SCALABLE VIDEO CODING AT VERY LOW BIT RATES EMPLOYING RESOLUTION PYRAMIDS

Klaus Illgner and Frank Müller

Institut für Elektrische Nachrichtentechnik
RWTH Aachen, 52056 Aachen, Germany
Tel: +49-241-80-7681; Fax: +49-241-8888-196
e-mail: {illgner,mueller}@ient.rwth-aachen.de

ABSTRACT

In this paper an approach for scalable video coding is described, based on the hybrid coding scheme. The scalability is achieved by decomposing the frames to be coded into a resolution pyramid. Motion estimation and compensation is performed at each level. The focus of the paper is to design motion estimation and compensation such, that the resulting pyramid of vector fields as well as the pyramid of prediction errors can be coded in an efficient fashion.

1 INTRODUCTION

Video coding schemes at very low bit rates are typically based on the hybrid coding principle, e.g. H.263, MPEG. Temporal redundancies are reduced by motion compensation. The remaining prediction error, also termed displaced frame difference, needs to be coded as well, since the motion based prediction fails in some regions. The advantage of the hybrid coding scheme, which is in fact a DPCM loop, is the low delay property (one frame). This characteristic is especially required for communications applications. A major drawback of the hybrid coding scheme is the lack of scalability. Scalability refers in this paper to spatial scalability. A video sequence of lower spatial resolution can be obtained by decoding only a part of the received bit stream.

Currently, there is strong demand for scalable video coding schemes. Due to the rapid development of chips and networks, video communications are of increasing interest. Bandwidth limitations, multipoint operation with receivers of different capabilities, and bandwidth dependent charging are some good reasons for scalable coding schemes. Also MPEG-4 announced strong demands for scalable video coding algorithms.

Almost all scalable codecs based on the hybrid coding scheme utilize multiple DPCM loops on different resolution levels. For decomposition of images into a multiresolution representation exist two main classes of approaches. One class is based on subband or wavelet decompositions [7][8][9]. Advantageous is, that there is no data expansion, since the subband images are critically sampled. Motion estimation and compensation is performed on the subband images, which have bandpass characteristics (except the highest and lowest level). The second class uses decompositions into resolution pyramids [1]. Due to the overcomplete data representation, a perfect reconstruction of the higher resolution levels of the pyramid can be obtained, even if the lower resolution levels contain quantization errors. Advantageous is the high degree of flexibility in the design of the filters.

The different scalable coding approaches vary in the degree of coupling the DPCM codecs of the resolution levels [2][3][7]. Typically, the displaced frame differences and the vector fields of the hybrid coders at the different resolution levels are coded independently in the main. Hence, the coding efficiency is limited. Furthermore, some codecs generate a single bit stream, which requires additional symbols for separating the parts in a single bit stream.

The scalable video coding approach described in this paper employs resolution pyramids and uses DPCM loop in each scalable resolution level. The main difference to all other approaches is, that the resolution pyramids of the motion vector fields and the displaced frame differences are encoded in an embedded fashion using zerotrees [4][6]. The key to efficient coding performance is a suitable design of motion estimation and compensation, which therefore is analyzed in more detail.

The complete codec is outlined in the next section, while in the third section the motion estimation and compensation is described.

2 OUTLINE OF THE SCHEME

In fig.1 the basic single resolution coding scheme is depicted[1]. The prediction error image after motion compensation is decomposed into a Gaussian pyramid $\mathbf{D} = \{d_t^{(0)}, d_t^{(1)}\}$ [2]. From this pyramid a Laplacian pyramid is derived, which is encoded in an efficient fashion

[1] The symbol $\boxed{\downarrow 2}$ denoted filtering and subsequent subsampling.

[2] Subscripts denote time a index while superscripts denote a level index.

