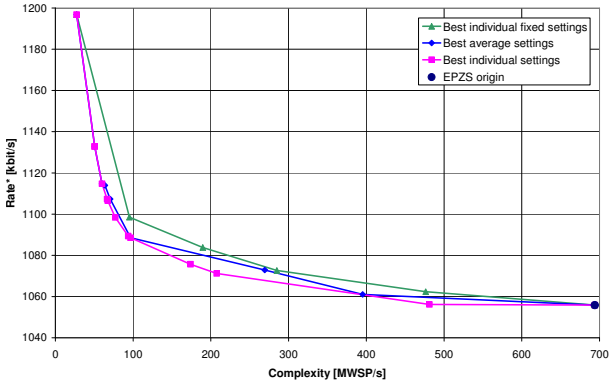

Fig. 2: R*-C curves for best settings, sequence *Cycling*.

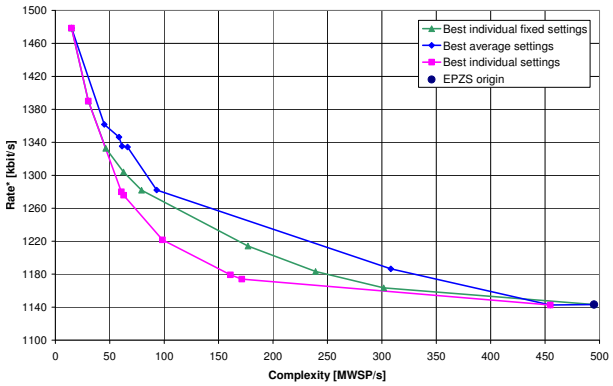

Fig. 3: R*-C curves for best settings, sequence *MobCal*.

the in-loop de-blocking filter and all block modes have been used. The GOP-structure follows the usual *IBBP* structure with a length of 12 frames. One hundred frames are used from each sequence.

The best settings in a rate*-complexity (R*-C) optimal sense are those with R*-C points that define the lower convex hull curve in the R*-C plane. This is illustrated in Figs. 2 and 3 for two of the initial test sequences. Generally, the results for *Barcelona* show similar tendencies as those for *Cycling*, wherefore only the latter is shown here. The *best individual settings* are found among all the individually-tested settings for each sequence. The so-called *best average settings* are found among all tested settings that the three initial test sequences have in common and averaging both the R* and the C values. In the two first figures, the results are depicted with the actual values (and not the averages) for the particular sequence, whereas the average values are used in Fig. 4. The *best individual fixed settings* for the three initial test sequences separately and for their best average are listed in Table 1 and are depicted in the three figures as well. The EPZS point of origin for the searches with all thresholds set to zero is marked with a filled circle. The distortion (D) values in these points of origin are 32.53 dB (*Cycling*), 28.90 dB (*MobCal*) and 32.92 dB, respectively. The results for the best average fixed settings with the *worst-case complexity* substituting the average C are also given in Fig. 4. The dotted curve in the same figure illustrates the average of all best points for the initial test sequences. Before averaging for this curve, linear interpolation has been used between adjacent best settings, if a setting was not available for a desired complexity value.

Focusing on the best individual rate*-complexity settings, it is recognized that all three initial test sequences seem to obtain a

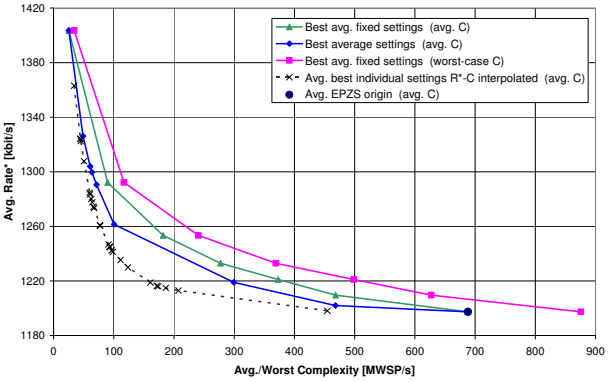

Fig. 4: R*-C curves for best settings, average over initial test set.

good compromise in coding performance at around 200 MWSP/s. This means that the complexity can be reduced by a factor of 3-5 with just a small increase (0.5-2.5%) in the rate* relative to that of the basic H.264-adapted EPZS algorithm. The latter reduces C by itself by a factor 300-500 relative to the FS algorithm, and thus a total reduction by more than a factor of 2000 is possible! It should be emphasized that the fixed settings are useful alternatives, as these settings only increase R* by a quite small amount (Table 1). For example, for $(r_{\mathrm{M}}, b_{\mathrm{M}}) = (2, 8)$, the rate* increases by less than 5% at a complexity of less than 200 MWSP/s. Pushing C lower than 100 MWSP/s affects the R* to increase more and is not advisable if high performance is desired. At the lowest complexity levels the visual quality also suffers with blocking-like artefacts.

