Wavelet-Based Very Low Bit-Rate Video Coding Using Image Warping and Overlapped Block Motion Compensation

Guido Heising, Detlev Marpe, Hans L. Cycon and Alexander P. Petukhov

Corresponding author:

Detlev Marpe

Heinrich-Hertz-Institute

Image Processing Department

Einsteinufer 37

10587 Berlin

Germany

G. Heising and D. Marpe are with the Image Processing Department, Heinrich-Hertz-Institute (HHI) for Communication Technology, Berlin, Germany. H. L. Cycon is with the Fachhochschule für Technik und Wirtschaft, Berlin, Germany. A. P. Petukhov is with the Dept. of Mathematics, University of South Carolina, Columbia, USA.

Abstract

This paper presents an algorithm for very low bit-rate video coding that combines new ideas in motion estimation, wavelet filter design, and wavelet-based coding techniques. A new motion compensation technique using image warping and overlapped block motion compensation is proposed to reduce temporal redundancies in a given image sequence. This combined motion model has the advantage of representing more complex motion than simple block matching schemes. To further improve the quality of the temporal prediction, an adaptive grid with variable density according to the varying motion activity of a given scene is generated. An adaptively switched high-quality texture interpolation is employed to cope with the problem of fractional displacements in such a way that both objective and subjective reconstruction quality is improved. Spatial decorrelation of the motion compensated residual images is performed using an one-parametric family of biorthogonal infinite impulse response (IIR) wavelet filters coupled with the highly efficient pre-coding scheme of 'partitioning, aggregation and conditional coding' (PACC). Experimental results demonstrate significant improvements in objective quality of 1.0–2.3 dB PSNR in comparison to the H.263+ test model TMN10 using advanced coding options. In addition, our intra coding method provides a performance gain of 0.5 dB PSNR on the average for a test suite of various still images when comparing to the emerging still image coding standard JPEG-2000.

1. Introduction

Traditionally, very low bit-rate hybrid video coding schemes consist of methods of motion estimation, motion compensation and transform coding which rely on block-based techniques. One of the most successful representatives of this generation of video coding schemes is adopted by the ITU-T low bit-rate video compression standard H.263 [2]. Although this video coding standard provides a high coding efficiency, it has the well-known drawback of suffering from blocking artifacts especially at very low bit-rates. In this work, we propose a hybrid video coding scheme which combines very effective non block-based or overlapping block-based techniques for temporal prediction with an efficient frame-based method of spatial decorrelation. As we will demonstrate, the proposed scheme significantly improves coding efficiency as well as visual quality when compared to the most efficient currently available video compression standard.

The temporal prediction scheme of our proposed hybrid coding approach utilizes two different motion models: a bilinear image warping motion model [7], [23] and an overlapped block motion compensation (OBMC) with cosine weighted window functions [21]. Although the image warping model induces a subdivision of a given image into square "blocks", the obtained motion vector field is smooth, especially across the block boundaries. However, in the presence of objects performing different movements, motion discontinuities along object boundaries are not well represented in the image warping model. In this case, OBMC may offer an instrument for a better prediction by means of a superposition of overlapping displaced blocks from the reference frame, each weighted by a smooth cosine window. Thus, we propose an adaptive block-wise switching between both motion models controlled by the objective of minimizing a given costfunction.

In addition to our previous work on this subject [9], [15], we propose to further improve the prediction by using an adaptive two-level quadtree grid instead of a fixed regular grid. Our approach is similar to the idea presented in [11]. However, in contrast to this method which uses the local variance of a given frame difference as a criterion to decide whether a block should be split into smaller ones, we propose a bottom-up algorithm for generating a hierarchical grid. Our method is based on a threshold decision with respect to the variation of local prediction error energy, where the threshold depends on the given motion activity. This approach enables us to merge grid points on an initially given fine grid in stationary areas or areas with locally uniform motion, thus reducing the overall cost of side information with little impairment of prediction quality.

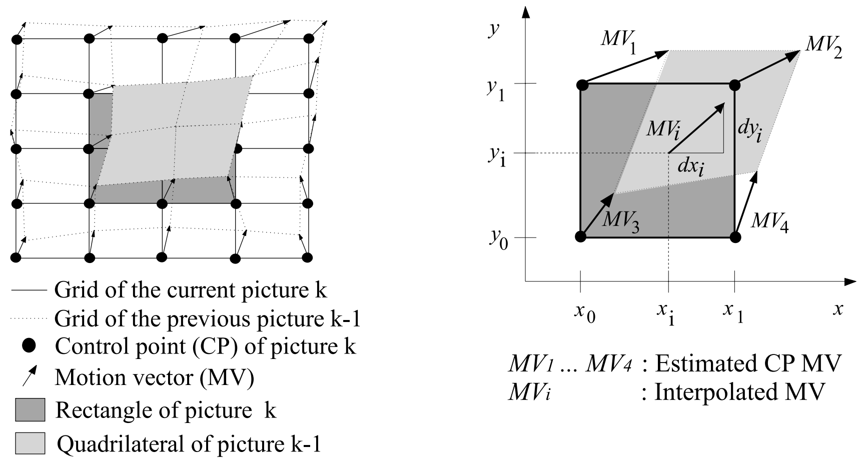
Prediction residual frames (interframes) as well as full frames (intraframes) are decorrelated using a discrete wavelet transform (DWT) which is realized by an appropriately designed pair of infinite impulse response (IIR) filters. In contrast to biorthogonal wavelets with compact support [5] most frequently used in the image coding community, biorthogonal bases generated by recursive filters [17] offer more flexibility for the adaptation to specific coding problems. To further enhance the transform coding scheme, an additional pre-coding part is employed where the quantized wavelet coefficients are pre-processed prior to arithmetic coding. This pre-coder is based on the concepts of partitioning, aggregation and conditional coding (PACC), a framework, which was first introduced in [13] and which has proven to be very efficient in still image compression and very low bit-rate video coding [14], [16].

The organization of the paper is as follows. Section 2 describes the motion models, the estimation of the model parameters and the motion compensation algorithm. In Section 3, we discuss the employed type of wavelet transform together with the PACC pre-coding framework and some related pre-coding methods. Section 4 contains experimental results comparing our proposed video coding scheme to the current test model TMN10 of H.263+ and for the case of pure intra coding to the verification model of JPEG-2000. Conclusions can be found in Section 5.

