USING MULTIPLE GLOBAL MOTION MODELS FOR IMPROVED BLOCK-BASED VIDEO CODING

Eckehard Steinbach, Thomas Wiegand, and Bernd Girod

Telecommunications Laboratory
University of Erlangen-Nuremberg
Cauerstrasse 7/NT, D-91058 Erlangen, Germany
email: [steinb|wiegand|girod]@nt.e-technik.uni-erlangen.de

ABSTRACT

A novel motion representation and estimation scheme is presented that leads to improved coding efficiency of block-based video coders like, e.g., H.263. The proposed scheme is based on motion-compensated prediction from more than one reference frame. The reference pictures are warped or globally motion-compensated versions of the previous frame. Affine motion models are used as warping parameters approximating the motion vector field. Motion compensation is performed using standard block matching in the multiple reference frames buffer. The frame reference and the affine motion parameters are transmitted as side information. The approach is incorporated into an H.263-based video codec at minor syntax changes and embedded into a rate-constrained motion estimation and macroblock mode decision frame work. In contrast to conventional global motion compensation, where one motion model is transmitted, we show that multiple global motion models are of benefit in terms of coding efficiency. Significant coding gains in comparison to TMN-10, the test model of H.263+, are achieved that provide bit-rate savings between 20 % and 35 % for the various sequences tested.

1. INTRODUCTION

Independently moving objects in combination with camera motion and focal length change lead to a complicated motion vector field that has to be approximated efficiently for motion-compensated prediction. Block-based motion compensation with variable block size is often used as a compromise between prediction gain and side information over a wide range of motions. Motion compensation can be further improved by more sophisticated motion models that allow for more than the two degrees of freedom of the translational model to represent the motion field for large regions with a small number of coefficients.

Karczewicz et al. show in [2] that the use of polynomial motion models in combination with a coarse segmentation of the video frames can be beneficial in terms of coding efficiency. In [2] a block-granular segmentation with associated motion coefficients is used to efficiently approximate the motion field. The motion coefficients are estimated such that they lead to an optimum representation of the motion field inside the corresponding image segment.

The methodology in this paper differs from [2] in that the motion models are not associated with a single image segment. Although the motion coefficients are determined on a finite sub-area of the image, the model can be employed at any position inside the frame. An efficient realization of this *segmentation-free* motion representation paradigm is the use of multiple reference frames for prediction. The use of multiple reference frames requires only very minor syntax changes of state-of-the-art video coding standards.

In our approach, the motion in the sequence is captured by a finite set of affine motion models. The motion models are used to fill a multiple reference picture buffer with warped versions of the previous decoded frame. The encoder selects the reference frame that is most suited for prediction of the current frame on a block-by-block basis. This idea has first been introduced in [5]. The motion model estimation procedure in [5], however, is computationally expensive since the model coefficient estimation is performed on a precomputed displacement vector field using a robust estimator for the identification of different object or camera motions. This paper describes a new motion estimation scheme that is of much lower complexity than the approach in [5] and additionally leads to higher prediction gain.

A related motion representation scheme has independently been developed by Lauzon and Dubois in [4]. They address the motion representation problem by means of a dictionary of polynomial motion models with additional information that specifies which motion model has to be used at a particular image position.

This paper is organized as follows. We first present the proposed multi-frame video coding architecture. We then describe the estimation of the global motion models and the reference picture warping. Then, the incorporation of the multi-frame prediction into a rate-constraint motion estimation and mode decision framework is explained. Finally rate-distortion plots for four test sequences are given that illustrate the superior rate-distortion performance of the proposed method in comparison to TMN-10, the test model for H.263+ [1]. We also show the average number of selected motion models as a function of bit-rate and underline our claim that motion models estimated in one area of the image are also of use in other regions.

0-7803-5467-2/99/ $ 10.00 © 1999 IEEE

2. MULTIPLE REFERENCE FRAME MOTION-COMPENSATED PREDICTION

The architecture of the motion-compensated predictor as illustrated in Fig. 1 is based on the prediction of blocks in the current frame from more than one reference frame. The reference pictures are warped versions of the previous decoded frame using affine motion parameters that are estimated with reference to the current frame. The various affine motion parameter sets are transmitted as side information to the decoder. These reference frames are utilized by the encoder to perform translational block-based motion-compensated prediction.