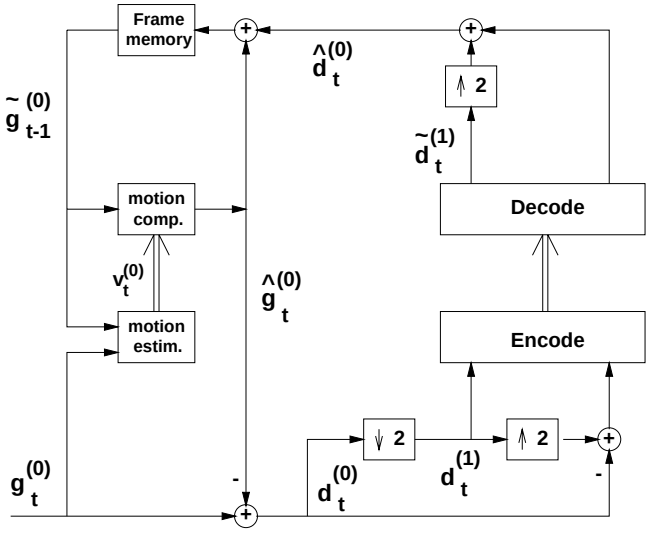

Figure 1: Block diagram of the single resolution encoder using a 2-level resolution pyramid

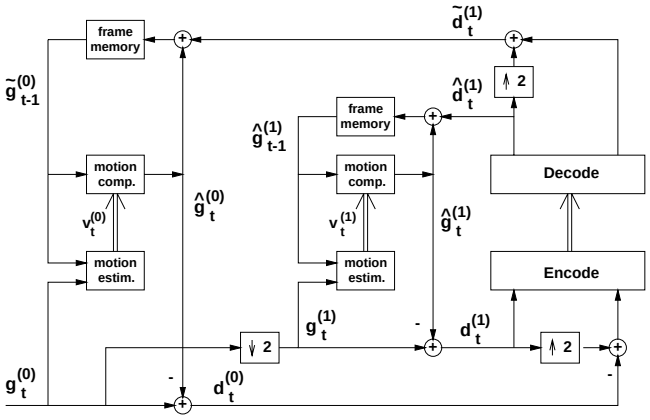

Figure 2: Block diagram of the 2-level scalable coding scheme

using zero-trees [6]. The motion vector field $\vec{v}_t^{(0)}$ is estimated between the the new frame $g_t^{(0)}$ and the previous decoded frame $\tilde{g}_{t-1}^{(0)}$. The decoded prediction error frame is denoted by $\tilde{d}_t^{(0)}$.
For a scalable multiresolution codec at very low bit rates at least the CIF and QCIF resolutions should be provided. The input frame is lowpass filtered and subsampled, resulting in a Gaussian pyramid of 2 levels. Motion estimation is performed at the highest resolution (CIF) as well as on the lower resolution level. Within each level of the pyramid a prediction is calculated employing motion compensation. The resulting prediction error levels form again a pyramid, which is encoded in an embedded fashion using [6]. Since the motion vector fields at the levels represent also a pyramid, the motion information is coded using the approach described in [4]. The block diagram of the coding scheme is depicted in fig. 2.
For the performance of the codec it is essential that the prediction error levels as well as the motion vector fields can be coded efficiently. The coding efficiency of the resolution pyramid of the motion vector fields depends on motion estimation while motion compensation determines the coding performance for the prediction error. Therefore, regarding these constraints motion compensation and estimation is analyzed in the next section.

