Improved Video Coding Using Long-term Global Motion Compensation

Aljoscha Smolić, Yuriy Vatis, Heiko Schwarz, Peter Kauff, Ulrich Gölz, and Thomas Wiegand
Fraunhofer Institute for Telecommunications, Heinrich-Hertz-Institut
Einsteinufer 37, 10587 Berlin, Germany

ABSTRACT

We present a new approach to video coding that utilizes video analysis based on global motion features. For each encoded frame a global motion parameter set is estimated with respect to several previously transmitted frames. Using this global motion parameter set, a new motion compensation technique inspired by super-resolution mosaicing is applied. Rate-distortion optimization is used to identify macroblocks that are only affected by global motion. For such macroblocks no prediction error is transmitted. They are purely reconstructed using long-term global motion-compensated prediction. Our results indicate significant bit-rate savings compared to a state of the art H.264/AVC video codec at the same visual quality.

Keywords: Video coding, global motion compensation, global motion estimation, video mosaicing, super-resolution

1. INTRODUCTION

Global motion compensation (GMC) is an important tool for a variety of video processing applications including for instance segmentation and coding. The basic idea is that a part of the visible 2-D motion within video sequences is caused by camera operation (translation, rotation, zoom). A common approach is to model this *global motion* by a parametric 2-D model. In the past, efficient algorithms have been developed to estimate the global motion parameters between consecutive frames or over a longer period [5]. Finally, the estimated parameters are used to compensate the global motion.

In video coding one goal is to predict parts of pictures where the picture-to-picture displacement can be described by global motion models in order to increase the coding efficiency. Such systems have been studied in numerous publications and it has been shown that the coding efficiency can be increased for sequences that contain significant global motion. As a result GMC has been adopted for the MPEG-4 video coding standard [1]. The subjective quality at a given bit-rate can be improved even more with a related technology called sprite coding [1, 5], which employs video mosaics for coding. However, the usage of sprite coding is quite restricted due to inherent constraints of these approaches.

In [6] we have presented an algorithm that exploits the spatial alias to produce mosaics and video with a higher spatial resolution than the original video. It is a combination of video mosaicing and super-resolution techniques. The resulting mosaics and video are much sharper and contain more details compared to the original. Further we have shown in [7] how these super-resolution mosaics and video can be used for long-term global motion compensation (LT-GMC). Our approach combines elements from both, GMC and sprite prediction, resulting in prediction results that are significantly better compared to standard GMC.

In this paper we present the integration of LT-GMC into an H.264/AVC video codec [2]. LT-GMC is applied on a macroblock basis to detect those macroblocks where motion can be described by global motion models. The decision is taken based on rate-distortion criteria, i.e., the operational control of the reference H.264/AVC encoder is used as detection module for global motion regions. The corresponding macroblocks are encoded in a special way without transmission of the prediction error. This results in significant bit-rate savings without affecting the visual quality of the decoded pictures. This approach is very similar in spirit to the algorithm presented in [3] that applies texture analysis and synthesis to detect picture regions with specific texture properties and to encode them in a special way.

In the next Section, we describe the generation of super-resolution mosaics. A signal-theoretic motivation of this algorithm is presented in Section 3. Section 4 describes a new video prediction and coding tool based on a super-resolution mosaicing approach. Experimental results are presented in section 5.

2. SUPER-RESOLUTION MOSAICING

As shown in Fig. 1, video mosaicing means warping and blending all pictures of a considered video sequence into a common reference coordinate system. This can be either the local sample coordinate system of one of the pictures or a further transformed coordinate system, for instance a cylinder or a sphere. In any case, accurate warping parameters with respect to the common coordinate system have to be determined. Otherwise artifacts would become visible in the mosaic picture. Hence, one prerequisite for the generation of mosaics from multiple pictures is an accurate description and estimation of the global motion.

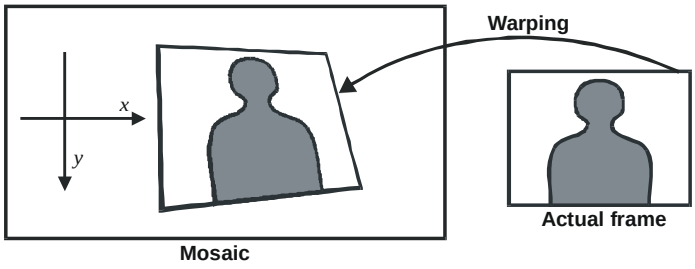

Fig. 1. Process of mosaicing: warping and blending all frames of a video sequence towards a common reference system, controlled by estimated global motion parameters

