MOTION-COMPENSATING LONG-TERM MEMORY PREDICTION

Thomas Wiegand, Xiaozheng Zhang, and Bernd Girod

ABSTRACT

Motion-compensating long-term memory prediction extends the spatial displacement utilized in block-based hybrid video coding by a variable time delay permitting the use of more frames than the previously decoded one for motion compensation. The long-term memory covers the decoded frames of some seconds at encoder and decoder. We investigate the influence of memory size in our motion compensation scheme and analyze the trade-off between the bit-rates spent for motion compensated prediction and residual coding. Simulation results are obtained by integrating long-term memory prediction into an H.263 codec. PSNR improvements up to 2 dB for the *Foreman* sequence and 1.5 dB for the *Mother-Daughter* sequence are demonstrated in comparison to the TMN-2.0 H.263 coder.

1. INTRODUCTION

The rate distortion efficiency of motion-compensating prediction (MCP) can be improved using several techniques such as filtering and increased prediction accuracy [1] or sophisticated encoding of the spatial displacement field such as polynomial motion field representation [2]. Besides these methods, MCP can be further improved by multi-hypothesis estimation approaches including overlapped block motion compensation and bidirectional prediction [3].

In most cases, motion compensation (MC) is carried out by employing the immediately preceding frame which is available as the reconstructed frame at encoder and decoder. Long-term statistical dependencies in the coded video sequence are not exploited in existing international standards for improving coding efficiency of the video codec.

An example for extensive use of long-term dependencies in video sequences can be found in the negotiable H.263+ option called "Reference Picture Selection Mode" (RPS mode), as specified in Annex N of H.263 [4]. This mode permits a modified inter-picture prediction called "NEWPRED" [5] to stop temporal error propagation due to transmission errors. The RPS mode requires the storage of several decoded frames in a picture memory. The encoder may select one of the picture memories to suppress the temporal error propagation due to the inter-frame coding based on backward channel messages sent from the decoder to inform the encoder which part of which pictures have been correctly decoded at the decoder. The RPS mode is designed to suppress the temporal error propagation and not to improve coding efficiency of the video coder. However, we have observed that an architecture very similar to NEWPRED can lead to significant coding gains when omiting the overhead information contained in the syntax of the RPS mode [4].

Techniques to use several reference pictures in order to improve coding efficiency of video codecs are being analyzed within the MPEG-4 standardization group. These techniques are called "Sprites", "Global Motion Compensation" (GMC) [6], and "Short Term Frame Memory/Long Term Frame Memory" (STFM/LTFM) prediction [7]. Furthermore, "Background Memory" prediction techniques have been around for a while, e.g. see [8]. Common to all these techniques is that the video encoder can choose between the immediately preceding reconstructed picture and a second picture either generated by the Sprite, GMC, STFM/LTFM, or the Background Memory technique. An interesting extension to background memory prediction techniques was proposed in [9], wherein the image sequence is represented by layers.

In general, the rationale for techniques with MC using multiple reference pictures for improving coding efficiency is of heuristic nature. It has been recognized that the performance of video coding strongly depends on the bit allocation in hybrid video coding. Known in literature is a Lagrangian formulation to the problem of optimum bit allocation which we will also employ in this work.

2. MOTION-COMPENSATING LONG-TERM MEMORY PREDICTION

Our approach for exploiting long-term statistical dependencies is to extend the spatial displacement utilized in hybrid video coding by a variable time delay permitting the use of more frames than the previously decoded one for block-based motion compensation. With that, a long-term memory containing several seconds of the reconstructed image sequence can be used by the MCP. The frames inside the long-term memory, which is simultaneously built at the encoder's as well as the decoder's side, are addressed by a combination of the codes for the spatial displacement and the variable time delay. Hence, the transmission of the variable time delay requires additional bit-rate that has to be justified by improved MCP.

We view MCP as special case of entropy-constrained vector quantization (ECVQ) [10]. The image blocks to be encoded are "quantized" using individual code books that consist of image blocks of the same size in the previously decoded frames. A code book entry is addressed by the translational motion parameters which are entropy-coded. Consequently, the number of translational motion parameters, that is determined by the accuracy of the MCP and the motion search range, relates to the code book size in VQ. These considerations lead us directly to the insight that the effective long-term memory size is dependent on the rate-constraint imposed on the motion parameter codes as will be demonstrated by means of experimental results.