3 HIERARCHICAL MOTION COMPENSATION

Let g_t denote the frame to be coded at time t. Disregarding motion estimation and compensation for the moment, the predicted frame $\hat{g}_t$ is simply the previous frame g_{t-1}. The difference frame, also termed prediction error frame, $d_t = g_t - \hat{g}_t$ is coded.
It has been shown, that difference frames can efficiently be coded with resolution pyramids [6]. The resolution pyramid $\mathbf{D}_t$ is calculated by filtering and subsequent subsampling by a factor of 2 in each spatial dimension of each level

$$\mathbf{D}_t = \{d_t^{(0)}, d_t^{(1)}, \ldots, d_t^{(N-1)}\} \tag{1}$$

with

$$d_t^{(k)} = f_R\left(d_t^{(k-1)}\right), \quad d_t^{(0)} = d_t. \tag{2}$$

The filter is denoted by $f_R()$. Furthermore, let $\mathbf{G}_t$ and $\hat{\mathbf{G}}_t$ denote pyramid decompositions of the frames g_t and $\hat{g}_t$, respectively, employing the same filter $f_R()$.
As is shown in eq. 3, under the constraint of linear filters $f_R()$, $d_t^{(k)}$ can be derived also directly from the corresponding levels $g_t^{(k)}$ and $\hat{g}_t^{(k+1)}$ of the pyramids $\mathbf{G}_t$ and $\hat{\mathbf{G}}_t$:

$$\begin{aligned} d_t^{(k+1)} &= f_R\left(d_t^{(k)}\right) \\ &= f_R\left(g_t^{(k)} - \hat{g}_t^{(k)}\right) \\ &= f_R\left(g_t^{(k)}\right) - f_R\left(\hat{g}_t^{(k)}\right) \\ &= g_t^{(k+1)} - \hat{g}_t^{(k+1)}. \end{aligned} \tag{3}$$

More compact, D can be obtained by

$$\mathbf{D}_t = \mathbf{G}_t - \hat{\mathbf{G}}_t. \tag{4}$$

Regarding a single resolution approach eq. 4 is independent of the method for motion estimation and compensation, since the motion compensated frame $\hat{g}_t$ is decomposed.
Much more important, this property provides also a basis for the development of a scalable video coding scheme. From eq. 3 follows, that optimal scalability is achieved, if the motion compensated frame at level k, denoted by

$$\hat{g}_t^{(k)} = m_c(g_{t-1}^{(k)}, \vec{v}_t^{(k)}), \tag{5}$$

followed by filtering, is equivalent to motion compensation of the filtered frame

$$f_R(\hat{g}_t^{(k)}) \equiv m_c(f_R(g_{t-1}^{(k)}), \vec{v}_t^{(k+1)}). \quad (6)$$

m_c denotes the motion compensation function. The relation between the motion vector field used for motion compensation at level k, denoted by $\vec{v}_t^{(k)}$, and $k+1$ can be arbitrary and needs not to be determined for the moment.
Analyzing typical motion compensation schemes in the literature almost always some kind of mapping is used. The predicted frame $\hat{g}$ at location (x, y) is the sum of i small pattern signals $\tilde{g}_i$

$$\hat{g}(x, y) = \sum_i \tilde{g}_i(x - \Delta x, y - \Delta y) \cdot w_i(x, y). \quad (7)$$

Each pattern might represent a single pixel, a region or a block of a reference frame at the displaced location $(x - \Delta x, y - \Delta y)$. The window function determines the size and the weight of the patterns, e.g. for block matching w_i is a rectangular step function. By filtering the predicted frame, aliasing is introduced by the window w_i. Hence, perfect equivalence in eq. 6 can not be obtained.
However, since the aliasing effect is limited in case of large displaced regions, block based motion compensation is a useful approximation. Another advantage is the simplicity and robustness of block matching. Furthermore, overlapped block compensation uses a smoothing window instead of step function window. Hence, the aliasing effects are reduced.