When N motion parameter sets are transmitted, we utilize $M = N + 1$ reference frames that can be selected to independently predict each block by the encoder. Blocks in the reference frame buffer are addressed by a combination of the code words for the spatial displacement and a frame selection parameter that has to be transmitted to the decoder as well. Hence, the transmission of several motion parameter sets and the frame selection parameters potentially increases the bit-rate. But, if the improvements obtained by finding a good match in our extended motion search range make up for the extra bit-rate, we gain coding efficiency of our video codec.

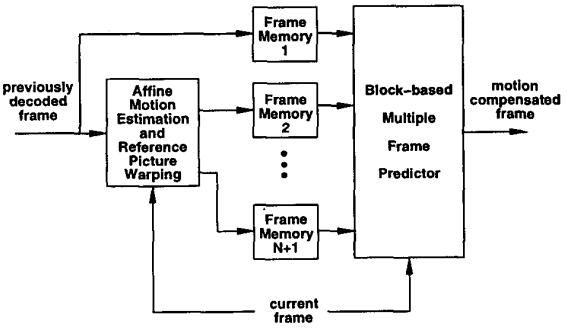

Figure 1: Motion-compensated predictor using M reference frames that are obtained by using the previous decoded frame and $N = M - 1$ warped frames given the corresponding affine motion parameter sets.

Hence, instead of just searching over the previous decoded frame the block-based motion estimator also searches positions in the warped frames in Fig. 1 and transmits the corresponding spatial displacement and frame selection parameter.

Relating our approach to region-based coding with affine motion models, e.g., see [2], we note that we also transmit various motion parameter sets and that the various "regions" associated with these motion parameter sets are indicated by the frame selection parameter. But, the "regions" in our scheme do not have to be connected. They are restricted to the granularity of the fixed or variable block-size segmentation of the block-based video codec. Furthermore, each block belonging to a "region" may have an individual extra displacement vector. This is beneficial if the motion exhibited in the scene cannot be compensated by a few affine motion models. In addition, if the video scene does not lend itself to a description by various affine motion models, the coder drops into its fall-back mode which is translational block-based motion compensation using the previous decoded frame only.

3. AFFINE MOTION MODEL

The motion model employed in our investigation is an orthonormalized version of the well known affine (6 parameter) transformation model where the relationship between the model coefficients and the displacement vector $(d_x(\boldsymbol{a}, x, y), d_y(\boldsymbol{a}, x, y))$ at an image point $\mathbf{x_c} = (x_c, y_c)$ in the current frame is given as [2]

$$\begin{aligned} d_x(\boldsymbol{a}, x_c, y_c) &= a_1 f_1(x_c, y_c) + a_2 f_2(x_c, y_c) + a_3 f_3(x_c, y_c) \\ d_y(\boldsymbol{a}, x_c, y_c) &= a_4 f_1(x_c, y_c) + a_5 f_2(x_c, y_c) + a_6 f_3(x_c, y_c) \end{aligned} \quad (1)$$

with $\boldsymbol{a} = (a_1, a_2, a_3, a_4, a_5, a_6)^T$ being the model coefficient vector, and $f_i(x, y)$ the orthonormal basis functions. The corresponding image point $\mathbf{x_p} = (x_p, y_p)$ in the previous frame is therefore given as

$$(x_p, y_p) = (x_c + d_x(\boldsymbol{a}, x_c, y_c), y_c + d_y(\boldsymbol{a}, x_c, y_c)). \quad (2)$$

Following the approach presented in [2], the orthonormalized basis functions $f_i(x, y)$ can be derived as

$$\begin{aligned} f_1(x, y) &= \alpha_{00}\beta_{00} \\ f_2(x, y) &= \alpha_{00}(\beta_{10} + \beta_{11} y) \\ f_3(x, y) &= \alpha_{00}(\alpha_{10} + \alpha_{11} x) \end{aligned} \quad (3)$$

with

$$\begin{aligned} \alpha_{00} &= \sqrt{\frac{1}{L_X + 1}}, \quad \alpha_{10} = \sqrt{\frac{3L_X}{(L_X + 1)(L_X + 2)}}, \\ \alpha_{11} &= -2\sqrt{\frac{3}{L_X(L_X + 1)(L_X + 2)}}. \end{aligned} \quad (4)$$

The β_{ij} are computed in the same manner replacing the image width L_X in (4) with the image height L_Y. We adopt this orthonormalization of the affine motion model in order to make the reconstructed motion model less sensitive to quantization of the model coefficients. A uniform scalar quantizer is used in our implementation.

