28th Picture Coding Symposium, PCS2010, December 8-10, 2010, Nagoya, Japan

ADAPTIVE GLOBAL MOTION TEMPORAL PREDICTION FOR VIDEO CODING

Alexander Glantz, Andreas Krutz, and Thomas Sikora

Communication Systems Group
Technische Universität Berlin
Berlin, Germany

ABSTRACT

Depending on the content of a video sequence and the settings used for encoding it, the amount of bits spent for the transmission of motion vector information can be enormous and in some cases even take the largest fraction of the bit rate. This is not always necessary since often wide areas, i.e. background or large foreground regions, fit the same global motion. Additionally, a global motion model using sophisticated interpolation techniques can be a better representation of movement in these regions than a motion vector that has only quarter-pel accuracy. This is true especially if scaling, rotation or perspective transformation occur. This paper presents a novel prediction technique that is based on global motion compensation and temporal filtering of previously decoded pictures. The new approach is incorporated into an H.264/AVC reference software. The new encoder outperforms the reference by up to 14%.

Index Terms— H.264/AVC, video coding, global motion, temporal filtering, prediction

1. INTRODUCTION

The emergence of high definition (HD) or even ultra-high definition video content asks for new coding techniques that can cope with that amount of data. The existing state-of-the-art video coding standard H.264/AVC [1] aims at compressing high-quality video content at low bit rates. However, its performance is not sufficient for content that is at least five times larger than common resolutions (standard-definition television – SDTV – compared to 1080p full-HD video). Additionally, frame rates have increased from 25 up to 60 Hz.

Common hybrid video coding techniques remove temporal redundancy by estimating the motion of a macroblock to encode using the previously decoded pictures. Thus, only the residue of the prediction and a motion vector has to be transmitted for that block. Compared to spatial redundancy reduction (INTRA), this so-called INTER prediction significantly lowers the bit rate needed for transmission of the block. However, the amount of bits spent for the transmission of motion vector information can be enormous and in some cases even take the largest part of the bit rate. In the worst case, a full-HD video at 60 Hz sends more than 480,000 motion vectors a second – this does not account for sub-macroblocks. A common SDTV sequence at 25 Hz sends about 94% less motion vector information in the same time. Some research has been done on the reduction of that part of the bitstream. E.g., Kamp et al. have tried to reduce motion vectors and derive them at the decoder [2]. However, their approach is very different to the one presented herein and is therefore not further discussed.

The technique presented in this paper is based on the idea that the transmission of motion vector information is not always necessary since often wide areas, i.e. background or large foreground regions, fit the same global motion. Additionally, a global motion model using sophisticated interpolation techniques can be a better representation of movement in these regions than a set of motion vectors that have only quarter-pel accuracy. In [3], Wiegand et al. use higher-order global motion models to generate a set of reference frames for motion-compensated prediction (MCP). However, a temporal filtering of multiple references as in multihypothesis MCP [4] is not performed. This paper presents a novel prediction technique that is based on both, global motion compensation and temporal filtering. Following the authors' previous work in [5], a set of previously decoded pictures is transformed into the coordinate system of the current picture to encode. For that, global motion estimation (GME) is performed on the existing motion vector field using the Helmholtz Tradeoff Estimator [6]. The generated image stack is blended together using an adaptive pixel difference threshold method that minimizes the MSE between the filtered prediction signal and its original correspondence. The encoder then decides by means of rate-distortion (RD) optimization whether to use the new global motion temporal prediction (GMTP) or one of the common modes for a macroblock. Global motion parameters and pixel thresholds are sent as side information to the receiver to enable reconstruction.

This paper is organized as follows. Section 2 shows how a prediction signal for a picture to encode is built. In Section 3 the experimental setup is shown by incorporating the proposed technique into an existing H.264/AVC reference software. Section 4 describes the experimental results and the last section summarizes the paper.

978-1-4244-7133-1/10/$26.00 ©2010 IEEE

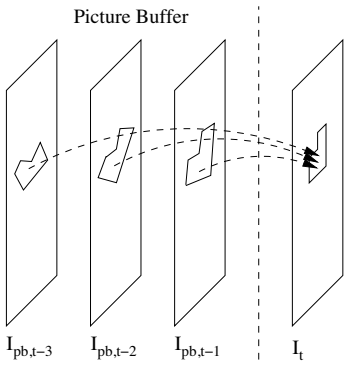


Fig. 1. Example for the generation of a prediction signal for the current picture I_t from previously decoded pictures. The pictures inside the picture buffer can be past and/or future pictures of the sequence depending on the GOP structure set in the encoder and are transformed into the coordinate system of the current picture.