2. Temporal Prediction Scheme

2.1 Motion Model

Two different motion models are employed in order to increase the quality of the predicted frames. A bilinear geometric transform [23] is used for warping prediction



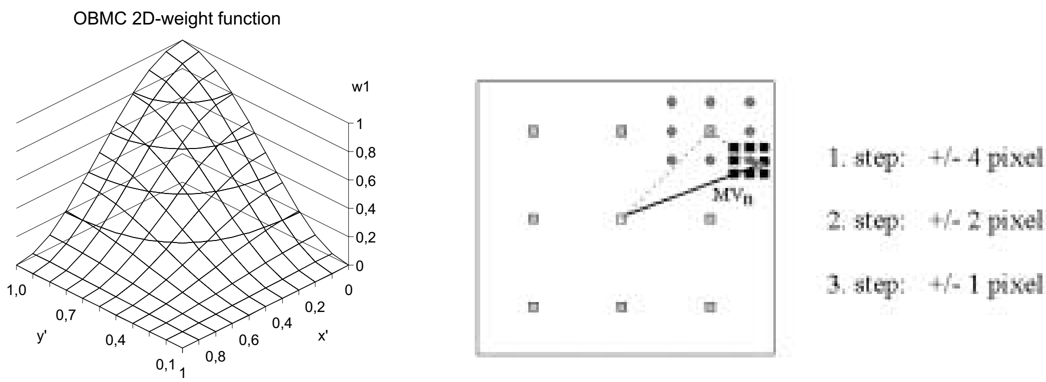
(a) (b)

Fig. 1. (a) Principle of image warping prediction. (b) Spatial interpolation of motion vectors.

and cosine weighted windows (Fig. 2 (a)) are utilized for OBMC [21]. According to Fig. 1 (a), the current frame k is subdivided into squares of 16 by 16 pels, thus obtaining a regular grid. In contrast to H.263 the positions of the motion vectors are on the vertices of the squares which leads to an additional row and column of motion vectors. The estimated vectors are used for warping predicition, OBMC or both (in case of neighboring blocks utilizing different motion models).

A dense motion vector field is achieved by using a bilinear geometric transform (warp) which smoothly varies over the image. In addition to translational motion the model can describe rotation, shear and change in scale. The 8 parameters of the bilinear transform can be described by the motion vectors of 4 points. Therefore the motion vectors assigned to the four vertices of a square are used. The motion vector $\boldsymbol{MV}_i$ inside a square block is interpolated from the four surrounding control point motion vectors $\boldsymbol{MV}_1 \ldots \boldsymbol{MV}_4$ ($cf.$ Fig. 1 (b)) using the following equation:

$$\begin{aligned} \boldsymbol{MV}_i = \begin{pmatrix} dx_i \\ dy_i \end{pmatrix} &= \begin{pmatrix} dx_1 \\ dy_1 \end{pmatrix} \cdot (\tilde{y}_i - \tilde{x}_i \cdot \tilde{y}_i) + \begin{pmatrix} dx_2 \\ dy_2 \end{pmatrix} \cdot (\tilde{x}_i \cdot \tilde{y}_i) \\ &+ \begin{pmatrix} dx_3 \\ dy_3 \end{pmatrix} \cdot (1 - \tilde{x}_i - \tilde{y}_i + \tilde{x}_i \cdot \tilde{y}_i) + \begin{pmatrix} dx_4 \\ dy_4 \end{pmatrix} \cdot (\tilde{x}_i - \tilde{x}_i \cdot \tilde{y}_i) \quad (1) \end{aligned}$$



(a) **(b)**

Fig. 2. (a) 2D weighting function w_1 for OBMC-based prediction. (b) Principle of n-step motion search (for $n = 3$).

where

$$\tilde{x_i} = \frac{x_i - x_0}{x_1 - x_0}, \quad \tilde{y_i} = \frac{y_i - y_0}{y_1 - y_0}.$$

This kind of warping prediction, also known as control grid interpolation [19], leads to a motion vector field without motion discontinuities. The smoothness across border lines is achieved by employing the same two motion vectors on both sides of the border, *e.g.* for $x_0 - x_i \approx \pm 0$ according to Eq. (1) only $\boldsymbol{MV}_1$ and $\boldsymbol{MV}_3$ contribute to $\boldsymbol{MV}_i$. To be more specific, the intensity $I(x_i, y_i, k)$ of frame k at a given position (x_i, y_i) is predicted by

$$p^{\text{warp}}(x_i, y_i, k) = \hat{I}(x_i + dx_i, y_i + dy_i, k-1), \tag{2}$$

where $\hat{I}(\cdot, \cdot, k-1)$ denotes the itensity of the decoded past frame $k-1$. Hereby blocking artifact free predictions are obtained, which is essential for the performance of the proposed loop filter described in Section 2.4 and the wavelet transform coding part (*cf.* Section 3).

However, in the case of sequences with differently moving objects, the warping model is not capable of dealing properly with motion discontinuities at the object borders. In these blocks, overlapped block motion compensation is employed by superimposing 4 predicted intensity values $p_1^{\text{trans}}, \ldots, p_4^{\text{trans}}$ using non linear weighting functions $w_1, \ldots, w_4$ (*cf.* Fig. 2 (a)):

$$\begin{aligned} p^{\text{obmc}}(x_i, y_i, k) &= \sum_{j=1}^{4} p_j^{\text{trans}}(x_i, y_i, k) \cdot w_j(x_i, y_i) \\ &\stackrel{\text{def}}{=} p_1^{\text{trans}} \cdot (\hat{y}_i - \hat{x}_i \cdot \hat{y}_i) + p_2^{\text{trans}} \cdot (\hat{x}_i \cdot \hat{y}_i) \\ &\quad + p_3^{\text{trans}} \cdot (1 - \hat{x}_i - \hat{y}_i + \hat{x}_i \cdot \hat{y}_i) + p_4^{\text{trans}} \cdot (\hat{x}_i - \hat{x}_i \cdot \hat{y}_i) \,, \end{aligned} \tag{3}$$

where

$$\hat{x}_i \stackrel{\text{def}}{=} \frac{1}{2}\left(1 - \cos\left(\pi \cdot \frac{x_i - x_0}{x_1 - x_0}\right)\right), \quad \hat{y}_i \stackrel{\text{def}}{=} \frac{1}{2}\left(1 - \cos\left(\pi \cdot \frac{y_i - y_0}{y_1 - y_0}\right)\right)$$

and

$$p_j^{\text{trans}} = p_j^{\text{trans}}(x_i, y_i, k) \stackrel{\text{def}}{=} \hat{I}(x_i + dx_j, y_i + dy_j, k - 1), \quad j = 1, \ldots, 4.$$

Thus, the four predicted values $p_1^{\text{trans}}, \ldots, p_4^{\text{trans}}$ are computed by employing the translational motion model with one of the motion vectors $\boldsymbol{MV}_1, \ldots, \boldsymbol{MV}_4$ of the four surrounding vertices for each prediction. For each block, one bit is transmitted indicating the used prediction type.