If the camera operation is restricted to zoom and rotation (i.e. no translation) and the pinhole camera model holds, the global motion can be exactly described by the perspective model:

$$x_1 = \frac{a_1 + a_2 x_0 + a_3 y_0}{1 + c_1 x_0 + c_2 y_0} \qquad y_1 = \frac{b_1 + b_2 x_0 + b_3 y_0}{1 + c_1 x_0 + c_2 y_0} . \tag{1}$$

These well known equations describe the transformation of a position in the reference picture (x_0, y_0) to a position in the actual picture (x_1, y_1). The transformation is controlled by a set of warping parameters with elements a_i (influencing x-coordinates), b_i (influencing y-coordinates), and c_i (influencing both coordinates), which can be combined in a warping parameter vector $\boldsymbol{\Xi}$. This description is inherently continuous with no restriction to any kind of sampling grid. In principle, the video mosaicing with super-resolution superposition presented in this paper works with any kind of projection (pure translation, affine, cylindrical, spherical). In this paper, we will consider the perspective model.
For a fast, reliable, and very accurate estimation of warping parameters, we utilize a combination of a differential and a feature-matching algorithm [5]. The differential algorithm relies on the iterative minimization of a postulated error measure with respect to the motion parameters:

$$\min_{\boldsymbol{\Xi}_t} E(\boldsymbol{\Xi}_t) = \min_{\boldsymbol{\Xi}_t} \sum_{\forall (x_0, y_0)} \left[M_{t-1}(x_0, y_0) - I_t(x_1, y_1) \right]^2 \tag{2}$$

where M is a mosaic, I denotes the intensity values of the actual picture and t denotes the time index. Equation (2) seeks for that set of motion parameters that best warps the actual picture onto the previous mosaic. This is the common problem formulation that is solved with a gradient-based, iterative algorithm. Stability and reliability is increased by incorporation of a robust M-estimator, in order to cope with differently moving foreground objects and other disturbances (lightning changes, shadows, noise, occlusions, etc.). In order to cope with large displacements between consecutive frames, we have incorporated an initial estimation of the translation using a feature-matching algorithm. The features are selected by a criterion that seeks for points that can be most reliably matched between pictures [5]. This initial stage provides an improved starting point (in accumulation to the previous estimated parameters $\boldsymbol{\Xi}_{t-1}$) for the differential algorithm (closer to the minimum), which ensures fast and reliable convergence of the iterative approach. The block diagram of the whole algorithm is shown in Fig. 2. The time-recursive structure provides an increased long-term stability, which is essential for accurate video mosaicing.

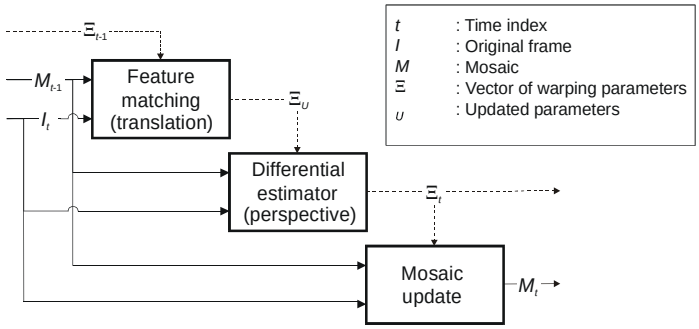

Fig. 2. Recursive and hierarchical algorithm for estimation of long-term warping parameters and video mosaic generation

After estimation of the warping parameters, the mosaic is constructed by warping the frames towards the common coordinate system. Since most of the visual information is visible in several pictures, a blending procedure has to be defined, for instance, averaging or median filtering, using only the most recent or first data. In the estimator loop in Fig. 2 we use the always the first available data, to be consistent with the recursive structure of the algorithm. Since the warping equations (1) represent continuous functions, traditional mosaicing requires the interpolation of fractional sample intensity values. Despite the usage of sophisticated interpolation filters, this leads to a low-pass effect which is further enhanced by the blending function. In addition, motion blur can have a bad impact on the mosaicing results. This is illustrated in Fig. 3. The left picture shows a detail from a conventionally generated mosaic. The middle picture shows a corresponding detail from an original frame. Note that the mosaic detail is wider, since it is warped. It looks blurred compared to the detail from the original picture. The right picture shows a corresponding detail from a super-resolution mosaic, which looks much sharper and contains more details (the white dots are explained below). The pictures have been scaled to the same size using the word processor's scaling utility.

