DECODER-SIDE BLOCK MOTION ESTIMATION FOR H.264 / MPEG-4 AVC BASED VIDEO CODING

Sven Klomp, Marco Munderloh, Yuri Vatis, Jörn Ostermann, Fellow, IEEE

Institut für Informationsverarbeitung
Leibniz Universität Hannover, Appelstr. 9A, 30167 Hannover, Germany
{klomp, munderl, vatis}@tnt.uni-hannover.de

ABSTRACT

In video coding standards like H.264 / MPEG-4 AVC, the encoder performs motion estimation in order to utilise temporal dependencies within a sequence. In addition to the rate of the residue, the encoder has to allocate bits for motion vectors required to compensate the motion at the decoder. This bit rate increases for smaller block sizes, since more motion vectors need to be transmitted. Therefore, motion compensation using dense motion vector field is not feasible for such an architecture.

This paper proposes to estimate motion for coding of B frames at the decoder. Using this decoder-side motion estimation, the transmission of the motion vectors is not necessary and the bit rate is reduced. Furthermore, prediction quality is higher in many cases resulting in a coding gain of up to 1.7 dB at low bit rates and 0.2 dB at higher bit rates.

***Index Terms*—** Motion compensation, Motion vector, Interpolation, H.264 / MPEG-4 AVC, B frames, Direct mode

I. INTRODUCTION

In current video coding solutions, such as MPEG-1,2,4 Video or ITU-T H.26x standards, the encoder estimates the motion in inter frames (P and B frames) and transmits the motion vectors and the residue to the decoder. Thus, temporal correlations between frames are exploited and compression is achieved. Due to block-based motion estimation, accurate compensation at object borders can only be achieved with small block sizes. However, the smaller the block, the more motion vectors have to be transmitted, resulting in a discrepancy to bit rate reduction. Therefore, the block size has a significant impact on compression performance and is limited to 4x4 pixel in H.264 / MPEG-4 AVC.

To overcome the drawbacks of segmentation and block-based motion compensation, motion estimation can be performed at the decoder. In this case, the predicted motion is already known at the decoder and thus, the transmission of the motion vectors can be omitted. Depending on the compensation accuracy, additional bits can be saved in the residue.

The proposed decoder-side motion estimation (DSME) is introduced in Section II. In Section III, the results obtained with this technique are presented and compared with current standards. Evaluation of the results and directions for further research are presented in Section IV. This paper finishes with conclusions in Section V.

II. DECODER-SIDE MOTION ESTIMATION

As mentioned above, the rate-distortion performance can be improved by performing motion estimation at the decoder. The decoder estimates the motion between two key frames, I or P, and interpolates the intermediate B frame using these motion vectors. In conventional motion estimation schemes, the motion vectors are selected by minimising the prediction error between the current frame and a reference frame. Therefore it might occur that the motion estimation algorithm finds motion vectors that produce the smallest residue but do not represent the true motion. Since DSME assumes constant motion to predict intermediate frames, those wrong motion vectors would induce high interpolation errors. Therefore, the motion estimation algorithm has to be redesigned for decoder-side motion estimation.

First, a full-search block matching algorithm estimates the motion vectors between a key frame and the consecutive one (I or P) with full-pel accuracy. Since this vector field will result in overlapped and uncovered areas after frame interpolation, the motion estimation scheme proposed in [1] is used: For each 16x16 block of the DSME frame, a vector is selected from the previously estimated candidates that intercepts the DSME frame closest to the centre of the block (Fig. 1(a)). This motion vector is used as the initial value for the bidirectional motion estimation in which the motion vector is refined in sup-pel accuracy with a smaller search range [2]. Since linear and constant motion is assumed between the key frames, the forward and backward motion vectors are symmetrical (Fig. 1(b)). In the last step, the motion vector field is smoothed by using weighted vector median (WVM) filters [3] in order to detect outliers.