By combining these two compensation schemes, namely warping prediction and OBMC, no blocking is introduced *a)* between adjacent blocks using the same prediction scheme and *b)* around vertices of neighboring blocks using different schemes. The latter is a result of using, on the one hand, the same vertex motion vectors for both schemes. On the other hand, both schemes converge to the same prediction in the vicinity of a vertex. For example, let us assume that inside the square of Fig. 1 (b) OBMC is used and outside warping prediction is employed. In that case, close to the upper left vertex warping prediction p^{warp} relies only on the motion vector $\boldsymbol{MV}_1$ while for OBMC only the predicted value p_1^{trans} of the respective vertex is used, which leads to equal predictions. Only little blocking artifacts may be introduced around the center of the border line of two adjacent blocks using two different prediction schemes.

2.2 Motion Estimation

For motion estimation the current picture is subdivided into squares, thus creating a control grid. For every control point (vertex of a square), a motion vector is estimated by *octagonal matching*, *i.e.* finding the minimum mean-squared error (MSE) between the predicted and the current four surrounding squares of a control point (Fig. 1 (a)).

In doing so the motion vectors of the eight neighboring control points are fixed and only the motion vector of the center control point is changed. Because of their interdependence, the motion vectors are iteratively refined. Therefore in each iteration the control points are scanned from top left to bottom right of the image. By using a n-step search algorithm, $n = 4$ iterations or, in case of larger motion, $n = 6$ iterations are performed with step-sizes varying in the set $\{16, 8, 4, 2, 1, 0.5\}$ for a six iteration search (Fig. 2 (b)). As the number of iterations n corresponds to the number of steps, the total number of tested candidate vectors per control point amounts to $8n + 1$.

In order to improve the prediction in presence of large motion a hierarchical motion estimation scheme changing the grid size from coarse to fine is employed, *e.g.* an initial grid size of 32×32 followed by 16×16 pels can be used. The vectors of the finer grid are initialized by bilinear interpolation and subsequent rounding to half pel accurate values of the prior estimated vectors on the coarse grid. This also leads to a more homogeneous motion vector field. To further force the estimation of smooth vector fields a Lagrangian multiplier λ^{estim} is used to choose the best control point motion vector considering the prediction error (in the MSE sense) and the local motion vector variance between the candidate vector $\boldsymbol{MV}_j$ and the eight motion vectors $\boldsymbol{MV}_k$ of its neighboring control points:

$$\boldsymbol{MV}_i^{\text{opt}} = \arg\min_{\boldsymbol{MV}_j} \left(\text{MSE}\left(\boldsymbol{MV}_j\right) + \lambda^{\text{estim}} \cdot \sum_{k=1}^{8} \left(\boldsymbol{MV}_k - \boldsymbol{MV}_j\right)^2 \right). \qquad (4)$$

The decision whether to use warping prediction or OBMC for a block is also based on Eq. (4). The prediction type yielding lower cost is chosen. To avoid a large increase of the computational workload, the MSE of the prediction error for both schemes, *i.e.* warping prediction and OBMC, is computed for the motion vector in question only under consideration of the bottom right subblock of the octagon (*cf.* Fig.1 (a)). The three other subblocks are using the prediction schemes which

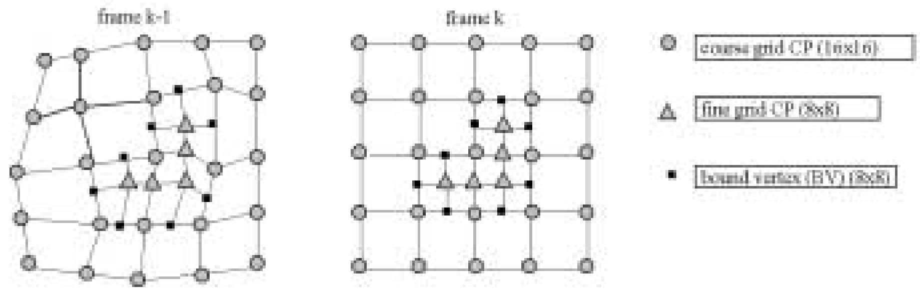


Fig. 3. Principle of warping prediction using a variable two-level quadtree grid with square blocks of size 8×8 and 16×16 pels.

were previously assigned during motion estimation of vectors at causal positions.

2.3 Hierarchical Control Grid Interpolation

The use of a regular grid of block size 16×16 pels imposes a severe constraint on the motion model with respect to the most common situation where both highly active and (quasi-) stationary regions are given. For a better adaptation to this kind of scenes, we consider an extension of our motion model based on an irregular grid with square blocks of size 8×8 and 16×16 pels. To simplify the discussion, let us first consider a pure image warping prediction. Figure 3 gives an illustration of the warping model based on such an irregular grid of squares. As usual, the prediction is obtained by relating the grid of the current frame k to a warped grid of the previous frame $k-1$. Note however, that the irregular grid consists of two kinds of vertices. The grid points which are connected to four other grid points are freely moving control points (CPs) as indicated in Fig. 3. The other type of grid points are bound vertices (BVs) forming a T-shaped connection between three other grid points. In contrast to a CP, the motion vector (MV) of a BV is bilinearly interpolated from the MVs of neighboring CPs, and hence need not be transmitted. The overall structure of the grid can be described by a quadtree [11].

The motion estimation process is a hierarchically organized scheme operating on successively refined regular grids of sizes 32×32, 16×16 and 8×8 pels in a way, it was described in the previous section. As a result of this first estimation step, a motion vector field on a fine control grid $\mathcal{G}_{8\times8}$ of size 8×8 pels is obtained. In a second step, CPs on the fine grid which are not grid points on the coarser grid $\mathcal{G}_{16\times16}$ of size 16×16 pels are examined in order to merge the four surrounding 8×8 blocks to one 16×16 block of those CPs which are related to homogeneously moving areas. Assuming a control point $c_i \in \mathcal{G}_{8\times8} \setminus \mathcal{G}_{16\times16}$ is given, we evaluate the local increase in prediction error ΔD_i (in terms of MSE) obtained by replacing the actual estimated motion vector $\boldsymbol{MV}_i^{\mathrm{opt}}$ of c_i by its bilinearly interpolated motion vector $\boldsymbol{MV}_i^{\mathrm{ipol}}$ given by the four motion vectors of its neighboring control points of $\mathcal{G}_{16\times16}$, *i.e.* $\Delta D_i \stackrel{\mathrm{def}}{=} D(\boldsymbol{MV}_i^{\mathrm{ipol}}) - D(\boldsymbol{MV}_i^{\mathrm{opt}})$. If the increase in prediction error can be controlled by the norm of the difference $||\Delta \boldsymbol{MV}_i||$ of these both candidates, *i.e.* if

$$\Delta D_i \leq \lambda^{\mathrm{merge}} \cdot ||\Delta \boldsymbol{MV}_i|| = \lambda^{\mathrm{merge}} \cdot ||\boldsymbol{MV}_i^{\mathrm{ipol}} - \boldsymbol{MV}_i^{\mathrm{opt}}|| \qquad (5)$$

with some given pre-determined threshold λ^{merge}, then $\boldsymbol{MV}_i^{\mathrm{opt}}$ is substituted by $\boldsymbol{MV}_i^{\mathrm{ipol}}$ and the related control point c_i is removed from the list of unrestricted control points, *i.e.* the set of potential candidate control points of the final irregular grid.