2. PREDICTION SIGNAL GENERATION

Since many video sequences have been recorded with a moving camera, the background region of these sequences is not fixed. Theoretically, this means that the picture buffer does not contain a set of equal signals only differing by the superimposed noise, which could easily be removed by averaging, but of a set of displaced image signals containing noise. This problem can be solved using higher-order motion models that account for the displacement a camera performs. This motion can then be compensated so that the signals are spatially aligned. The aligned representations can then be considered as equal signals differing only by quantization noise and blocking artifacts, respectively.

In the following, the process of deriving a prediction signal for the current picture is described. First, the approach for global motion estimation is outlined. Second, the deduced global motion parameters are used for picture alignment to be able to temporally filter the generated image stack.

2.1. Global motion estimation

The global motion estimation algorithm used herein derives a homography H for a pair of pictures from a motion vector field available in common MPEG video streams using a Helmholtz Tradeoff Estimator (HTE). The HTE is a robust estimator with the ability of detecting up to 80% of outliers in a given dataset for an underlying model using subsets. For every randomly chosen subset out of a given motion vector field, a perspective transformation matrix H is calculated. In the next step, every motion vector position is transformed using H. The λ-th percentile of the distances between estimated and true motion vector destinations computes a standard deviation that is used to part the subset into inliers and outliers. Here, λ depends on the desired outlier tolerance. All inliers are used to calculate a final homography H_s by least squares for every subset. The homography belonging to the subset with the highest rating in terms of amount of inliers vs. inlier variance is taken as the final homography. For further information see [6].

2.2. Picture alignment and temporal filtering

Fig. 1 shows the exemplary transformation process of three previously decoded pictures – $I_{\text{pb},t-3}$ to $I_{\text{pb},t-1}$ – from the decoded picture buffer into the coordinate system of the current picture I_t. For that, the so-called long-term motion between picture $I_{\text{pb},t-i}$ and I_t has to be known. Therefore, the short-term motion parameters as derived in Section 2.1 are accumulated following the work by Smolic et al. [7].

The transformation process that uses spline interpolation creates an image stack of spatially aligned pictures. Assuming, the pictures that have been transformed contain blocking artifacts created during the encoding process, these can be reduced using temporal filtering. However, for large buffer sizes filtering of the complete image stack is not always the best choice, since moving foreground objects and misestimations negatively influence the filtered result. Therefore, a frame adaptive approach is used to generate an optimal result for every pixel in the prediction signal. Given a pixel difference threshold t and the difference $\Delta(i, x, y)$ between pixels of two successive pictures $I_i(x, y)$ and $I_{i-1}(x, y)$ in the image stack, it is assumed that a sudden increase along i in $\Delta(i, x, y)$ accounts for changes that shall not be used for filtering. Thus, for every pixel to generate a prediction signal for, an array of pixels is created that contains only those values from the image stack in increasing temporal distance from the current picture until $\Delta(i, x, y) > t$. The final prediction signal pixel is generated by taking the average of that array.

To generate the optimal prediction signal for a picture, threshold values from $T = \{1, \ldots, 8\}$ are tested. The threshold that produces the lowest mean squared error (MSE) between the filtered and the original picture is taken as the final prediction signal.

3. INCORPORATION INTO H.264/AVC

Fig. 2 shows the GMTP approach incorporated into a common hybrid video coding environment. At the encoder side, global motion estimation is performed using the motion vector fields from common INTER prediction. This generates global motion models that can be used as described in Section 2.2. The encoder adaptively generates a prediction signal using the decoded picture buffer and the global motion parameters. The encoder then decides on a macroblock-basis which mode is used by means of rate-distortion optimization. This is done by minimization of the Lagrangian cost function which is given by $min(J)$, where $J = D + \lambda R$. Here, D is the macroblock distortion for a given mode and R is the bit