4. MOTION PARAMETER ESTIMATION

Since we expect multiple independently moving objects in combination with camera motion and focal length change, the motion models have to be estimated accordingly. In order to avoid the need for an a priori segmentation of the scene, the image is subdivided into blocks of fixed size and a preselected number of motion models N is estimated from the previous decoded frame with respect to the current original frame. A typical number of motion models would be $N = 32$. For the 32 motion models, a QCIF image is subdivided into 32 blocks yielding blocks of size 22×36. For each measurement rectangle the motion coefficient estimation is performed in two steps.

- Initial half-pel accurate translational estimate using block matching.

- Affine refinement using an image intensity gradient based approach.

The first step is performed in order to robustly deal with large displacements. The translational estimate is used to motion-compensate the current frame towards the previous decoded frame as an initialization for the second step. Assuming constant brightness of the continuous differentiable image intensity $I(x,y,t)$ along the vector $(d_x, d_y, 1)$ allows to use the well-known optical flow constraint

$$\frac{\partial I(x,y,t)}{\partial x}d_x + \frac{\partial I(x,y,t)}{\partial y}d_y = -\frac{\partial I(x,y,t)}{\partial t}. \quad (5)$$

This constraint is combined with the affine displacement description in (1) leading to

$$\frac{\partial I}{\partial x}(a_1 f_1 + a_2 f_2 + a_3 f_3) + \frac{\partial I}{\partial y}(a_4 f_4 + a_5 f_5 + a_6 f_6) = -\frac{\partial I}{\partial t}. \quad (6)$$

Rearrangement leads to the following linear equation with the 6 unknowns $a_1 \ldots a_6$

$$\left(\tfrac{\partial I}{\partial x}f_1, \tfrac{\partial I}{\partial x}f_2, \tfrac{\partial I}{\partial x}f_3, \tfrac{\partial I}{\partial y}f_4, \tfrac{\partial I}{\partial y}f_5, \tfrac{\partial I}{\partial y}f_6\right) \begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \\ a_5 \\ a_6 \end{pmatrix} = \quad (7)$$

$$= -\tfrac{\partial I(x,y,t)}{\partial t}.$$

Setting up this equation at each pixel position inside the measurement rectangle leads to an over-determined set of linear equations that is solved in the least squares sense using, e.g., the *pseudo inverse* technique. Due to the inherent linearization in (5) it is beneficial to perform the parameter estimation in (7) multiple times in order to minimize the linearization error. Our experiments showed that two estimation steps give good results. Using more than two iterations leads to no further noticeable improvement. The repeated coefficient estimation in (7) requires the motion compensation of the current frame I_c towards the previous frame I_p. Since the coefficients of the affine motion models are estimated from the current frame I_c towards the previous frame I_p, motion compensation of the current frame towards the previous frame has to be performed using the inverse orthonormalized affine motion model. Rewriting (1) in matrix notation leads to

$$\begin{pmatrix} x_p \\ y_p \end{pmatrix} = \begin{pmatrix} x_c + d_x \\ y_c + d_y \end{pmatrix} = \mathbf{A}\begin{pmatrix} x_c \\ y_c \end{pmatrix} + \mathbf{b} \quad (8)$$

with

$$A = \begin{pmatrix} (1 + a_3\alpha_{00}\alpha_{11}) & a_2\alpha_{00}\beta_{11} \\ a_6\alpha_{00}\alpha_{11} & (1 + a_5\alpha_{00}\beta_{11}) \end{pmatrix} \quad (9)$$

and

$$b = \begin{pmatrix} a_1\alpha_{00}\beta_{00} + a_2\alpha_{00}\beta_{10} + a_3\alpha_{00}\alpha_{10} \\ a_4\alpha_{00}\beta_{00} + a_5\alpha_{00}\beta_{10} + a_6\alpha_{00}\alpha_{10} \end{pmatrix}. \quad (10)$$