In order to determine the spatial displacement vector (d_x, d_y) and the time delay d_t, we conduct a standard block matching procedure. The criterion for the block motion search is the minimization of the Lagrangian cost function $J(\boldsymbol{d}) = D(\boldsymbol{d}) + \lambda R(\boldsymbol{d} - \boldsymbol{p})$, where $D(\boldsymbol{d})$ is a distortion measure for a given motion vector $\boldsymbol{d} = (d_x, d_y, d_t)$, such as the L_1 norm of the displaced frame difference, and $R(\boldsymbol{d} - \boldsymbol{p})$ is the bit-rate associated with a particular choice of the spatial displacement and time delay given its predictor $\boldsymbol{p} = (p_x, p_y, p_t)$. In this work, we set $p_t = 0$ for simplicity reasons. The Lagrange parameter λ imposes the rate-constraint.

In order to transmit the time delay d_t, we have generated a Huffman code table. For that, a set of 10 QCIF training sequences each with 10 seconds of video is encoded at 10 frames/s. While encoding, histograms are gathered on the time delay parameter to design the Huffman codes which are employed in the next encoding step. The loop is performed until convergence is reached, i.e., the changes in the overall Lagrangian costs become small. The spatial displacements (d_x, d_y) are transmitted using the H.263 MVD table [4].

The predictor for the spatial displacement (p_x, p_y) is computed using displacement vectors taken from a region of support (ROS). The ROS includes previously coded blocks that are close spatially and temporally. First, the time delay d_t for the current block is transmitted. Then, the spatial displacements assigned to blocks in the ROS are selected in case their time delay coincides with the time delay of the current block. The result is sorted in descending order of the correlations between the spatial displacement parameters of the current block and the blocks of the ROS. We measured these correlations off-line for the set of training sequences by conventional block matching with the immediately preceding frames.

The predictor is formed by taking the median from the first three of the sorted spatial displacement vectors. In case there are less than three displacement vectors available, only the first displacement vector is used as predictor if it exists. Otherwise we set the predictor $(p_x, p_y) = (0, 0)$. More details on prediction techniques for long-term memory MCP systems can be found in [11].

3. INTEGRATION INTO H.263

In order to evaluate the proposed technique the long-term memory MCP is integrated into an H.263 video codec. For that, the H.263 inter-prediction modes INTER, INTER-4V, and UNCODED[1] are extended to long-term memory MC. The INTER and UNCODED mode are assigned one code word representing the variable time delay for the entire macroblock. The INTER-4V utilizes four time parameters each associated to one of the four 8×8 motion vectors.

To run our H.263 as well as our long-term memory coder, we have implemented a modified encoding strategy as utilized by the TMN-2.0 coder, the test model for the H.263 standard.[2] Our encoding strategy differs for the motion estimation and the mode decision, where our scheme is motivated by rate-distortion theory.

In principle, the problem of optimum bit allocation to the motion vectors and the residual coding in any hybrid video coder is a non-separable problem requiring a high amount of computation. To circumvent this joint optimization, we split the problem into two parts: the motion estimation and the mode decision.

The motion estimation is performed as described above using the minimization of the Lagrangian cost function. For each frame the best motion vector using

[1] The UNCODED mode is an INTER mode for which the COD bit indicates copying the macroblock from the previous frame without residual coding [4].

[2] The TMN-2.0 codec is available via anonymous ftp to `bonde.nta.no`.

the L_1 norm distortion measure of the prediction error is found by full search on integer-pel positions followed by half-pel refinement. The integer-pel search is conducted over the range $[-16\ldots15] \times [-16\ldots15]$ pels. The impact of overlapped block motion compensation is neglected in the motion estimation.

Given the displacements for a particular mode that may be UNCODED, INTER or INTER-4V we are computing the overall rate distortion costs. The distortion is computed using the L_2 norm, and the rate is computed including the rates of macroblock headers, motion parameters, and DCT quantization coefficients. In case of long-term memory MCP, the motion estimation followed by the mode decision as described is conducted for each frame in the frame buffer.

Since there are now two Lagrangian cost functions to be minimized, we employ two different Lagrange multipliers: one for the motion search (λ_{motion}), the other one for the mode decision (λ_{mode}). Furthermore, the distortion measures differ because of complexity reasons. Hence, the selection of the Lagrange parameters remains rather difficult in our coder. In this work, we employ the heuristic $\lambda_{motion} = \sqrt{\lambda_{mode}}$, which appears to be sufficient. The parameter λ_{mode} itself is derived from the rate distortion curve that we computed using the TMN-2.0 H.263 coder.