The underlying assumption of criterion (5) is that, given a homogeneous area of the motion vector field, the prediction error $D(\boldsymbol{MV}_i^{\mathrm{opt}})$ as a function of the optimal motion vector $\boldsymbol{MV}_i^{\mathrm{opt}}$ should be a locally smooth function with bounded variation and, hence, the interpolated motion vector $\boldsymbol{MV}_i^{\mathrm{ipol}}$ may be a good approximation of the optimal choice $\boldsymbol{MV}_i^{\mathrm{opt}}$ on the fine grid. If however, on the other hand, $\Delta D_i/||\Delta \boldsymbol{MV}_i|| > \lambda^{\mathrm{merge}}$ holds for the given threshold λ^{merge}, there is a strong evidence that either there is a local singularity in the motion vector field or $\boldsymbol{MV}_i^{\mathrm{opt}}$ represents a strong local minimum of $D(\cdot)$, so that in both cases, it is a reasonable choice to keep the related control point c_i as a candidate of the final irregular grid. After each control point on $\mathcal{G}_{8\times8} \setminus \mathcal{G}_{16\times16}$ has been examined according to relation (5), a third and final merging step checks whether each of the kept control point on the fine grid $\mathcal{G}_{8\times8}$ remains an unrestricted CP with its own MV or becomes a bound vertex. This decision is based on the neighborhood of each control point c_i in such a way, that, if one of the 4 immediate neighbors of c_i in $\mathcal{G}_{8\times8}$ is a fixed or so-called 'merged' control point represented by a bilinearly interpolated MV, c_i becomes a bound vertex. Note that bound vertices are not allowed to move freely but rather are collinear with two of its three neighboring grid points (*cf.* Fig. 3).

The result of this 3-step merging process is a segmentation into 8×8 and 16×16 square blocks where the motion parameters, *i.e.* the motion vectors, the decisions on the predicion type (warping or OBMC) and the texture interpolation type (as described below) have to be refined. This is done in a final estimation process by using the same type of estimation routines as described before in Sec. 2.2.

2.4 Loop Filter

A warping predictor utilizes a higher amount of fractional motion which leads to a strong lowpass effect when a bilinear pixel interpolation is employed. A high quality texture interpolation reduces the lowpass effect of the first order filtering of the bilinear texture interpolation. Hereby the peak signal-to-noise ratio (PSNR) can be increased up to 1.0–1.5 dB, or equivalently, bit savings up to 30% can be achieved compared to bilinear pixel interpolation [8].

Best results were obtained by using the predicted pixel at half or quarter pel position next to the unknown value for prediction. Tensor products of one-dimensional 10-tap hamming weighted sinc interpolation filters with pre-stored coefficients are used to compute the predictions at quarter pel positions.

As this interpolation scheme emphasizes the high frequency image components the predicted image gets much sharper and less high frequency coefficients must be coded. Unfortunately, also unavoidable coding artifacts are more visible. Therefore, a fully non-block based coding scheme such as the proposed warping predictor and the wavelet-based residual coder should be used to avoid any blocking artifacts. In order to also reduce the ringing effects a block-wise adaptive interpolation scheme is employed. In this scheme, after the motion vectors are estimated it is adaptively decided for each square block whether to use the high quality or the bilinear texture interpolation. The MSE is used as a criterion for the decision, which leads to a slightly increased PSNR and an improved visual quality. This scheme can also be viewed as an adaptive loop filter. As additional side information, one bit per block is transmitted indicating the type of loop filter.

3. Wavelet-Based Coding

Encoding of the motion compensated P-frames as well as of initial intraframes (I-frames) is performed by means of a conventionally structured transform coding scheme. First, a wavelet transform is applied to an entire frame. Uniform scalar quantization with a central dead-zone around zero similar to that designed for H.263 is then used to map the dynamic range of wavelet coefficients to a reduced alphabet of decision levels. The actual statistical coding of the quantized wavelet coefficients is complemented by a 3-stage pre-coding process based on the PACC coding principle. In this section, we give a brief description of this coding framework along with a discussion of some issues concerning construction and choice of an appropriate family of wavelet filters.

3.1 Biorthogonal Wavelet Bases Associated with Recursive Filters

Orthogonal wavelet bases, associated with recurcive filters, were investigated in detail by Herley and Vetterly [10]. In video and image processing biorthogonal pairs of bases consisting of functions having odd or even symmetry are usually more effective than orthogonal non-symmetric wavelet bases. For example, in wavelet-based image compression the so-called 9/7-wavelet with compact support, which was constructed in [5], is most frequently used. The wavelet bases proposed in [10], which combine orthogonality and symmetry, suffer from the fact that their adaptation to a concrete transform coding application is rather problematic due to inherently severe constraints. To overcome this problem, we propose an alternative approach.

The standard approach of constructing a dual pair of biorthogonal bases consists in reducing this problem to finding solutions of the matrix equation

$$M(z)\tilde{M}^T(z^{-1}) = 2I, \tag{6}$$

where I is the identity matrix, and

$$M(z) = \begin{pmatrix} h(z) & h(-z) \\ g(z) & g(-z) \end{pmatrix}, \quad \tilde{M}(z) = \begin{pmatrix} \tilde{h}(z) & \tilde{h}(-z) \\ \tilde{g}(z) & \tilde{g}(-z) \end{pmatrix}$$

are so-called 'modulation matrices', whose components in our case are rational functions, satisfying the normalizing conditions $h(1) = \tilde{h}(1) = \sqrt{2}$. Here, and in the following, filters h and g denote low-pass and high-pass filters of the decomposition algorithm, respectively, while $\tilde{h}$ and $\tilde{g}$ denote the corresponding filters for reconstruction.