3.1 Motion Estimation

The aim for coding applications should be not to obtain the best estimate only for the highest resolution level, but also good predictions for all resolution levels. However, motion estimation needs also to be constrained to the coding costs of the motion vector fields. This can quite easily be accomplished by hierarchical top-down motion estimation. The vector field obtained at level k is refined by $\Delta\vec{v}_t^{(k-1)}$ at the next lower level $k-1$

$$\vec{v}_t^{(k-1)} = f_E\left(\vec{v}_t^{(k)}\right) + \Delta\vec{v}_t^{(k-1)}. \quad (8)$$

The search area size for block matching at the next lower level is equivalent to the refinement $\Delta\vec{v}_t^{(k-1)}$, and determines the coding costs. Enlarging will provide a better prediction, but at extra costs for coding of the motion vector field. $f_E()$ denotes an appropriate expand and scaling function. The design of the function is non-trivial in general, due to discontinuities of the vector field at boundaries of moving objects.
Note, that here is no obvious relation between the motion vectors $\vec{v}^{(k)}$ at different levels k. Considering for instance pure translational motion as the simplest motion model, the displacements caused by motion remain after lowpass filtering. However, due to aliasing (eq. 7), a motion estimator at a lower resolution level will hardly be able to detect the same displacements.

4 REALIZATION

For coding a frame g_t, the frame is first decomposed into a resolution pyramid $\mathbf{G}_t$ consisting of N levels employing linear filters. A pyramid representation of the previous reconstructed frame $\tilde{\mathbf{G}}_t$ is already available from coding the last frame. Based on these pyramids, multiresolution block matching is performed starting at the level $N_m < N$, with N_m denoting the number of scalability layers. The aim for motion estimation is to obtain a sufficient good prediction for each level $k \in \{0, \ldots, N_m - 1\}$ of the pyramid. Therefore, the estimation criterion incorporates a smoothness constraint and includes overlapped block motion compensation [5]. The effects are reduced coding costs (smooth vector field) and reduced artifacts (overlapped window), and hence a reduced prediction error. The displaced frame difference at each level $0 \le k < N_m$ is given by

$$d_t^{(k)} = g_t^{(k)} - \hat{g}_t^{(k)}, \quad \text{with} \quad \hat{g}_t^{(k)} = \tilde{g}_{t-1}^{(k)}(\vec{v}_t^{(k)}).$$

The vector field obtained at level k is refined at the next lower level $k-1$ according to eq. 8 by block matching considering the vector field of level k. The block size at the next lower level is enlarged by a factor 2. Hence, the expand function $f_E()$ reduces to a scaling function, which scales the vectors by 2. Finally, a multi-resolution representation of the vector field is also obtained. Although the vector fields form no true pyramid any more, they still can be coded efficiently by the approach presented in [4].
To obtain a resolution pyramid $\mathbf{D}_t$ of displaced frame differences of N levels, the remaining levels $N_m \le k < N$ are calculated according to eq. 3 by

$$d_t^{(k)} = g_t^{(k)} - f_R(\tilde{g}_{t-1}^{(k-1)}. \quad (9)$$

It turned out, that the displaced frame differences $d_t^{(k)}, 0 \le k < N_m$ have similar properties as the pyramid levels, which are obtained by filtering the predicted error frame obtained by motion estimation and compensation at the lowest level. Hence, following the coding scheme described in [6] the pyramid $\mathbf{D}_t$ can be efficiently coded.

5 RESULTS

A 2-level coding scheme has been simulated, the higher resolution at CIF and the lower resolution at QCIF format. The frames #163 and #166 of the test sequence *silent* have been used to obtain the results.
For comparison the CIF and QCIF resolution frames have been coded with the single resolution codec, as depicted in fig.1. Table 1 summarizes the results. Motion is estimated between the original frames, which are also used for motion compensation. The block size is 16×16

at CIF resolution and 8×8 at QCIF resolution. The motion vector field has been coded employing [4]. For coding of the displaced frame difference a 5-level (4-level for QCIF) resolution pyramid has been used [6].

format	$\frac{\mathrm{R}(\vec{v})}{[\mathrm{bit}]}$	$\frac{\mathrm{PSNR}(\hat{g}_t)}{[\mathrm{dB}]}$	$\frac{\mathrm{R}(d)}{[\mathrm{bit}]}$	$\frac{\mathrm{PSNR}(\tilde{g}_t)}{[\mathrm{dB}]}$
CIF	1895	32.05	6060	35.08
QCIF	1722	33.67	4104	37.05