Fig. 3. Left: detail from a conventionally generated mosaic, middle: detail from an original frame, right: detail from a super-resolution mosaic

As already indicated in Fig. 3, the visual quality of video mosaics can be significantly improved by application of super-resolution superposition. Fig. 4 illustrates the approach.
The samples of the video frames are transformed into a mosaic of double resolution in both directions, controlled by the estimated warping parameters, which have to be scaled accordingly. First, the frame at time instant t_0 is written into the mosaic of double resolution leaving empty sample positions in the mosaic at time instant t_0. Second, a half-sample diagonal shift from frame t_0 to frame t_1 is assumed. In this case, the corresponding samples in the mosaic are filled since they fall directly onto an integer-sample position in the mosaic. Note, that in these experiments we do not interpolate any mosaic sample. Only mosaic samples that are directly hit by warped video frame samples are updated accordingly, using the intensity (or color) value of the video frame sample. In practice we use a tolerance range of e.g. ± 0.2 sample units in the mosaic. This rule is employed to preserve the original sharpness of the video sequence by avoiding spatial interpolation. The theoretical motivation of our approach is given in section 3. The remaining empty sample positions are filled when processing more frames.

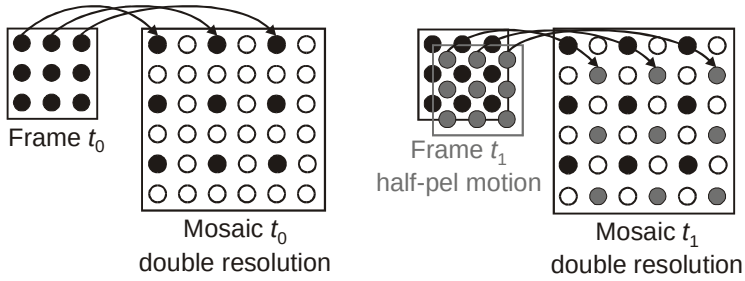

Fig. 4. Principle of video mosaicing with super-resolution superposition: transformation of video samples into mosaic of double resolution in both directions, without interpolation of intermediate intensity or color values

Fig. 5 illustrates an example for the successive filling of a super-resolution video mosaic. In the beginning (top-left picture), the mosaic is only sparsely filled since only one frame is written into the mosaic, leaving ¾ of the positions of the mosaic empty, which are displayed in white. As the camera moves (rotation and zoom), the intermediate positions are filled, only if they are directly hit by the sample accurate displacement vector that is given via the warping parameters (eq. (1)). The other pictures in Fig. 5 show intermediate stages of the process after 5, 15, and 49 frames of the video sequence.

The pattern of samples that have not been filled-in by the mosaicing depends on the global motion in the scene. Thus our approach would fail, if only full sample translational motion is contained in the video sequence. The number of frames needed to completely fill the area covered by the first frame also depends on the global motion in the scene. We have found experimentally, that 50 frames are clearly sufficient in most cases. Remaining holes in the super-resolution mosaic (see white dots in Fig. 3 right, Fig. 5 bottom-right) can be filled by standard interpolation methods, if required by the application.

Fig. 5. Successive filling of a super-resolution mosaic (704x480 sample), for sequence Stefan (30 Hz, 352x240 sample), results when processing 1, 5, 15, 49 frames are shown

It is obvious that the estimation of the warping parameters needs to be very accurate and robust, since a higher-resolution picture is generated and errors would become even more visible. Our estimation algorithm described above fulfills these requirements as the presented results prove. Note, that the warping parameter estimation works completely independent from the super-resolution superposition process. It only provides the estimated parameters. This means that the mosaic generated in the recursive estimator loop (Fig. 2) is a normal-resolution mosaic.

The super-resolution mosaic generation can be repeated for every time instant of the video sequence, using for instance past 50 frames. The result is a sequence of super-resolution mosaics, i.e. super-resolution video. This is illustrated in Fig. 6.

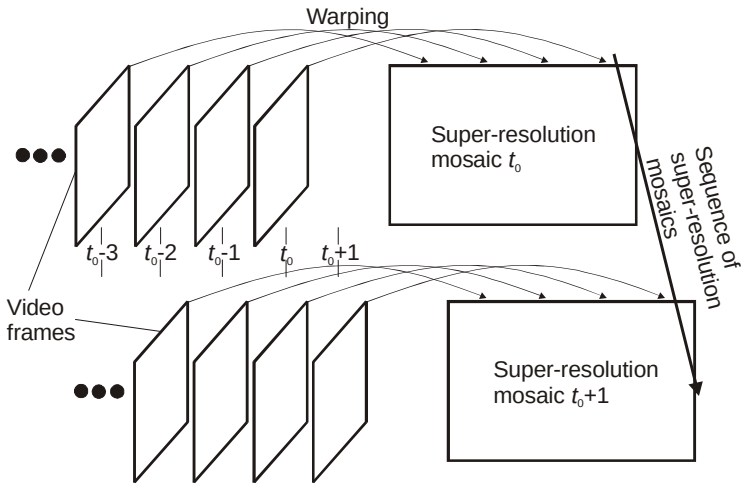

Fig. 6. Generation of a sequence of super-resolution mosaics (i.e. super-resolution video) from a video sequence