4. SIMULATION RESULTS

In this section we demonstrate the performance of the proposed approach. Figs. 1 and 2 show the results obtained for the test sequences *Foreman* and *Mother-Daughter*, respectively. These sequences where not part of the training set. The Huffman codes for the time delay are trained for various memory sizes but only for one λ. The coder is run with constant quantizer when coding 100 frames at 10 frames/s. All results are generated from decoded bit streams.

The upper plots of Figs. 1 and 2 show the average PSNR from reconstructed frames produced by the TMN-2.0 codec, our rate distortion optimized H.263 codec and the long-term memory prediction codec vs. overall bit-rate. The size of the long-term memory is selected as 2, 5, 10, and 50 frames. The curve is generated by varying the Lagrange parameter and the DCT quantization parameter accordingly. Hence, the points marked with "+" in the plots relate to values computed from entire sequences. The long-term memory buffer is built up simultaneously at encoder and decoder by reconstructed frames. The results are obtained by measuring over frames 50...100, to avoid the effects at the beginning of the sequence.

The impact of rate-constrained encoding strategy is visible when comparing our H.263 codec with TMN-2.0. We noticed that the usage of the full search range $[-16\ldots15] \times [-16\ldots15]$ for the 8×8 displacement vectors provides most of the gain for our H.263 codec. The TMN-2.0 coder only permits the use of half-pel positions that surround the previously found 16×16 displacement vector, which is searched in the range $[-16\ldots15] \times [-16\ldots15]$. We have observed that extending the search range for the 8×8 displacement vectors leads to improved coding performance for our rate-constrained motion estimation, whereas for the TMN-2.0 we get worse results, since no rate-constraint is employed. This effect is even stronger, in case of long-term memory MCP.

When comparing long-term memory MCP to TMN-2.0, the PSNR gains achieved are about 2 dB for the *Foreman* sequence and 1.5 dB for *Mother-Daughter* when using a memory of 50 frames. For both sequences, a PSNR gain of 0.6 dB is due to our rate distortion optimization while the rest comes from the use of long-term memory prediction. These results demonstrate that utilizing the long-term memory we get an improved motion-compensating prediction scheme. The gains tend to vanish for very low bit-rates. This is in line with our interpretation of MCP as ECVQ, since as the effective code book size for ECVQ becomes smaller also the effective long-term memory size becomes smaller when the bit-rate is reduced.

The lower plots of Figs. 1 and 2 show the amount of bit-rate used for the motion parameters vs. overall bit-rate. As the long-term memory size increases, the amount of bit-rate for the motion parameters increases. But this increase is well compensated by the reduction in bit-rate for the residual coding. The decrease of the motion bit-rate for TMN-2.0 coder measured on the *Foreman* sequence results from the fact, that motion estimation is performed using the reconstructed frames (for TMN-2.0 as well as for our coder). As bit-rate decreases these reconstructed frames get noisier and since the regularization by the rate-constraint is missing for the TMN-2.0, the estimates for the motion data get noisier requiring a higher bit-rate.

5. CONCLUSIONS

By using motion-compensated long-term memory prediction we obtain a significantly improved video codec in terms of rate distortion performance. The gains are achieved at the expense of increased computational complexity and memory requirement. Main ingredient for the successful use of motion-compensated long-term memory prediction is the rate-constrained encoding strategy.

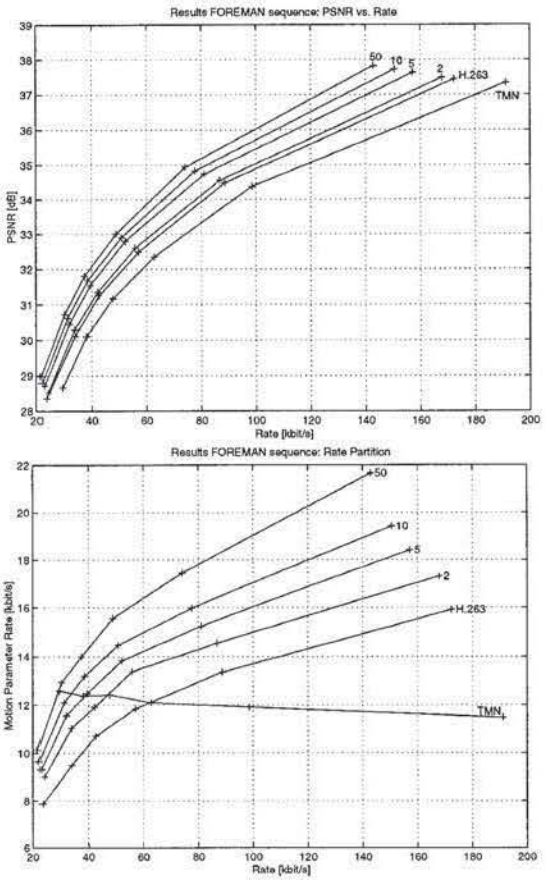

Figure 1: Results for the sequence *Foreman*.