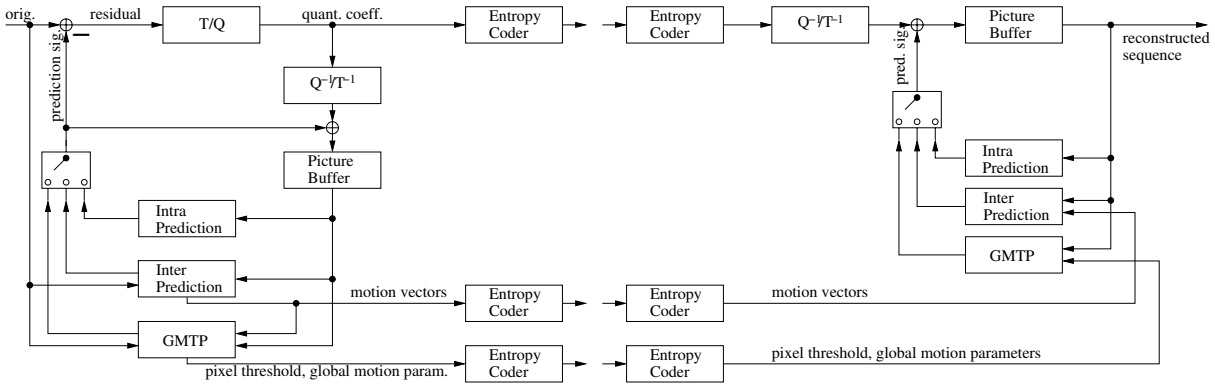


Fig. 2. GMTP within a hybrid video coding environment: GMTP is performed using global motion estimation based on existing motion vector fields besides common prediction modes. The encoder decides whether to use GMTP or common prediction on a macroblock-basis by rate-distortion optimization.

rate needed for its transmission. The new GMTP mode is included in that minimization problem. However, the Lagrange parameter λ has not been adjusted in the experimental setup, cf. Section 4. Possibly, the results could be further enhanced if λ was adjusted to the new extended set of modes.

As side information, the encoder sends a set of eight global motion parameters – i.e. the well-known perspective motion model – per frame without further optimization. Four byte floating point precision is used per parameter resulting in a total additional bit rate of $4 \times 8 \times 8 = 256$ bits per frame which can easily be surpassed by the amount saved in motion vector information. The pixel threshold values per frame are differentially encoded using signed Exp-Golomb coding.

The decoder receives global motion parameters and pixel thresholds besides the common bitstream and can therefore easily reconstruct the video sequence.

Sequence	Resolution	Frames/Hz	BD-rate	BD-PSNR
Basketball	1024×576	200/25	-1.19%	0.06 dB
BasketballDrive	1920×1080	500/50	-1.86%	0.07 dB
BBC-Pan-13	720×576	110/25	-14.87%	0.81 dB
BQSquare	416×240	600/60	-2.38%	0.08 dB
BQTerrace	1920×1080	600/60	-8.80%	0.21 dB
Desert	720×400	240/25	-5.29%	0.25 dB
ParkScene	1920×1080	240/24	-0.26%	0.01 dB
PartyScene	832×480	500/50	-3.97%	0.16 dB
Traffic	2560×1600	300/30	-3.13%	0.13 dB
Average			-4.64%	0.20 dB

Table 1. Test sequences and experimental evaluation in terms of BD-rate and BD-PSNR [8].

4. EXPERIMENTAL EVALUATION

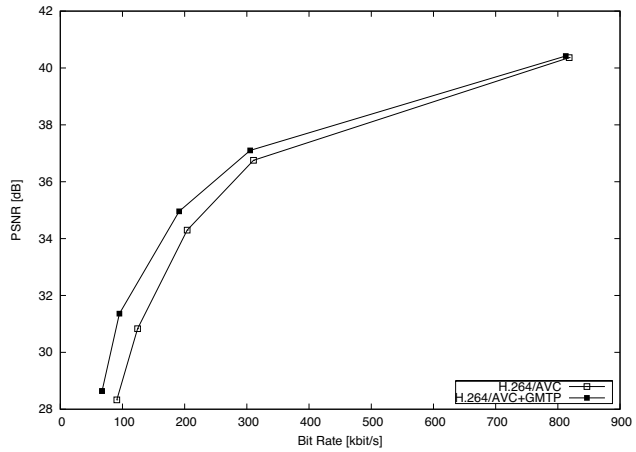
For experimental evaluation, the GMTP approach has been incorporated into the H.264/AVC reference software JM 17.0. As coding setting, a low delay setting has been used employing the H.264/AVC High Profile. This setting includes an IPPP GOP structure, CABAC entropy coding, RD optimization in high complexity mode, EPZS motion estimation with a search range of 128×128, 8×8 transforms enabled and QPPSlice $\in \{27, 33, 37, 43, 47\}$ (QPISlice $=$ QPPSlice $- 1$). The test sequences used are shown in Table 1 besides the experimental results computed following [8]. Fig. 3 shows two exemplary rate-distortion curves.

On average, bit rate savings of about 4.6% could be reached with a maximum saving of 14.87% for the *BBC-Pan-13* sequence. The amount of bits saved strongly depends on the content of the scene. In some sequences as in *BBC-Pan-13*, a large fraction of a picture shows background regions that can easily be represented by one single global motion model. Therefore, a huge number of macroblocks in *BBC-Pan-13* was predicted using the new GMTP mode.