Fig. 7 shows on the left side details from 3 consecutive original frames of test sequence Stefan. The right side shows the corresponding details from the corresponding super-resolution mosaics. The pictures are scaled to same size using the word processor's utility. The details from the super-resolution video mosaics clearly show superior visual quality. They are much sharper, contain much more details, appear much less blocky, and most of all aliasing is highly reduced. Looking at original video in motion, annoying flicker artifacts can be noticed, that result from block structure artifacts, as shown on the left-hand side of Fig. 7, changing over time. In the sequence of super-resolution mosaics these block structure artifacts are drastically reduced, as shown on the right-hand side of Fig. 7, due to the spatial alias elimination capability of our algorithm. Therefore the resulting video sequence is flicker free.

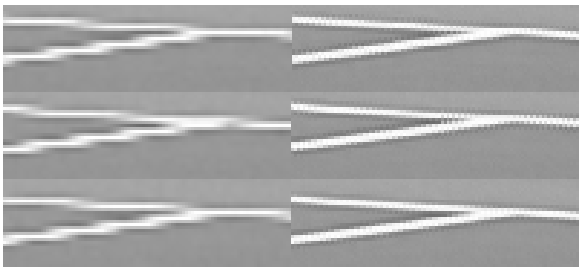

Fig. 7. Left: details from consecutive original frames of sequence Stefan, right: corresponding details from consecutive super-resolution mosaics

3. THEORETICAL MOTIVATION

As mentioned above, the strength of our approach relies on the elimination of the spatial alias of the video sequence by global motion-compensated super-resolution superposition. This section provides a theoretical analysis of the algorithm described technically in the previous section. The derivation is provided for the case of a pure translational displacement with constant velocity. For more complex types of motion the resulting equations become very difficult to be handled analytically. Moreover, to further simplify the notation we present the derivation for a 2-D signal. Note that the extension to 3-D signals is straightforward.

Consider a continuous 2-D signal $g(x,t)$ that is derived from a 1-D signal $s(x\text{-}vt)$ that is displaced in the spatial dimension x with constant velocity v, as illustrated in Fig. 8:

$$g(x,t) = s(x - vt) \tag{3}$$

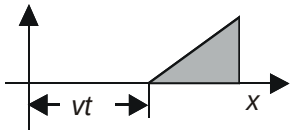

Fig. 8 The 1-D signal displaced in *x*-direction with constant velocity *v*

The spectra of the analog signal after spatial and temporal Fourier-transform (f_x, f_t denote the spatial and temporal frequency respectively) is given by (for details on signal theory refer for instance to [4])

$$G(f_x, f_t) = S(f_x) \cdot \delta(f_t + f_x v) \tag{4}$$

Herein δ denotes the Dirac function. Due to the masking properties of the Dirac function, the region of support is given by $f_t = -v \cdot f_x$. Fig. 9 illustrates the spectra in the space-time frequency domain for different velocities $v_2 > v_1 > v_0$.

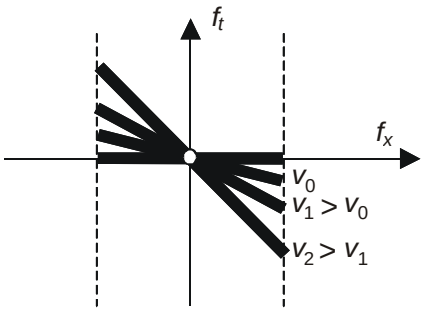

Fig. 9. Illustration of the space-time spectra for different velocities

The sampling of a continuous signal with sampling distance *X* can be expressed as multiplication with the Dirac function series

$$\delta_X(x) = \sum_{k=-\infty}^{\infty} \delta(x - kX) \tag{5}$$

and with that, we get in the 2-D case

$$\begin{aligned} g_d(x) &= s(x - vt) \cdot \delta_X(x) \cdot \delta_T(t) \\ &= \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} s(kX - vnT) \cdot \delta(x - kX) \cdot \delta(t - nT) \end{aligned} \tag{6}$$

The subscript d indicates a discrete signal. The spatial and temporal summation index is denoted by k and n, respectively. The spectrum G_d of the sampled signal is given as the convolution of the base spectrum of the continuous signal G with a frequency domain Dirac impulse series $\delta_{f_X}(f_x) \cdot \delta_{f_T}(f_t)$ with spatial and temporal sampling frequency $f_X = 1/X$, $f_T = 1/T$

$$G_d(f_x, f_t) = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} S(f_x - kf_X) \cdot \delta[f_t - nf_T + v \cdot (f_x - kf_X)] \tag{7}$$