Table 1: Results of single resolution encoding

The results of the 2-level scalable coding scheme are given in table 2. The PSNR after motion compensation is for both levels comparable, although the complexity is significantly reduces due to hierarchical block matching. The slightly differing PSNR is caused by a different lowpass filter used to obtain the single resolution QCIF frame. The rate necessary to transmit the motion vector field is lower than the sum of the rates for encoding the single resolution vector fields. Interestingly, the update coding reaches the same PSNR for both layers as in case of the single resolution coding. Again, the data rate is significantly reduced. Overall, the scalable coding scheme saves about $\approx 38\%$ bit rate compared to the data rate required to encode the frames independently at two single resolutions.

$\frac{\mathrm{R}(\vec{v})}{[\mathrm{bit}]}$	PSNR/[dB]		$\frac{\mathrm{R}(\mathbf{D})}{[\mathrm{bit}]}$	PSNR/[dB]	
	$\hat{g}_t^{(0)}$	$\hat{g}_t^{(1)}$		$\tilde{g}_t^{(0)}$	$\tilde{g}_t^{(1)}$
2854	32.72	33.85	5939	35.08	37.05

Table 2: Results of 2-level scalable coding

6 SUMMARY

An efficient scalable video coding scheme has been described. The input frames are decomposed into a Gaussian pyramid. Each level is designed as an individual DPCM loop. However, motion estimation is performed in a hierarchical fashion, such that the displacement estimates of higher levels are used as initial estimates at the next lower level. The resulting resolution pyramid of motion vector fields as well as the pyramid of prediction errors are encoded in an embedded fashion. The experiments proofed the efficiency of the scalable coding scheme. The rate to achieve the same PSNR at two resolution levels is much lower compared to the sum of two independently encoded frames of different resolution.

References

[1] P. J. Burt and E. H. Adelson, "The Laplacian pyramid as a compact image code," *IEEE Trans. Commun.*, vol. 31, pp. 532–540, Apr. 1983.

[2] B. Girod, U. Horn, and B. Belzer, "Scalable video coding with multiscale motion compensation and unequal error protection," in *Proc. of Symposium on Multimedia Communications and Video Coding*, (New York City, USA), Oct. 1995.

[3] T. Hanamura, W. Kameyama, and H. Tominaga, "Hierarchical coding scheme of video signals with scalability and compatibility," *Signal Processing: Image Communication*, vol. 5, pp. 159–184, Feb. 1993.

[4] K. Illgner and F. Müller, "Hierachical coding of motion vector fields," in *Proceedings IEEE Intern. Conference on Image Processing ICIP'95*, vol. I, (Washington DC, USA), pp. 566–569, Oct. 1995.

[5] K. Illgner and F. Müller, "Motion estimation using overlapped block motion compensation and Gibbs-modeled vector fields," in *Proc. 9$^{\mathrm{th}}$ Workshop on Image and Multidimensional Signal Processing (IMDSP 96)*, (Belize City, Belize), Mar. 1996.

[6] F. Müller and K. Illgner, "Embedded pyramid coding of displaced frame differences," in *Proceedings IEEE Intern. Conference on Image Processing ICIP'95*, vol. III, (Washington DC, USA), pp. 216–219, Oct. 1995.

[7] T. Naveen and J. W. Woods, "Motion compensated multiresolution transmission of high definition video," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 4, pp. 29–41, Feb. 1994.

[8] K. M. Uz, M. Vetterli, and D. J. LeGall, "Interpolative multiresolution coding of advanced television with compatible subchannels," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 1, pp. 86–99, Mar. 1991.

[9] Y.-Q. Zhang and S. Zafar, "Motion-compensated wavelet transform coding for color vidoe compression," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 2, pp. 285–296, Sept. 1992.