Finally, the DSME frame is predicted with bilinear interpolation using the motion vector field. This DSME frame is than fed into the reference lists (list0, list1) of the H.264 / MPEG-4 AVC coder as shown in Fig. 2 (a). The coder is now able to use the DSME frame as reference for each macroblock. As the DSME frame is a prediction

978-1-4244-3828-0/09/$25.00 ©2009 IEEE

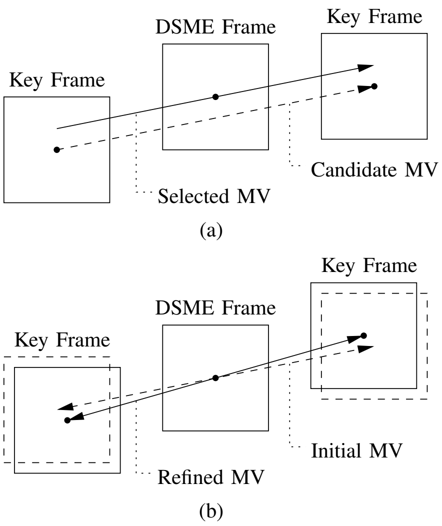

Fig. 1. Motion vector selection (a); refinement (b) from [1]

for the current frame to be encoded, the residual is smaller in many cases and thus, less bits have to be transmitted. Since H.264 / MPEG-4 AVC signals the index of the selected reference with different code word sizes, coding gain is dependant on the position of the DSME frame in the reference lists as shown in Section IV.

Experiments have shown that for low bit rates, it is better to use the DSME frame as decoded picture, without coding the remaining residual. Therefore, a hybrid approach is used where the encoder is deciding either to send the whole frame as a modified B frame as described above, or just to signal to use the DSME frame (Fig. 2 (b)). In that case, the frame is called pure DSME frame, since no additional information like prediction error or motion estimation parameters are sent to the decoder. The rate-distortion optimised decision implemented in the reference H.264 / MPEG-4 AVC encoder [4] is used to select the mode with minimum Lagrangian cost.

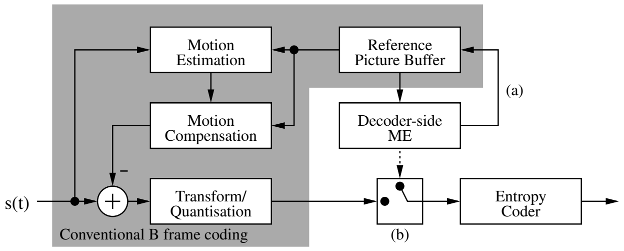

Fig. 2. DSME architecture with reference frame insertion (a) and pure DSME frame coding (b)

III. EXPERIMENTAL RESULTS

To evaluate the performance, the operational rate-distortion curves of the proposed method (DSME) and the H.264 / MPEG-4 AVC reference software JM [5] (H.264) are compared for several sequences (Fig. 3-7). The used test sequences have a spacial resolution of 352x288 pixel (CIF) with a frame rate of 30 fps and are coded with the GOP structure set to I-B-P-B-P. Additionally, a limit for decoder-side motion estimation is plotted assuming the distortion in the DSME frames is the same as in the neighbouring reference frames used for the interpolation without transmitting any data for the DSME frame.

Performance gains of up to 1.5 dB are achieved with decoder-side motion estimation for the Foreman sequence (Fig. 3) at very low bit rates. The predicted DSME frame is accurate enough to replace the B frame in the hybrid approach and bit rate is saved. However, the non-rigid motion impedes precise motion estimation needed for higher quality and the gain almost vanishes at high bit rates.