Equation (7) describes the periodic repetition of the base spectrum with the corresponding sampling frequencies in the spatio-temporal frequency domain. Fig. 10 illustrates the periodic spectrum under investigation for the velocity of

$$v = \frac{1}{2}\text{pel/frame}, \text{ i.e. } \frac{f_t}{f_T} = -\frac{1}{2}\frac{f_x}{f_X}.$$

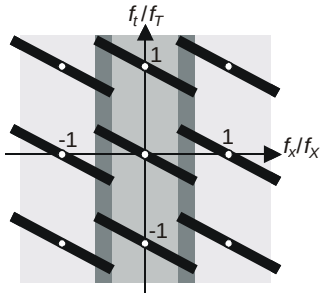

Fig. 10. Spatio-temporal spectrum of the sampled function with v = ½ pel/frame. As can be seen, the region of support includes spatio-temporal alias components (dark gray area) and therefore, alias free reconstruction is not possible

The algorithm described in Section 2 can be regarded as a motion-compensated superposition of the sampled signal $g(x,t) = s(x - vt)$

- at location x at time t
- and location x-vT at time t-T

More precisely, the algorithm produces the following signal

$$g_v(x,t) = \left[s(x-vt)\cdot\delta_X(x)\cdot\delta_T(t)\right] * \left[\delta(x)\cdot\delta(t) + \delta(x-vT)\cdot\delta(t-T)\right] \quad (8)$$

Convolution with the first term $\delta(x)\cdot\delta(t)$ causes an identity mapping, the second term $\delta(x-vT)\cdot\delta(t-T)$ clips the displaced signal at time t-T. For sake of simplicity further calculations utilize $v = 0.5$ pel/frame while the extension to other velocities is straight forward. To evaluate Equation (8), we introduce in a first step spatial oversampling by a factor of 2, i.e., $(X \rightarrow X/2)$ at time t and t-T followed by introduction of zeros, i.e. every second sample is set to zero. The single spectra are compressed in the spatial frequency domain by a factor of 2 and the intensity is scaled by a factor of ½. In other words the spectrum is a spatially scaled version of Fig. 10 with $f_{X/2} = 2 \quad f_X$.

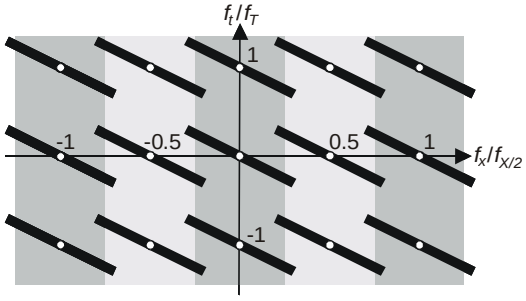

Fig. 11. Spatio-temporal spectrum after oversampling and introduction of zeros

The spectrum of the superimposed signal according to Equation (8) is given by

$$G_v(f_x, f_x) = \frac{1}{2}G_d(f_x, f_x)\cdot H(f_x, f_x) \quad (9)$$

This is the scaled product of the spectrum of the original sampled signal and a transfer function H. The impulse response in the time domain is given by

$$h(x,t) = \left[\delta(x)\cdot\delta(t) + \delta(x - vT)\cdot\delta(t - T)\right] \tag{10}$$

and its Fourier-transform by

$$H(f_x, f_t) = 1 + e^{-j2\pi f_x vT}\cdot e^{-j2\pi f_t T} \tag{11}$$

We now introduce the normalized velocity $\tilde{v}$

$$v = \tilde{v}\,\frac{\text{pixel}}{\text{frame}} = \tilde{v}\,\frac{X}{T} = \tilde{v}\,\frac{f_T}{f_X} \tag{12}$$

The Dirac-function in equation (7) is only non-zero, if the argument is zero. With that and Equation (12) we can derive the following relationship:

$$\frac{f_t}{f_T} = n + \tilde{v}k - \tilde{v}\,\frac{f_x}{f_X} \tag{13}$$

If we use equation (12) and (13) to reformulate the transfer function we get:

$$H(f_x, f_t) = 1 + e^{-j2\pi(n+\tilde{v}k)} = 1 + \cos\left[2\pi(n+\tilde{v}k)\right] - j\sin\left[2\pi(n+\tilde{v}k)\right] \tag{14}$$

In the description of the algorithm in Section 2 we have shown that the superposition is only carried out if the normalized velocity satisfies $\tilde{v} = m + 1/2$, where m is a positive or negative integer value. Then we finally get:

$$H(f_x, f_t) = \begin{cases} 2, & \text{if } k \text{ even} \\ 0, & \text{if } k \text{ odd} \end{cases} \tag{15}$$

This means that the repetitions of the base spectrum are eliminated, if k is odd and that they are retained if k is even. The resulting spectrum for our example is shown in Fig. 12. Obviously it is now possible to reconstruct the base spectrum from the oversampled and superposed signal without spatial alias artifacts. This effect of elimination of certain spectra, explained here for the simple case of a 1-D signal that undergoes translational motion, is the theoretical basis of our super-resolution video mosaicing algorithm.