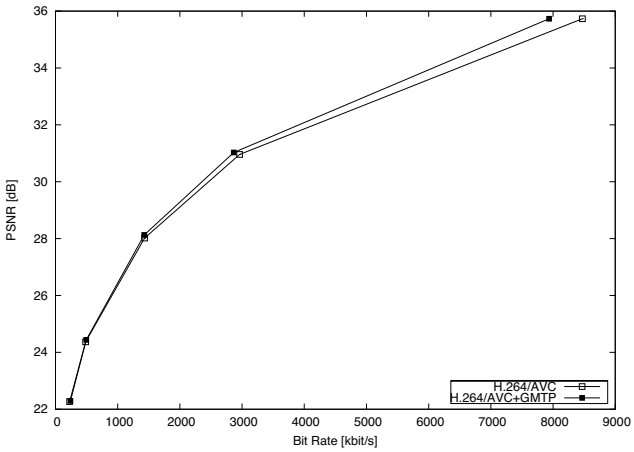
Fig. 4 shows exemplary statistical data for the *BBC-Pan-13* sequence. Depicted is the percentage of motion vector information in the complete bitstream for common H.264/AVC coding (black) and H.264/AVC coding using GMTP (gray). This corresponds to the expected results: Depending on the quantization parameter, motion vector information could be decreased up to 50%.

5. SUMMARY

We presented a novel prediction technique for hybrid video coding that is based on global motion compensation and temporal filtering. Global motion estimation is performed between consecutive pictures using the motion vector fields from the bitstream and a very robust Helmholtz estimator.



(a) Sequence *BBC-Pan-13*



(b) Sequence *PartyScene*

Fig. 3. Exemplary rate-distortion results. Depicted are curves for coding using common H.264/AVC as reference (H.264/AVC) and coding using H.264/AVC with GMTP mode enabled (H.264/AVC+GMTP).

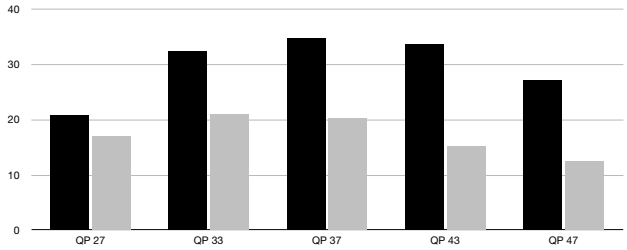


Fig. 4. Percentage of motion vector information in bitstream for the *BBC-Pan-13* sequence for various values of QP. Coding with H.264/AVC (black) compared to coding with H.264/AVC+GMTP (gray). The bits needed for transmission of global motion parameters are included.

This generates global motion parameters that enable the encoder to build an image stack for a picture, which then can be filtered using an adaptive pixel difference threshold method. The encoder decides on a macroblock-basis which coding mode is used and thereby reduces the amount of motion vector information sent to the receiver by up to 50%. The new coding method significantly outperforms existing state-of-the-art techniques.

6. REFERENCES

[1] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra, "Overview of the H.264/AVC video coding standard," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 13, no. 7, pp. 560–576, Jul 2003.

[2] S. Kamp, M. Evertz, and M. Wien, "Decoder side motion vector derivation for inter frame video coding," in *Proceedings of the 15th IEEE International Conference on Image Processing (ICIP 2008)*, Oct 2008, pp. 1120 –1123.

[3] T. Wiegand, E. Steinbach, and B. Girod, "Affine multipicture motion-compensated prediction," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 15, no. 2, pp. 197–209, Feb 2005.

[4] B. Girod, "Efficiency analysis of multihypothesis motion-compensated prediction for video coding," *Image Processing, IEEE Transactions on*, vol. 9, no. 2, pp. 173–183, Feb 2000.

[5] A. Glantz, A. Krutz, M. Haller, and T. Sikora, "Video Coding using Global Motion Temporal Filtering," in *Proceedings of the 16th IEEE International Conference on Image Processing (ICIP 2009)*, Cairo, Egypt, Nov 2009.

[6] M. Tok, A. Glantz, M.G. Arvanitidou, A. Krutz, and T. Sikora, "Compressed Domain Global Motion Estimation using the Helmholtz Tradeoff Estimator," in *Proceedings of the 17th IEEE International Conference on Image Processing (ICIP 2010) (to appear)*, Hong Kong, Sep 2010.

[7] A. Smolic, T. Sikora, and J.-R. Ohm, "Long-term global motion estimation and its application for sprite coding, content description, and segmentation," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 9, no. 8, pp. 1227–1242, Dec 1999.

[8] G. Bjontegaard, "Calculation of average PSNR differences between RD-curves," *ITU-T SG16/Q.6 VCEG document VCEG-M33*, Mar 2001.