In the polynomial case, the relations

$$g(z) = z^{-1}\tilde{h}(-z^{-1}),\ \tilde{g}(z) = z^{-1}h(-z^{-1})$$

necessarily hold. However, for rational solutions of Eq. (6) the last conditions can be violated. We consider an one-parameter family of filters

$$h(z) = \frac{1}{\sqrt{2}}(1+z), \tag{7}$$

$$\tilde{h}(z) = \frac{(2+a)(z^{-1}+3+3z+z^2)(z^{-1}+b+z)}{4\sqrt{2}(2+b)(z^{-2}+a+z^2)}, \tag{8}$$

$$g(z) = \frac{(2+a)(z^{-1}-3+3z-z^2)(-z^{-1}+b-z)}{4\sqrt{2}(2+b)}, \tag{9}$$

$$\tilde{g}(z) = \frac{1}{\sqrt{2}}\frac{1-z^{-1}}{z^{-2}+a+z^2}, \tag{10}$$

where $b = \frac{4a-8}{6-a}, \quad |a| > 2, \quad a \neq 6$. In the case of $a = 6$ we get a more simple pair of filters

$$h(z) = \frac{1}{\sqrt{2}}(1+z), \quad \tilde{h}(z) = \sqrt{2}\frac{z^{-1}+3+3z+z^2}{z^{-2}+6+z^2}.$$

The case $|a| \leq 2$ leads to unstable recursive filters because in this case the absolute value of roots of the denominator equals 1.

We note that the choice of the filters g and $\tilde{g}$ in Eqs. (9,10) is unusual since denominator $z^{-2}+a+z^2$ has moved from g to $\tilde{g}$. From the point of view of wavelet terminology, this modification does not lead neither to a change of the dual pair of 'multiresolution analyses' nor to a change of wavelet spaces. However, it modifies the underlying wavelet bases.

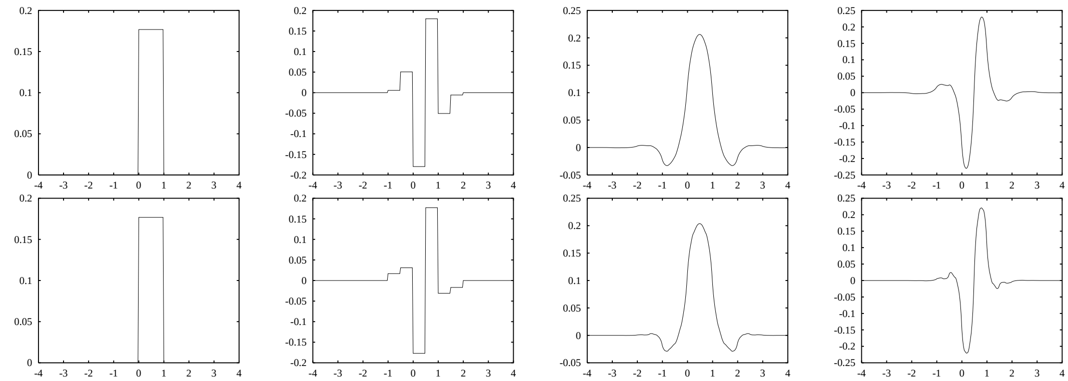


Fig. 4. From left to right: scaling function of analysis, analyzing wavelet, scaling function of synthesis, and synthesizing wavelet used for I-frame coding (top row) and P-frame coding (bottom row).

The nature and statistics of the images to be transformed is different for intra- and interframes. In intraframe mode we have to encode ordinary still images whereas interframe mode deals with prediction residual images (P-frames). Thus the bases for both modes were optimized separately. Our numerical simulations gave optimal values $a = 8$ for intraframe and $a = 25$ for interframe mode in relations (7)–(10). Graphs of the optimal basis functions are presented in Fig. 4.

3.2 Review of the PACC Pre-Coding Framework

For encoding the quantized wavelet coefficients, we follow the conceptual ideas presented in [13]. Next, we give a brief review of the involved techniques. For more details, the readers are referred to [13], [14].

As shown in the block diagram of Fig. 5, an initial 'partitioning' stage divides each frame of quantized coefficients into three sub-sources: a significance map, indicating the position of significant coefficients, a magnitude map, holding the absolute values of significant coefficients, and a sign map with the phase information of the wavelet coefficients. Note, that all three sub-sources inherit the subband structure from the quantized wavelet decomposition, so that there is another partition of each subsource according to the given subband structure.

In a second stage, the pre-coder performs an 'aggregation' of insignificant coeffi-

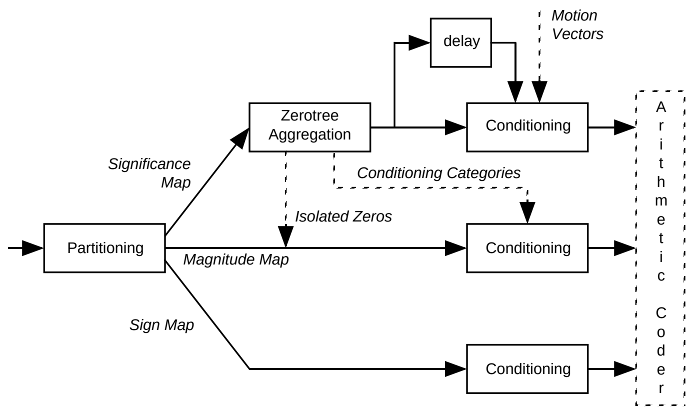


Fig. 5. Schematic representation of the PACC pre-coder.

cients using the 'zerotree' related data structure [12], [13] connecting insignificant wavelet coefficients which share the same spatial location across different scales of the octave-band decomposition. Note however, that in contrast to other zerotree-based coding methods, only zerotree roots localized in the lowest frequency bands are considered which guarantee a sufficient coding benefit.

The final 'conditioning' stage of the pre-coder supplies the elements of each source with a 'context', *i.e.* an appropriate model for the actual coding process in the arithmetic coder. Heuristically designed prototype templates are used for conditioning of elements of the significance map. Typically, they consist of two parts, where the first part is a causal neighborhood of the actual significance state c depending on scale and orientation of a given band. Except for the lowest frequency bands, the second part of the template utilizes additional information of the next upper level, *i.e.* significance information on the next coarser scale, which is given by the neighbors of the parent of the actual significance state c, thus allowing a "prediction" of the non-causal neighborhood of c. The processing of subbands is performed in the order from lowest to highest frequency bands, and the partitioned data of each band is processed such that the significance information is coded (and decoded) first.

This permits the construction of special conditioning states for the coding of magnitudes using the local significance information. Thus, the actual conditioning of magnitudes is performed by classifying significant coefficients according to the local variance estimated by the significance of their 8-neighborhood. For conditional coding of sign information a higher-order Markov model is used whose states are built of two preceding signs of a given sign event with respect to the orientation of the related band [13], [14].

Motion model parameters and all symbols generated by the pre-coder are encoded using an adaptive arithmetic coder [22]. For a fast adaptation of the models to the actual statistics, the non-binary symbols like magnitudes of coefficients or motion vector components are first mapped to binary symbols with length proportional to their expected probability distribution. For intra- and interframe coding separate models are used. Consecutive P-frames as well as consecutive motion vector fields are encoded using the updated related models of the previous P-frame and motion vector field, respectively.