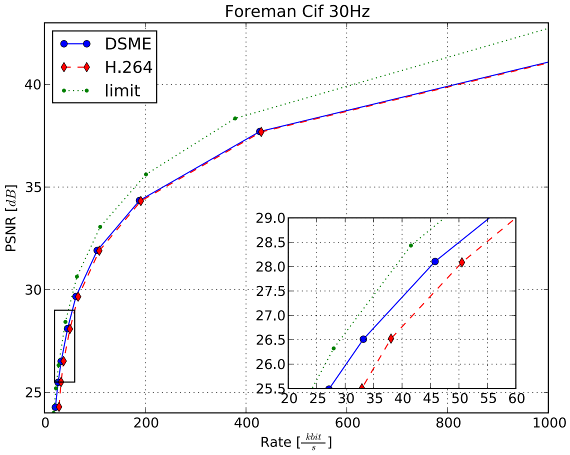

Fig. 3. RD performance for the Foreman CIF Sequence (150 frames)

The Flowergarden sequence (Fig. 4) with its smooth pan is very suitable for the DSME approach, since motion can be predicted accurately. The performance is 1.0 dB above the H.264 reference for lower rates and also yields up to 0.15 dB for high bit rates.

Although the City sequence (Fig. 5) exhibits similar characteristics as the Flowergarden sequence, the gain at low bit rates is with 0.5 dB well below the others. This is due to the H.264 / MPEG-4 AVC direct mode, which performs very well for this sequence. However, the gains of up to 0.2 dB for coding the sequence with higher quality are very promising.

Another interesting observation can be made with the Hall & Monitor sequence (Fig. 6). It is a surveillance scenario

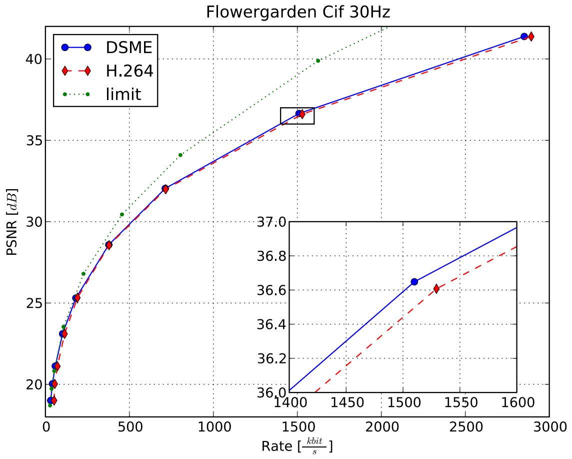

Fig. 4. RD performance for the Flowergarden CIF Sequence (125 frames)

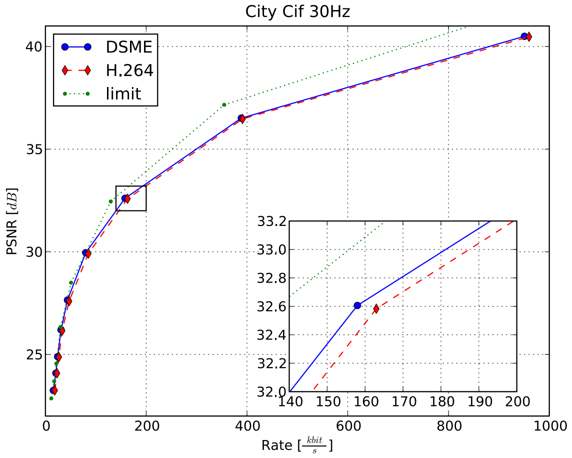

Fig. 5. RD performance for the City CIF Sequence (150 frames)

with a mounted camera where some people are walking through a corridor. Although only small parts of the frame are moving, and thus allowing the H.264 / MPEG-4 AVC coder to encode the static background in direct mode with low bit rates, DSME gains 1.7 dB. For high rates the amount of motion vectors becomes negligible and the gain vanishes.

Obviously, the gain of DSME decreases for sequences with deforming objects and where motion is non-linear. The results for the Soccer sequence (Fig. 7) are well below the other sequences. This is due to the non-linear non-rigid motion, which makes motion estimation unreliable.