Solving for the inverse displacement vector $\mathbf{d}_\mathbf{x}^\mathbf{I} = \mathbf{x_c} - \mathbf{x_p}$ leads to

$$\begin{pmatrix} d_x^I \\ d_y^I \end{pmatrix} = (\mathbf{A}^{-1} - \mathbf{E})\begin{pmatrix} x_p \\ y_p \end{pmatrix} - \mathbf{A}^{-1}\mathbf{b}, \quad (11)$$

with $\mathbf{E}$ being the unit matrix. After the first affine estimation step, the current frame is warped towards the previous frame using the inverse displacements in (11) and the refinement of the motion parameters in the next iteration step is computed between the warped current and the previous frame.

4.1. Reference Picture Warping

For each estimated motion model $\mathbf{a}_i$, $i = 1 \ldots N$, we warp the entire previous decoded frame towards the current frame to be coded. The reference picture warping is achieved using (1). Non-integer displacements are computed using cubic spline interpolation which turns out to be more efficient than bi-linear interpolation as the motion model becomes more complex [2]. We therefore obtain N new reference frames that can be used for block-based prediction of the current frame as illustrated in Fig. 1. Together with the previous decoded frame itself, $M = N + 1$ reference frames can be selected from the encoder for motion compensated prediction.

4.2. Rate-distortion Efficient Selection of Reference Frames

Having assembled the $M = N + 1$ frames in the multi-frame buffer, we want to determine what number of motion models is efficient in terms of rate-distortion performance [6],[7]. This is achieved by running a multi-frame H.263 coder in baseline mode [3]. The decision, which frames to prune from the multi-frame buffer is made using the following strategy:

1. Compute the Lagrangian cost $J = D + \lambda R$ ([6],[7]) when encoding a 16×16 block using a particular reference frame and store the cost values in an array. D represents the measured distortion and is computed as the *summed squared intensity difference* between the reconstructed and the original block. R represents the bit-rate required for the block.
2. For each 16×16 block select the reference frame that has the minimum Lagrangian cost value.
3. Sort the reference frames according to the frequency of their selection.
4. Starting with the least popular reference frame, test the utility of each reference frame. This is done by evaluating the rate-distortion improvement obtained by removing this reference frame. For those blocks that referenced the removed frame the best replacements in terms of Lagrangian costs among the more popular reference frames are selected. If no rate-distortion improvement is observed, the frame is kept in the reference buffer and the procedure is repeated for the next reference frame.

5. EXPERIMENTAL RESULTS

The multiple reference picture coding approach is integrated into an H.263-based codec at very minor syntax changes. The H.263 syntax is extended by the affine motion model parameters that are transmitted in the picture header. Further, the macroblock syntax is extended by a picture reference parameter enabling multi-frame motion compensation. Experiments are conducted using the QCIF sequences *Mobile & Calendar, Stefan, Foreman,* and *News*. A total of 100 frames of the sequences are encoded at 10 Hz varying the DCT quantizer over values $4, 5, 7, 10, 15, 25$. Bit-streams are generated that are decodeable producing the same PSNR values as at the encoder. The data of the first INTRA frame are excluded from the results.

5.1. R-D Plots

We first show rate-distortion curves for the proposed coder in comparison to the H.263 test model. The following abbreviations will be used to indicate the two cases:

- **H.263:** The result produced by the H.263 test model, TMN-10, using Annexes D,F,I,J, and T.
- **MRPC:** As **H.263** with additional multiple reference picture coding using at most 32 affine motion models.

Figs 2-5 show the rate-distortion curves for the four test sequences. It can be observed that considerable bit-rate savings at the same quality are achievable using the proposed algorithm. For the sequence *Mobile & Calendar* the bit-rate savings are up to 35 % at low bit-rates.

5.2. Coding Statistics

In Fig. 6, we depict the average number of motion models chosen by the MRPC codec. The number of estimated motion models is again 32. The number of motion models finally chosen from this set strongly depends on the bit-rate constraint as can be observed from Fig. 6. Based on these results we can conclude that more than one warped reference frame and with that several motion models are beneficial in terms of coding efficiency.