4. Experimental Results

Simulations were carried out using the QCIF test sequences *Container*, *News* and *Foreman* (30 Hz, 300 frames, 176×144 pels). As a reference system we used the state-of-the-art test model TMN10 [4] of the ITU-T H.263+ [3] standardization project. The distinctive feature of TMN10 compared to its predecessors consists of an enhanced encoding strategy which rather follows a rate-distortion optimizing framework than relying on simple threshold decisions in the motion estimation stage [20]. There are several modes of operation permitted by a H.263 compliant encoder as defined by a number of negotiable options in the annexes of H.263. For our experiments using the TMN10 coder, the unrestricted motion vector mode (Annex D), the advanced prediction mode (Annex F), the improved intra coding mode (Annex I), the deblocking filter mode (Annex J), and the modified quantization mode (Annex T) have been enabled. For each run on a whole sequence of both the TMN10 reference coder and our proposed coding scheme, the quantization step-size of intra- and interframe coding was fixed. Note, that all coding results were generated from decoded bit-streams.

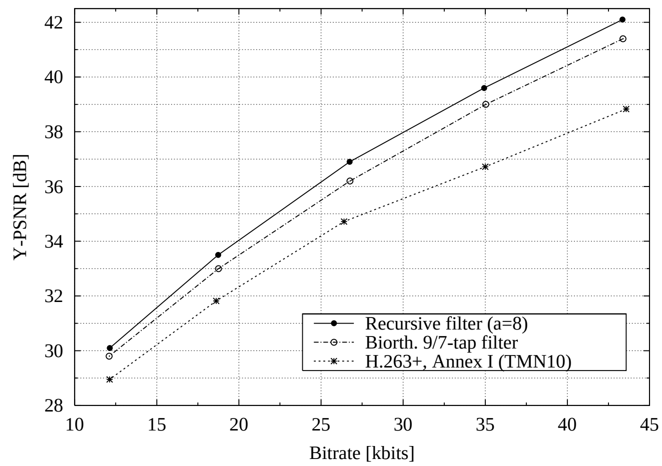


Fig. 6. Y-PSNR [dB] over bit-rate [kbits] obtained by our proposed method with two different filter choices and the TMN10 coder using the first frame of the QCIF *News* sequence.

4.1 Intraframe and Still Image Coding Performance

In this section, we report on the experimental performance evaluation of the intraframe coding method of our wavelet-based coding scheme. Fig. 6 shows the results for encoding the first frame of the *News* sequence in intra-mode. As can be seen from the graph, a coding gain of 0.3–0.7 dB PSNR on the luminance (Y) component was achieved by our proposed coding scheme supplied with the optimal choice of recursive filters ($a = 8$) compared to the same scheme using the biorthogonal 9/7-tap filter [5]. By using the recursive filter we also obtained visible improvements in reconstruction quality due to less ringing artifacts, especially at very low bit-rates.

Compared to the advanced intra coding mode (Annex I) of H.263+, our proposed new scheme showed a consistently better performance for different test sequences at various bit-rates. As shown in Fig. 6 for the first frame of the *News* sequence, a PSNR gain of 0.7–3.0 dB was achieved for the luminance component where higher coding gains were obtained at higher bit-rates. Subjectively, comparing the reconstructed I-frames of our proposed coder to those of TMN10 at the same bitrate, we observed an improved quality in favor of our proposed scheme especially at low and medium rates where the reconstructions of our coder were much sharper and free of any

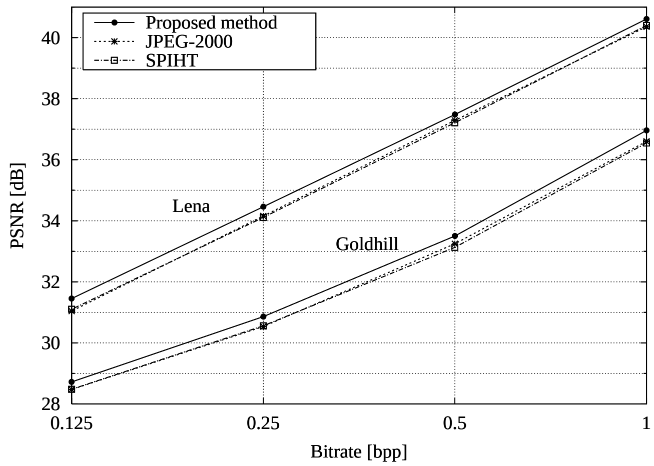


Fig. 7. PSNR [dB] over bit-rate [bpp] obtained by our proposed method and the wavelet-based still image coding schemes SPIHT and JPEG-2000 using the standard grayscale test images *Lena* and *Goldhill* of size 512×512 pels.

blocking artifacts.

Another experiment aimed at evaluating the rate-distortion (R-D) performance of our proposed method in the domain of still image coding. For this purpose, we used as reference systems the SPIHT-coder of Said and Pearlman [18] and the Verification Model (VM, Version 6.0) [1] of the ISO still image standardization project JPEG-2000, where the latter was driven in single-layer lossy coding mode with default parameters. For a fair comparison of the different wavelet-based coding methods, both reference schemes and our proposed coding scheme were operating with the same 9/7-tap filter-bank of [5].

In Fig. 7, coding results for the standard grayscale images *Lena* and *Goldhill* are presented showing that our proposed method achieved PSNR improvements of 0.2–0.5 dB over a bit-rate ranging from 0.125 to 1.0 bpp when comparing to state-of-the-art still image coders.

Table 1 provides additional coding results which give evidence of the fact that the superior coding efficiency of our intra coding method is not confined to image material of special kind. Compared to the emerging JPEG-2000 still image standard our coding method achieved an average gain in PSNR of approx. 0.5 dB for a

TABLE 1

RATE [bpp] *vs.* PSNR [dB] OBTAINED BY THE PROPOSED METHOD AND THE JPEG-2000 VM FOR A SET OF JPEG-2000 TEST IMAGES.

Image	**Rate [bpp]**	**Prop. method**	**JPEG-2000**	**Gain [dB]**
Hotel	1.0	38.74	38.48	0.26
720×576 pels	0.5	34.58	34.15	0.43
8 bpp	0.25	30.71	30.29	0.42
Bike	1.0	38.39	38.13	0.26
2048×2560 pels	0.5	33.71	33.55	0.16
8 bpp	0.25	29.79	29.65	0.14
CT	0.5	55.63	55.17	0.46
512×512 pels	0.25	48.94	48.28	0.66
12 bpp	0.125	42.80	41.88	0.92
Landsat	2.0	31.54	30.66	0.88
1024×1024 pels	1.0	25.95	25.43	0.52
8 bpp	0.5	22.65	22.23	0.42
			Average gain [dB] :	**0.46**

test set with pictures of different type (landscape, portrait, medical, satellite, etc.), resolution (up to 5 Megapixels) and bit depth (8–12 bpp).