It is important to note that the best R*-C settings obtained for the initial test sequences do not use the same values for the thresholds. Where *Cycling* and *Barcelona* display similar behaviour permitting large threshold values for the multiple reference frame early-stop criterion and smaller values for the early-block-stopping criteria, the *MobCal* sequence shows almost the opposite behaviour (Table 1 illustrates this tendency). This fact combined with the advantage of controlling the resource usage in practical situations suggests an *adaptive* scheme to accomodate this content-dependency of the ME by adaptively adjusting the thresholds of the early-stopping criteria. The potential gain of such a scheme is illustrated by the average curves in the previous figures when compared with the best individual R*-C performances. The average curves also indicate the potential coding losses experienced if only a single set of pre-defined thresholds is applied to any material without adapting the thresholds to the content. Thus an adaptive scheme will provide some robustness.

The R*-C performance has also been evaluated for five other SDTV sequences (the evaluation test set): *Tabletennis*, *Parkrun* [7], *Stockholm* [7], *Rafting* and *Football*. Averaged results using the best average fixed settings from the initial test set are also depicted in Table 1 along with the average loss ($\Delta R^*_{\mathrm{avg.}}$) relative to the best individual fixed setting for each sequence of the evaluation set. The previous R*-C performances are confirmed by these results, where e.g. good compromises exist close to 200 MWSP/s with just 0.62-1.58% increase in R* relative to the base EPZS performance.

Introducing *rate control* instead of using the fixed QP values changes the second parameter of interest to be the distortion. A similar technique as described previously for the (modified) rate term can be used for reducing the rate-distortion-complexity problem to only involve the distortion and the complexity parameters. The invariable small perturbations in the rate (for small test se-

Table 1: List of the best rate*-complexity results using only the *fixed* settings for the three initial test set sequences. The complexity (C) is measured in the unit of MWSP/s and the modified rate (R^*) is given in kbits/s. The ME settings are given as the max. number of reference frames (r_{M}) and the min. block type used (b_{M}). The far-right columns are for the averages of the results for the five evaluation test sequences.

	Cycling			***MobCal***			***Barcelona***			**Best avg. initial test set**			**Evaluation set best avg.**			
No.	C	R^*	$(r_{\mathrm{M}}, b_{\mathrm{M}})$	C	R^*	$(r_{\mathrm{M}}, b_{\mathrm{M}})$	C	R^*	$(r_{\mathrm{M}}, b_{\mathrm{M}})$	$C_{\mathrm{avg.}}$	$R^*_{\mathrm{avg.}}$	$(r_{\mathrm{M}}, b_{\mathrm{M}})$	$C_{\mathrm{avg.}}$	$R^*_{\mathrm{avg.}}$	$(r_{\mathrm{M}}, b_{\mathrm{M}})$	$\Delta R^*_{\mathrm{avg.}}$
1	27	1197	$(1, 16)$	15	1479	$(1, 16)$	34	1535	$(1, 16)$	25	1404	$(1, 16)$	20	1260	$(1, 16)$	0.00%
2	95	1099	$(1, 8)$	30	1390	$(2, 16)$	117	1411	$(1, 8)$	90	1292	$(1, 8)$	74	1167	$(1, 8)$	0.00%
3	190	1084	$(2, 8)$	46	1333	$(3, 16)$	175	1400	$(1, 4)$	182	1253	$(2, 8)$	148	1154	$(2, 8)$	0.93%
4	285	1073	$(2, 4)$	63	1304	$(4, 16)$	349	1396	$(2, 4)$	277	1233	$(3, 8)$	225	1146	$(3, 8)$	0.93%
5	476	1062	$(5, 8)$	79	1282	$(5, 16)$	525	1393	$(3, 4)$	373	1221	$(4, 8)$	301	1144	$(4, 8)$	1.12%
6	694	1056	$(5, 4)$	177	1214	$(3, 8)$	702	1392	$(4, 4)$	468	1210	$(5, 8)$	379	1146	$(5, 8)$	1.49%
7				239	1183	$(4, 8)$				688	1197	$(5, 4)$	589	1124	$(5, 4)$	0.00%
8				302	1163	$(5, 8)$										
9				495	1143	$(5, 4)$										