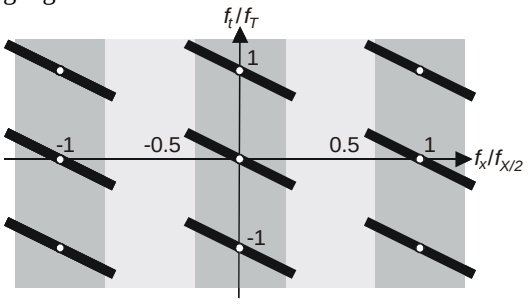

Fig. 12. Spatio-temporal spectrum after global-motion-compensated super-resolution superposition , v = ½ pel/frame

4. INTEGRATION INTO H.264/AVC

In this section we describe a new video prediction and coding method based on a super-resolution mosaicing approach and it's integration into an H.264/AVC codec. The approach is macroblock-based as necessary for integration into H.264/AVC and uses the multi-frame prediction tool. In [7] we have demonstrated that such super-resolution mosaics can significantly improve the prediction performance in terms of the residual error compared to standard GMC.

4.1 Macroblock-based LT-GMC

The multi-frame prediction tool [8] of H.264/AVC typically provides 5 decoded reference pictures to be used for motion compensation at the decoder. As a first step the global motion of these 5 pictures is estimated with respect to the actual picture to be encoded using the robust recursive algorithm described before [5]. The composition of the super-resolution prediction is done as follows. Each sample of a predicted macroblock is warped towards all reference frames using the respective global motion parameters as illustrated in Fig. 13. In general the transformed positions do not fall onto integer positions in the sample raster of the reference frames, but somewhere in between 4 neighbouring samples. A distance can be calculated to each of the neighbouring samples. The minimum distance is recorded. Such a minimum distance d_i is calculated for all 5 reference frames and finally the minimum of d_i is determined. As a result we get the closest distance of a warped sample to be predicted to a sample of the reference frames, which can be within any of the 5 available reference frames. The intensity or colour value used for prediction of the current macroblock samples is calculated by bilinear interpolation within this "closest" reference frame.

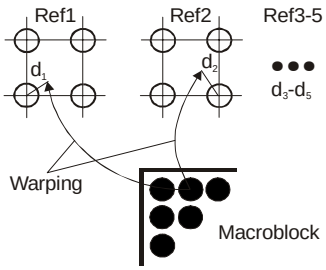

Fig. 13. Generation of prediction for a macroblock

4.2 Mode Decision and Encoding

The H.264/AVC codec supports a variety of different macroblock coding modes. In principle each encoder implementation has the freedom to decide about the encoding modes using any kind of proprietary operational control. The official reference encoder implementation applies a Lagrangian coder control as described in [9] that optimizes encoding decisions with respect to their rate-distortion efficiency.
Given the quantization parameter QP for a macroblock k as well as the motion vectors and reference indices for all motion-compensated macroblock modes of this macroblock, the mode decision for the macroblock k proceeds by minimizing the Lagrangian functional

$$J_k(p \mid QP,\lambda) = D_k(p \mid QP) + \lambda \cdot R_k(p \mid QP), \tag{16}$$

where the macroblock mode p is varied over the set of possible macroblock modes given by the slice type. The distortion $D_k(p \mid QP)$ is measured as the sum of squared differences (SSD) between the original samples of the macroblock k and their reconstructions that are obtained by using the coding mode p (with associated motion parameters) and the quantization parameter QP. The rate term $R_k(p \mid QP)$ specifies the number of bits associated with choosing the coding mode p and the quantization parameter QP including the bits for encoding the macroblock header, the motion parameters, and all transform coefficient blocks. The Lagrangian multiplier λ is determined according to

$$\lambda = 0.85 \cdot 2^{\wedge}(QP / 3 - 4). \tag{17}$$

We have introduced a new macroblock mode that utilizes LT-GMC as described in the previous section. This new mode is included in the mode decision process, where the distortion and rate terms are calculated as if the LT-GMC mode

comprises the transmission of a residual signal. However, for the encoding of LT-GMC macroblocks we introduce a heuristic assumption that is justified by intuition and the results. If LT-GMC results in the minimum of the cost functions, i.e. it is the best coding mode, then it is very likely that the corresponding macroblock is affected by global motion and even aliasing may occur. Then we assume that it can be reconstructed without transmission of the prediction error and of course without transmission of local motion vectors. This results in a reduction of the bit-rate without affecting the visual quality although the PSNR performance drops. This algorithm is in the spirit of sprite coding schemes that also perform best if the content is purely reconstructed from previously transmitted content without transmission of texture updates [5]. In a sense the operational control of H.264/AVC is used as recognition tool for global motion regions and performs an automatic segmentation of the content.