4.2 Performance Evaluation of Different Motion Models

In order to compare the prediction capabilities of the different motion models, an experiment was performed using the *Foreman* sequence such that the last original frame was used for motion estimation and compensation. Four different motion models have been tested, i.e. bilinear warping, affine warping (by subdividing each square into a top left and bottom right triangle), OBMC and BMC (block motion compensation). For the latter two in addition to the iterative n-step search algorithm an integer pel accuracy full search followed by testing the half-pel positions of the 8-neighborhood of the best candidate vector was performed (*obmc_fs*, *bmc_fs*). Note, that for all tests (with exception of those related to the full search method)

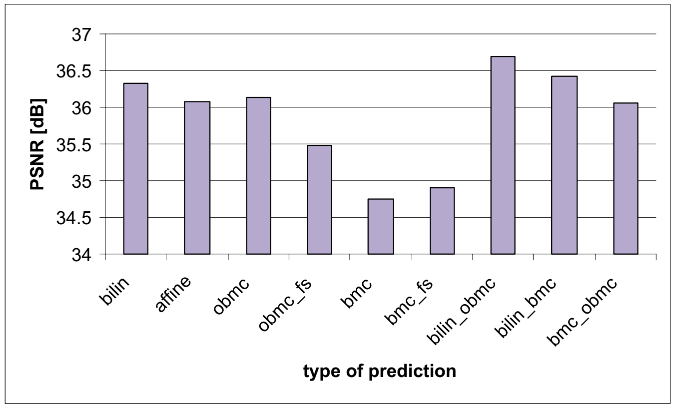


Fig. 8. Averaged PSNR [dB] of the luminance prediction versus type of prediction using the QCIF *Foreman* sequence at the original frame rate of 30 Hz.

the same motion vector positions (control points, *cf.* Fig. 1 (a)), regular grid or block size of 16×16 pels and iterative motion estimation technique was employed. The only difference was the underlying motion model during estimation and compensation. In the graph of Fig. 8, the (average) PSNR of the luminance prediction is plotted against the prediction type. As can be seen from the graph, the bilinear warping model performs best followed by OBMC, affine warping and BMC. The full search strategy (*fs*) leads to a slightly better prediction for BMC compared to the iterative search method, whereas OBMC performs much better when combined with the iterative motion estimation scheme. So the motion estimation technique originally developed for warping prediction also performs well for OBMC. By using the proposed combination of bilinear warping and OBMC (*bilin_obmc*) the prediction can be further improved. Combining OBMC with BMC or bilinear warping with BMC (*bmc_obmc*, *bilin_bmc*) leads to worse predictions.

4.3 Combined Intra- and Interframe Coding Performance

In the third and final part of our experiments, we evaluated the performance of our full video coding scheme using the three QCIF sequence. For this purpose, the original sequences were temporally subsampled to yield a target frame rate of 10 Hz, and only the first frame was coded in intra-mode. Figures 9 and 10 show the rate-distortion graphs as a result of these coding experiments. The R-D-curves were

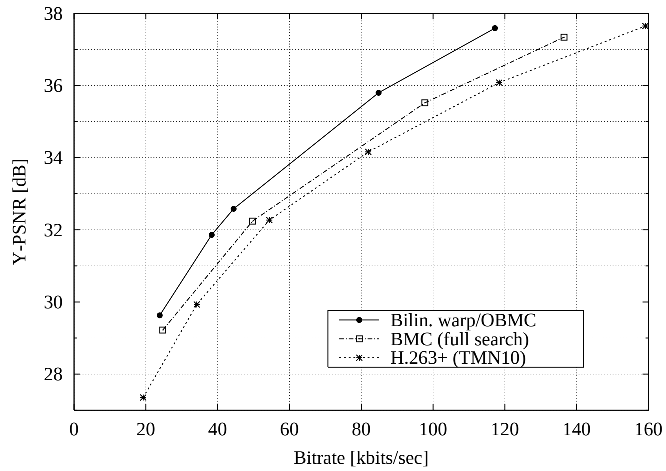


Fig. 9. Averaged Y-PSNR [dB] over bit-rate [kbits/sec] using the QCIF *Foreman* sequence at a frame rate of 10 Hz.

plotted using averaged Y-PSNR over bit-rate, where the latter was calculated as the arithmetic mean of total bits per frame (including the first I-frame) multiplied by the desired frame rate of 10 Hz. Note, that the H.263+ test model TMN10 was operating with Annexes D, F, I, J, and T.

Compared to the TMN10 reference coder, our proposed coding scheme using the combined motion model of warping prediction and OBMC achieved a gain in average PSNR of 1.0–1.75 dB for the very active *Foreman* sequence (*cf*. Fig. 9). By using simple block motion compensation (BMC) with full search in combination with our proposed wavelet-based coding method, only marginal advantages of 0.25–0.5 dB PSNR over the TMN10 coder were obtained. Figures 10 and 11 show the results for our experiments using the *Container* and *News* sequence, respectively. Employing a fixed regular control grid of size 16×16 pels for the combined warping/OBMC predictor a gain of 1.25–1.7 and 0.7–1.3 dB in average PSNR relative to the reference scheme was achieved for *Container* and *News*, respectively. Our proposed coding scheme supplied with the variable, two-level quadtree control grid yielded an additional PSNR gain of approx. 0.5 dB for *Container* and 1 dB for *News*.

Furthermore, the visual quality of the reconstructed frames appeared to be much improved. However, at low bit-rates the reconstructed frames of our proposed

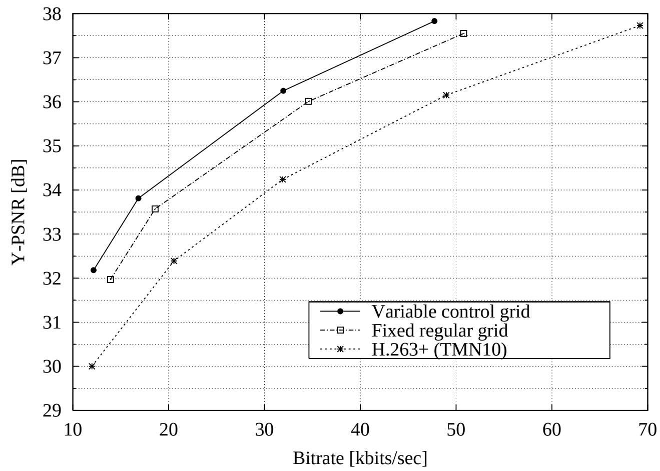


Fig. 10. Averaged Y-PSNR [dB] over bit-rate [kbits/sec] using the QCIF *Container* sequence at a frame rate of 10 Hz.