As described in Section 4 the motion models are estimated inside a small measurement window, but are used to warp entire reference frames. In order to illustrate that locally estimated motion parameter sets are also beneficial outside the estimation area we show in Table 1 the percentage of Macroblocks (MB) for which the encoder selects motion models that stem from a different (outside) or the same (inside) image region. The data are gathered for the 4 test sequences and all quantization runs. The results are split for the three different modes (UNCODED, INTER, INTER4V) that predict the current MB from the multiple reference frames. It can be seen that the largest number of motion models selected for the MBs are estimated in a different area of the image indicating their validity outside the measurement sub-region.

6. CONCLUSIONS

In this paper we demonstrate that the usage of multiple global motion models can be beneficial for the rate-distortion performance of block-based video coding. Significant coding gains in comparison to TMN-10, the test model of H.263+, are achieved that provide bit-rate savings between 20 % and 35 % for the various sequences tested. For the high motion sequence *Mobile & Calendar* bit-rate savings are 35 % corresponding to 2.2 dB gain in PSNR. We use orthonormalized affine motion models to warp additional reference frames that are employed for block-based prediction of the current frame. Since typically more than one motion cluster is present in the frame or the global motion cannot be described by one affine model we split the frame into N rectangular regions and estimate a motion model for this region using an image intensity gradient based approach. The encoder decides whether a particular model is beneficial in the rate-distortion sense at the desired bit-rate. Although the motion models are estimated on a finite sub-region of the

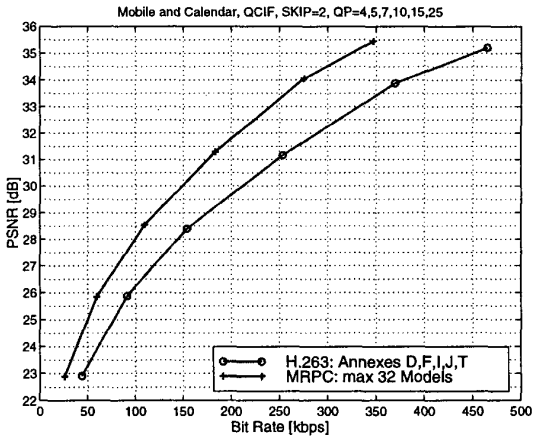

Figure 2: PSNR vs. overall bit-rate for the sequence *Mobile & Calendar* for the H.263 anchor and the proposed multiple reference picture coding scheme.

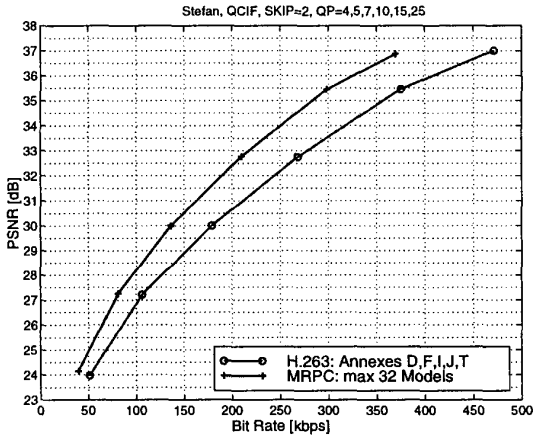

Figure 3: PSNR vs. overall bit-rate for the sequence *Stefan* for the H.263 anchor and the proposed multiple reference picture coding scheme.

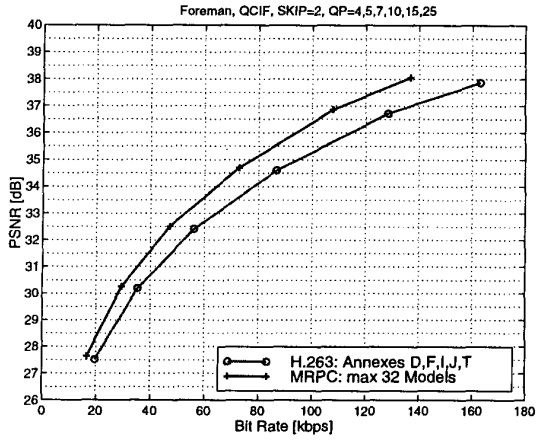

Figure 4: PSNR vs. overall bit-rate for the sequence *Foreman* for the H.263 anchor and the proposed multiple reference picture scheme.