Since the algorithm is fully integrated into the recursive encoding coding loop, errors cannot accumulate. If there are inaccuracies, LT-GMC will most likely not be the best coding mode and the errors will be corrected by transmission of prediction errors using other modes. Therefore our approach incorporates the advantages of standard GMC over sprite coding, i.e. being fully automatic, integrated and suitable for any type of content without relying on a priori knowledge (e.g. segmentation) of the content.

Our algorithm saves bits for transform coefficients and motion vectors, but on the other hand it introduces overhead for transmission of the global motion parameters. Having 40 parameters per encoded frame results e.g. in 960 parameters per second for a 25 Hz video sequence. Fortunately these parameters are highly redundant, and we have developed an efficient compression scheme. We first normalize the parameters to the picture dimensions. Knowing that the global motion cannot change discontinuously over time, we can predict the parameters from previously transmitted ones. The remaining residuals are γ-distributed and we apply scalar quantization and encode them using an exp-Golomb code. The scheme has been integrated into the H.264/AVC reference software (version JM2.1) and so far only for P frame encoding is implemented.

5. EXPERIMENTAL RESULTS

We have tested the algorithm with a variety of test sequences, where we used P frame only coding. Fig. 14 shows decoded pictures of the proposed and the reference codec for test sequence City (high definition 1280x720 samples, 30 frames). No visual difference was observed by us in any of the decoded frames. The reference picture was encoded with 445 kbit/s resulting in a PSNR of 40.3 dB. With LT-GMC we get 355 kbit and a PSNR of 39.3 dB. This means that we get a loss in PSNR of 1.0 dB which is not visible but expected since we do not perform waveform coding. On the other hand we get a bit-rate saving of 20.2% at the same visual quality.

Fig. 14. Coding results for City, top: baseline H.264/AVC, bottom: with LT-GMC

The macroblock assignment mask is shown in Fig. 15. In total 1028 LT-GMC macroblocks have been assigned which corresponds to 28.5% of the total number. The sequence is captured by a camera in a helicopter flying over New York City. The skyscraper in the front is relatively close to the camera. Due to the translational motion of the camera it does not coincide with the global motion (motion parallax) and can therefore be regarded as a differently moving foreground object. Nevertheless the global motion is estimated very accurately since a lot of LT-GMC macroblocks are assigned in the background. But the LT-GMC mode is not assigned to any of the foreground macroblocks. This example nicely illustrates the properties of our algorithm and validates accuracy and efficiency.

Fig. 15. Assigned macroblock modes for the example in Fig. 14 bottom, grey: LT-GMC, black: other modes

Fig. 16 shows the relative bit-rate savings when assuming that the coding at equal QP values results in the same visual quality. There is a significant gain for all bit-rates. The largest gain of up to 26.5 % is achieved for lowest bit-rates. Then the gain drops to 9.5 % for low bit-rates. For medium and high bit-rates the gain is relatively constant between 14 % and 16 %. 2 different effects influence the slope of the curve. At high and medium bit-rates the total savings mainly come from savings on texture updates. Savings on motion vectors are less important at these bit-rates. The relative savings on texture updates decrease to low bit-rates. At very low bit-rates the savings on motion vectors become important resulting in large total savings.

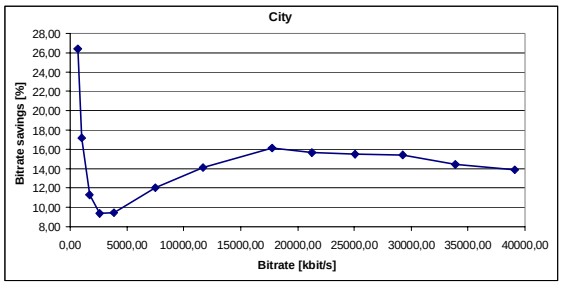

Fig. 16. Relative bit-rate savings of LT-GMC codec compared H.264/AVC codec at the same visual quality over bit-rate for City

These gains were achieved for high-resolution formats. The gain decreases with the resolution of the source video. Fig. 17 shows the bit-rate savings for Mobile & Calendar (SIF 352x240 samples, 300 frames). The gain decreases from about 5 % at highest bit-rates to negative values at lowest bit-rates. In this case the overhead for the global motion parameters becomes significant at low bit-rates and therefore a loss rather than a gain is obtained compared to the result in Fig. 16. This overhead is quite constant over the bit-rate and over the resolution of the source video. For high definition resolution it can be neglected, but for SIF resolution it becomes important at low overall bit-rates.