IV. EVALUATION

As seen in Section III, the performance can be improved by estimating the motion at the decoder. The main gain at low bit rates is achieved by reducing the amount of bits sent to the decoder. However, a closer look at the rate-distortion curves reveals that the quality also increases. Since the algorithms for spatial and temporal direct mode [6] are simple and almost no prediction error is transmitted due to the coarse quantisation, the decoder-side motion estimation is able to outperform the conventional B frames in quality.

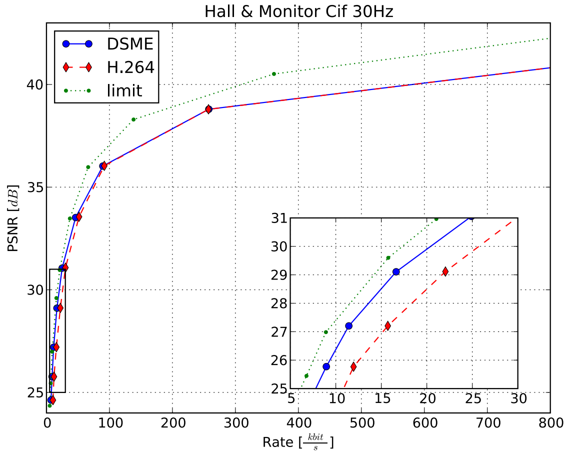

Fig. 6. RD performance for the Hall & Monitor CIF Sequence (165 frames)

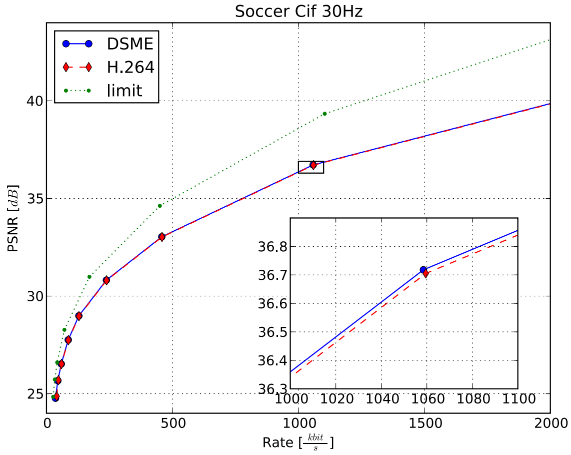

Fig. 7. RD performance for the Soccer CIF Sequence (150 frames)

Since, no prediction error is coded for pure DSME frames, the desired quality cannot be provided at higher bit rates. Thus, the encoder decides to transmit all frames as modified B frames in case of fine quantisation (lower quantisation parameter) as shown in Fig. 8.

Therefore, all gains achieved for high rates are due to the modified reference lists. As mentioned in Section II, the performance depends on the position of the DSME frame

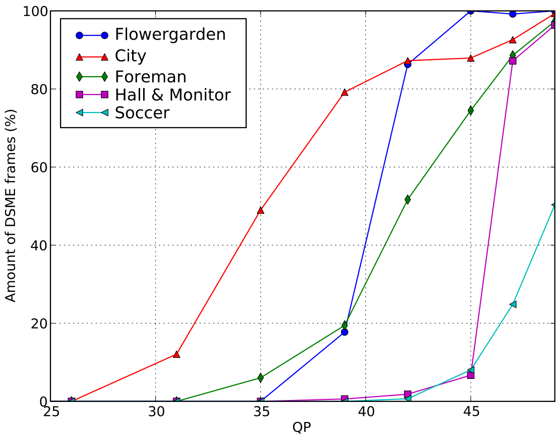

Fig. 8. Amount of B frames coded as pure DSME frame for various quantisation parameters QP

within the reference lists. In Fig. 9, the bit rate reduction compared to the JM reference software is shown for different positions of the DSME frame.