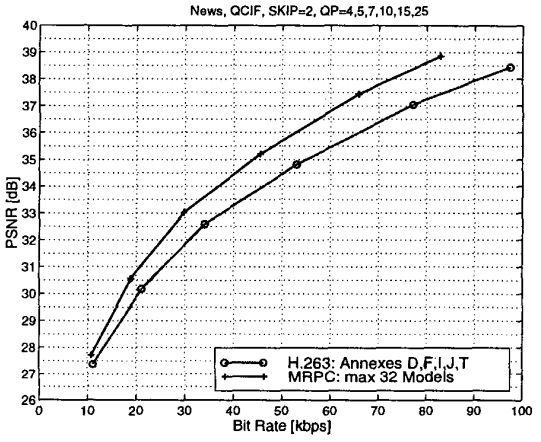

Figure 5: PSNR vs. overall bit-rate for the sequence *News* for the H.263 anchor and the proposed multiple reference picture scheme.

frame, they are often valid and therefore beneficial outside the estimation window. The scheme presented is this paper has been submitted as a proposal to ITU-T/SG16/Q15 for potential inclusion into H.263++ and H.26L.

Interestingly, additional coding efficiency can be observed when combining the approach presented in this paper with prediction from multiple past decoded frames [8]. While in this paper only the most recent decoded frame is employed for affine warping of additional reference frames, in [8] affine warping and motion compensation is conducted from several past decoded frames.

7. REFERENCES

[1] ITU-T Recommendation H.263 Version 2 (H.263+), "Video Coding for Low Bitrate Communication", Jan. 1998.

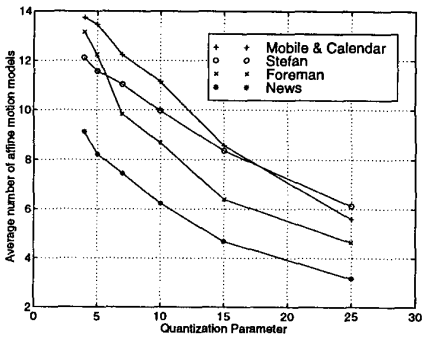

Figure 6: Average number of selected affine motion models as a function of the quantization parameter.

	mode percentage	inside	outside
UNCODED	46.00 %	21.3 %	78.7 %
INTER	34.94 %	19.7 %	80.3 %
INTER4V	19.06 %	11.3 %	88.7 %

Table 1: **First column:** Mode percentage for the 3 modes 'UNCODED', 'INTER', and 'INTER4V'. **Second column:** percentage of MBs that fall inside the local model estimation area. **Third column:** percentage of MBs that select a model which has been estimated outside the MB.

[2] M. Karczewicz, J. Niewęgłowski, and P. Haavisto, "Video Coding Using Motion Compensation with Polynomial Motion Vector Fields", *Signal Processing: Image Communication*, vol. 10, pp. 63-91, 1997.

[3] Thomas Wiegand, Xiaozheng Zhang, and Bernd Girod, "Long-Term Memory Motion-Compensated Prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 70-84, February 1999.

[4] D. Lauzon and E. Dubois, "Representation and Estimation of Motion using a Dictionary of Models," *Proc. International Conference on Acoustics, Speech and Signal Processing ICASSP '98*, pp. 2585-2588, 1998.

[5] T. Wiegand, E. Steinbach, Axel Stensrud, B. Girod, "Multiple Reference Picture Video Coding Using Polynomial Motion Models," *Visual Communication and Image Processing VCIP' 98*, pp. 134-145, Jan. 1998.

[6] G. J. Sullivan and R. L. Baker, "Rate-Distortion Optimized Motion Compensation for Video Compression Using Fixed or Variable Size Blocks", in *GLOBECOM'91*, 1991, pp. 85-90.

[7] B. Girod, "Rate-constrained Motion Estimation", *in Visual Communication and Image Processing VCIP '94*, A.K.Katsaggelos (ed.), Proc. SPIE vol. 2308, pp. 1026-1034, September 1994.

[8] T. Wiegand, E. Steinbach, B. Girod, "Long-Term Memory Prediction Using Affine Motion Compensation," *Proc. International Conference on Image Processing ICIP '99*, Kobe, Japan, October 1999.