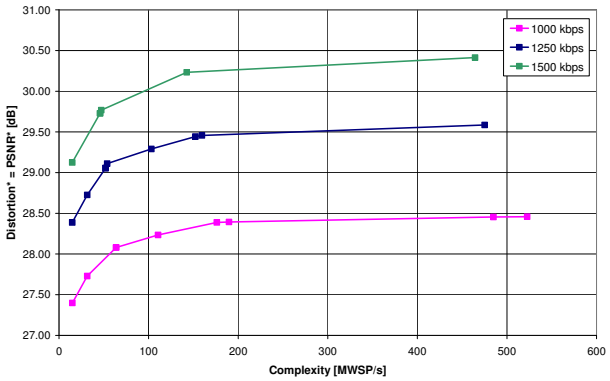

Fig. 5: D^*-C curves for best settings using rate control, *MobCal*.

quences) due to the buffer of the rate-control system can thus be translated into a *modified distortion* term, or *distortion** (D^*).

Three different bitrates (1.0, 1.25 and 1.5 Mbits/s) within the range of interest have been investigated using the rate-control algorithm of [8] merged into the modified H.264 reference software. For simplicity, the previously found global slopes for fixed QP values have been re-used with a reasonable outcome. Only the best individual settings for each initial test sequence obtained in the previous rate*-complexity experiments have been tested. Results are examplified in Fig. 5 for the *MobCal* sequence with similar results obtained for the other two initial test sequences. At 1.5 Mbits/s the visual quality was evaluated to be adequate for broadcast applications. These results help verify the optimization scheme, and it is recognized that a good performance compromise exists below 200 MWSP/s with negligible reduction of D^* in comparison with not using any early-stopping. It is expected that the optimal threshold values depend on and should be optimized for each QP. Even then, the parallel displacement of the curves in Fig. 5 suggests that a set of nearly-optimized threshold values may be applicable across a small range of bitrates (here 1.0-1.5 Mbits/s).

5. CONCLUSIONS

An operational method has been presented for optimizing the integer motion estimation part of an H.264 encoder with respect to rate-distortion vs. complexity. The complexity is measured as the number of visited search positions weighted by the size of each block type. The three-parameter problem is converted into a more tractable problem by a simple approximative elimination of either the rate or the distortion term by converting the small differences to be part of the other term. An H.264-enhanced implementation based on the fast EPZS motion estimation algorithm has been introduced and applied to interlaced SDTV material in the target bitrate range of 1.0-1.5 Mbits/s. This algorithm employs three early-stop criteria controlled by simple pre-defined and constant thresholds, which allow stopping the motion search after 16×16 and 8×8 block types and also for each additional reference frame. Results from simultaneous rate-distortion vs. complexity optimization of these three thresholds have been given with indications of the applicability of the method. For instance, a speed-up of 3-5 times compared to the basic H.264-adapted EPZS was achieved with an average rate increase of less than 1.5%.

6. ACKNOWLEDGMENTS

The authors would like to thank for valuable contributions from Eskil Faber and Jan H. Thomsen, Scientific-Atlanta, Denmark.

7. REFERENCES

[1] T. Wiegand and G. J. Sullivan (ed.), "Draft ITU-T Rec. and FDIS (ITU-T Rec. H.264-ISO/IEC 14496-10 AVC)," *H.264/MPEG-4 AVC (JVT-doc. JVT-G050)*, Mar. 2003.

[2] T. Wiegand et al., "Overview of the H.264/AVC video coding standard," *IEEE Trans. Circuits and Systems for Video Tech.*, vol. 13, no. 7, pp. 1–19, July 2003.

[3] A. M. Tourapis, "Enhanced predictive zonal search for single and multiple frame motion estimation," *VCIP 2002, Proc. of SPIE*, vol. 4671, pp. 1069–1079, 2002.

[4] P. Yin et al., "Fast mode decision and motion estimation for JVT/H.264," *Proc. ICIP-2003*, vol. I, pp. 901–904, 2003.

[5] D. Alfonso et al., "Detailed rate-distortion analysis of H.264 video coding standard and comparison to MPEG-2/4," *VCIP 2003, Proc. of SPIE*, vol. 5150, pp. 891–902, 2003.

[6] T. Wiegand and B. Girod, *Multi-Frame Motion-Compensated Prediction for Video Transmission*, Kluwer, Boston, 2001.

[7] `ftp://ftp.ldv.e-technik.tu-muenchen.de/pub/test_sequences/`.

[8] Z. Li et al., "Adaptive basic unit layer rate control for JVT," *JVT-doc. JVT-G012*, Mar. 2003.