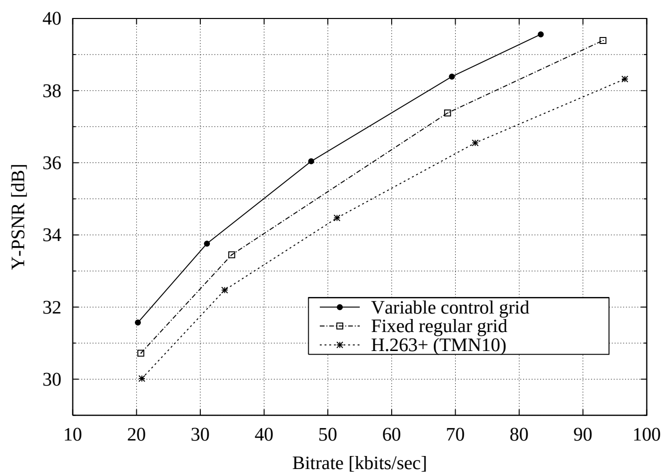


Fig. 11. Averaged Y-PSNR [dB] over bit-rate [kbits/sec] using the QCIF *News* sequence at a frame rate of 10 Hz.

scheme were still suffering from some ringing noise. We are currently investigating how to reduce these artifacts by incorporating a local intra coding mode and by studying the influence of different choices of wavelet filters.

5. Conclusions

A new blocking artifact free video coding scheme combining warping based prediction, overlapped motion compensation and wavelet-based residual coding has been presented. The proposed algorithm uses the key techniques of an adaptive grid partition with variable density and an adaptive texture interpolation method in the temporal predictor. Switching between the image warping model and the overlapped block motion compensation allows to deal efficiently with the problem of motion discontinuities. Intraframe and residual frame coding is performed using a specifically tailored class of biorthogonal recursive wavelet filters along with the efficient coding strategy of partitioning, aggregation and context-based conditional coding. Overall, the proposed coding method has proven to yield significant coding gains both with respect to objective and subjective evaluation. Speeding up the motion estimation is subject to our future research.

Acknowledgment

Parts of this work were supported by Deutsche Telekom Berkom GmbH, Germany, and by Russian Foundation for Basic Research under Grant No. 97-01-00443.

References

[1] ISO/IEC JTC 1/SC 29/WG 1, 'JPEG 2000 Verification Model Version 6.0', Document N 1575, Jan. 2000.

[2] ITU-T Recommendation H.263 'Video Coding for Low Bit-Rate Communication', Version 1, Mar. 1996.

[3] ITU-T Recommendation H.263 (Version 2) 'Video Coding for Low Bit-Rate Communication', Jan. 1998.

[4] ITU-T Study Group 16, Question 15, 'Video Codec Test Model Near Term, Version 10, TMN10 (Draft 1)', Document Q15-D-65, Apr. 1998.

[5] A. COHEN, I. DAUBECHIES and J.-C. FEAUVEAU, 'Biorthogonal Bases of Compactly Supported Wavelets', *Comm. on Pure and Appl. Math.*, 1992, **45**, pp. 485–560.

[6] G. CÔTÉ, B. EROL, M. GALLANT, and F. KOSSENTINI, 'H.263+: Video Coding at Low Bit-Rates', *IEEE Trans. on Circuits and Systems for Video Techn.*, 1998, **8** (7) pp. 849–866.

[7] G. HEISING and G. RUHL, 'Video Coding Using Spatial Extrapolation Based Motion Field Segmentation', *IEEE Int. Conf. Image Proc.*, 1996, **2**, pp. 482–485.

[8] G. HEISING, 'Blocking Artifact Free Video Coding by Combining Warping Based Prediction with Wavelet Error Coding', *Proc. Picture Coding Symposium (PCS) '97*, 1997, pp. 309–314.

[9] G. HEISING, D. MARPE and H. L. CYCON: 'A Wavelet-Based Video Coding Scheme Using Image Warping', *Proc. Int. Conf. Image Proc. (ICIP) 1998*, 1998.

[10] C. HERLEY and M. VETTERLI, 'Wavelet and Recursive Filter Banks', 1993, *IEEE Trans. on Signal Proc.* **41** (8) pp. 2536–2556.

[11] C.-L. HUANG and C.-Y. HSU, 'A New Motion Compensation Method for Image Sequence Coding Using Hierarchical Grid Interpolation', *IEEE Trans. on Circuits and Systems for Video Techn.*, 1994, **4** (1), pp. 42–51.

[12] A. LEWIS and G. KNOWLES, 'Image Compression Using the 2D Wavelet Transform', *IEEE Trans. on Image Proc.*, **1** (2), 1992, pp. 244–250.

[13] D. MARPE and H. L. CYCON, 'Efficient Pre-Coding Techniques for Wavelet-Based Image Compression', 1997, *Proc. PCS '97*, pp. 45–50.

[14] D. MARPE and H. L. CYCON, 'Very Low Bit-Rate Video Coding Using Wavelet-Based Techniques', *IEEE Trans. on Circ. and Sys. for Video Techn.*, 1999, **9** (1), pp. 85–94.

[15] D. MARPE, G. HEISING, H. L. CYCON and A. P. PETUKHOV: 'Video Coding Using A Bilinear Image Warping Model and Wavelet- Based Residual Coding', *Proc. SPIE* Vol. 3813, Wavelet Application in Signal and Image Processing VI, Juli 1999, pp. 401-408.

[16] D. MARPE, G. BLÄTTERMANN, J. RICKE, and P. MAASS, 'A Two-Layered Wavelet-Based Algorithm for Efficient Lossless and Lossy Image Compression', to appear in *IEEE Trans. on Circ. and Sys. for Video Techn.*.

[17] A. P. PETUKHOV, 'Biorthogonal Wavelet Bases with Rational Masks and Their Application', Preprint 1998.

[18] A. SAID and W. A. PEARLMAN, 'A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees', *IEEE Trans. on Circ. and Sys. for Video Techn.*, **6** (3), 1996, pp. 243–250.

[19] G. J. SULLIVAN and R. L. BAKER, 'Motion Compensation for Video Compression Using Control Grid Interpolation', *Proc. Int. Conf. ASSP*, Toronto, Canada, May 1991, pp. 2713–2716.

[20] G. J. SULLIVAN and T. WIEGAND, 'Rate-Distortion Optimization for Video Compression', *IEEE Signal Processing Mag.*, 1998, **15**, pp. 74–90.

[21] H. WATANABE and S. SINGHAL, 'Windowed Motion Compensation', *Proc. SPIE Conf. Visual Comm. Image Proc.*, **1605**, 1991, pp. 582–589.

[22] I. WITTEN, R. NEAL and J. CLEARY, 'Arithmetic Coding for Data Compression', *Comm. ACM*, **30**, 1987, pp. 520–540.

[23] G. WOLBERG, 'Digital Image Warping', (IEEE Computer Society Press, Los Alamitos, CA, 1990.)

List of Figures

List of Tables