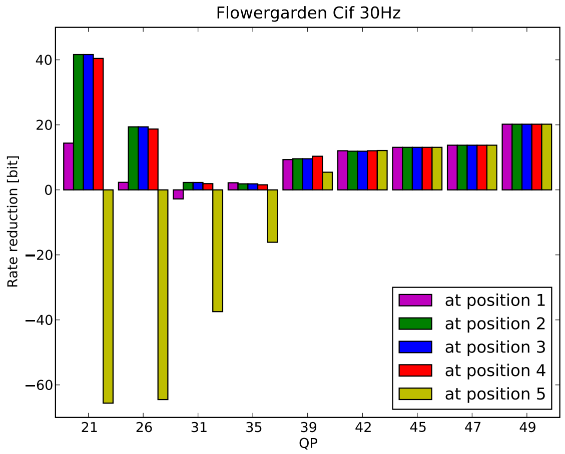

Fig. 9. Difference of the H.264 and DSME bit rate for various positions of the DSME frame within the reference list for Flowergarden

Using high quantisation parameters, the rate reduction is independent of the position since all frames are encoded as pure DSME frames (Fig. 8) and thus, the reference lists are not used. For higher qualities, the position becomes more important.

If the DSME frame is inserted in front of all other reference pictures at position 1, the bit rate savings are low. This is due the fact that the encoder often selects blocks of the temporal adjacent frame as reference. If it is moved to the second position in the list, the encoder needs more bits in signalling it to the decoder. If inserted at position 5, the DSME frame replaces the reference frame directly following the current frame. Since that frame is often used as reference, the DSME approach is worse than the H.264 reference.

Evaluations with several sequences have shown that inserting the DSME frame at the second position gives the best overall results.

Decoder-side motion estimation will become more valuable when incorporating more accurate motion estimation and compensation techniques like dense motion fields or object segmentation. This is possible because the motion information is not transmitted but is only used inside the decoder to improve the quality of the motion estimation.

V. CONCLUSIONS

In this paper, the benefits of decoder-side motion estimation are investigated and compared to the common H.264 / MPEG-4 AVC coding, in which the encoder performs all motion estimation tasks solely. The main advantage of DSME is that there is no need to transmit motion information, thus resulting in smaller bit rates. At low bit rates, even the transmission of the residue can be omitted.

Compared to conventional B frame coding, the current approach achieves coding gains of up to 1.7 dB or up to 27% bit rate reduction at lower quality. At higher bit rates, the gains of 0.2 dB or 6% bit rate reduction are still very promising. If the simple block-based motion estimation is replaced by more complex algorithms like dense motion field algorithms or motion estimation of segmented objects, the performance should increase further.

VI. REFERENCES

[1] J. Ascenso, C. Brites, and F. Pereira, “Improving frame interpolation with spatial motion smoothing for pixel domain distributed video coding,” in *5th EURASIP*, Slovak Republic, July 2005.

[2] S. Klomp, Y. Vatis, and J. Ostermann, “Side information interpolation with sub-pel motion compensation for wyner-ziv decoder,” in *Proceedings of the Int. Conf. on Signal Processing and Multimedia Applications*, Setúbal, Portugal, August 2006, pp. 178–182.

[3] L. Alparone, M. Barni, F. Bartolini, and V. Cappellini, “Adaptive weighted vector-median filters for motion fields smoothing,” in *IEEE ICASSP*, Georgia, USA, May 1996.

[4] G. J. Sullivan and T. Wiegand, “Rate-distortion optimization for video compression,” *IEEE Signal Processing Mag.*, vol. 15, no. 11, pp. 74–90, November 1998.

[5] “H.264 / MPEG-4 AVC reference software JM,” Website, available online at http://iphome.hhi.de/suehring/tml/.

[6] I. E. G. Richardson, *H.264 and MPEG-4 Video Compression*. West Sussex, England: John Wiley & Sons Ltd., 2003, ch. 6.5.1.4.