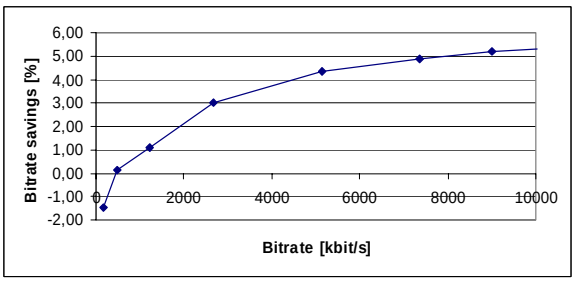

Fig. 17. Relative bit-rate savings of LT-GMC codec compared H.264/AVC codec at the same visual quality over bit-rate for Mobile & Calendar

An example mode assignment mask for Mobile & Calendar is show in Fig. 18. In total 100 macroblocks are coded with LT-GMC. The average over the sequence is about 20-30 for Mobile & Calendar. Apparently the relative size of the macroblocks is larger compared to the high-resolution examples, which results in a lower probability for assignment of the LT-GMC mode.

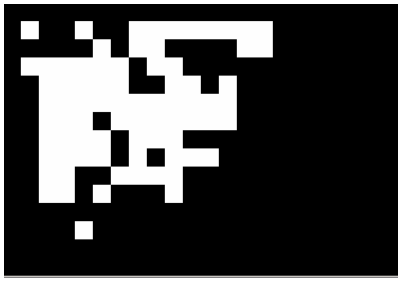

Fig. 18. Example macroblock modes for Mobile & Calendar, grey: LT-GMC, black: other modes

6. CONCLUSIONS AND FUTURE WORK

We have presented a new approach to video coding. It relies on an automatic detection of global motion regions, which are encoded with a special algorithm. It works best for high-resolution sequences that are dominated by global motion. In the worst case, if no LT-GMC macroblocks are assigned, the algorithm falls back to standard H.264/AVC. In such cases the transmission of global motion parameters should be switched off resulting in a neglectable overhead. This still needs to be implemented. Also the integration of LT-GMC for B-frames is still an open issue. Finally, we need to conduct formal subjective tests to accurately evaluate our results.

REFERENCES

1. ISO/IEC 14496, Part 2 (Visual), Ammendment 1 “Information Technology - Coding of Audio-Visual Objects (MPEG-4)”, February 2000.
2. Joint Video Team of ITU-T and ISO/IEC JTC 1, “Draft ITU T Recommendation and Final Draft International Standard of Joint Video Specification (ITU-T Rec. H.264 | ISO/IEC 14496-10 AVC),” Joint Video Team (JVT) of ISO/IEC MPEG and ITU-T VCEG, JVT-G050, March 2003.
3. P. Ndjiki-Nya, B. Makai, A. Smolić, H. Schwarz, and T. Wiegand, “Improved H264/AVC Coding Using Texture Analysis and Synthesis”, Proc. ICIP’03, IEEE International Conference on Image Processing, Barcelona, Spain, September 2003.
4. A.V.Oppenheim, R.W. Schafer, “Discrete-Time Signal Processing”, Prentice-Hall, Englewood Cliffs, NJ, USA, 1989.
5. A. Smolić, T. Sikora, and J.-R. Ohm, “Long-Term Global Motion Estimation and its Application for Sprite Coding, Content Description and Segmentation”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 8, pp. 1227-1242, December 1999.
6. A. Smolić, and T. Wiegand, “High-Resolution Video Mosaicing”, Proc. ICIP’01, IEEE International Conference on Image Processing, Thessaloniki, Greece, October 2001.
7. A. Smolić, Y. Vatis, and T. Wiegand, “Long-Term Global Motion Compensation Applying Super-Resolution Mosaics”, Proc. ISCE’02, IEEE International Symposium on Consumer Electronics, Erfurt, Germany, September 2002.
8. T. Wiegand, X. Zhang, and B. Girod, “Long-Term Memory Motion-Compensated Prediction”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 9, No. 1, pp. 70-84, Feb. 1999.
9. T. Wiegand, H. Schwarz, A. Joch, F. Kossentini, and G. J. Sullivan, “Rate-Constrained Coder Control and Comparison of Video Coding Standards”, IEEE Transactions on Circuits and Systems for Video Technology, Vol. 13, No. 7, pp. 688-703, July 2003.