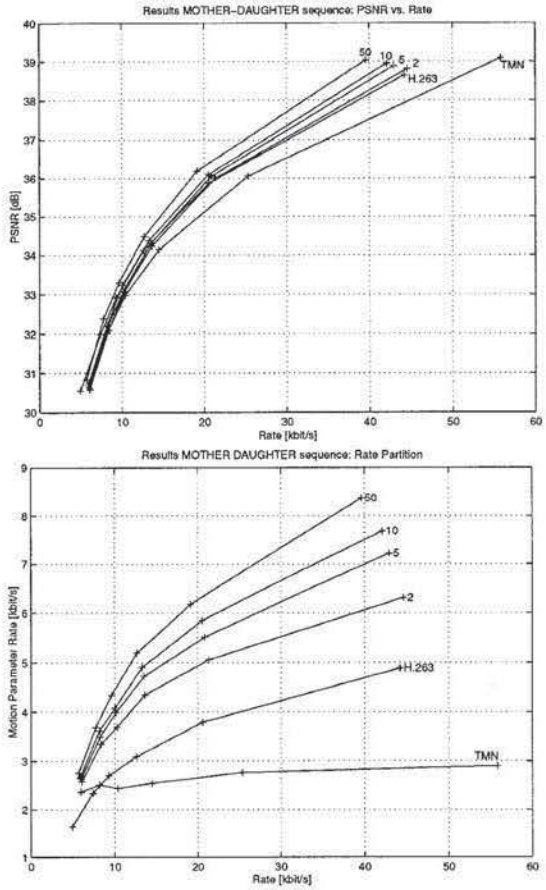

Figure 2: Results for the sequence *Mother-Daughter*.

6. ACKNOWLEDGEMENTS

Thanks to Eckehard Steinbach, Uwe Horn, Niko Färber, and Klaus Stuhlmüller for helpful discussions.

7. REFERENCES

[1] B. Girod, "Motion-Compensating Prediction with Fractional-Pel Accuracy", *TR-COM*, vol. 41, no. 4, pp. 604–612, Apr. 1993.

[2] ISO/IEC JTC1/SC29/WG11 MPEG96/M0904, "Nokia research center: Proposal for efficient coding", Submitted to Video Subgroup, July 1996.

[3] M. T. Orchard and G. J. Sullivan, "Overlapped Block Motion Compensation: An Estimation-Theoretic Approach", *TR-IP*, vol. 3, no. 5, pp. 693–699, Sept. 1994.

[4] ITU-T Recommendation H.263, "Video Coding for Low Bitrate Communication", Draft, Dec. 1995.

[5] ITU-T, SG15/WP15/1, LBC-95-033, Telenor R&D, "An Error Resilience Method Based on Back Channel Signaling and FEC", Jan. 1996.

[6] ISO/IEC JTC1/SC29/WG11 MPEG96/N1648, "Core Experiment on Sprites and GMC", Apr. 1997.

[7] ISO/IEC JTC1/SC29/WG11 MPEG96/M0654, "Core Experiment of Video Coding with Block-Partitioning and Adaptive Selection of Two Frame Memories (STFM / LTFM)", Dec. 1996.

[8] D. Hepper, "Efficiency Analysis and Application of Uncovered Background Prediction in a Low Bit Rate Image Coder", *TR-COM*, vol. 38, no. 9, pp. 1578–1584, Sept. 1990.

[9] J. Y. A. Wang and E. H. Adelson, "Representing Moving Images with Layers", *TR-IP*, vol. 3, no. 5, pp. 625–638, Sept. 1994.

[10] P. A. Chou, T. Lookabaugh, and R. M. Gray, "Entropy-Constrained Vector Quantization", *TR-ASSP*, vol. 37, no. 1, pp. 31–42, Jan. 1989.

[11] T. Wiegand, X. Zhang, and B. Girod, "Block-Based Hybrid Video Coding Using Motion-Compensated Long-Term Memory Prediction", in *Proc. PCS*, Berlin, Germany, Sept. 1997, To be published.