A NEW BIDIRECTIONALLY MOTION-COMPENSATED ORTHOGONAL TRANSFORM FOR VIDEO CODING

Markus Flierl and Bernd Girod

Max Planck Center for Visual Computing and Communication
Stanford University, Stanford, CA 94305
mflierl@stanford.edu

ABSTRACT

Motion-compensated lifted wavelets have received much interest for video compression. While they are biorthogonal, they may substantially deviate from orthonormality due to motion compensation, even if based on an orthogonal or near-orthogonal wavelet. A temporal transform for video sequences that maintains orthonormality while permitting flexible motion compensation would be very desirable. We have recently introduced such a transform for unidirectional motion compensation from one previous frame. In this paper, we extend this idea to bidirectional motion compensation. Orthonormality is maintained for arbitrary integer-pixel motion compensation by cascading a sequence of incremental orthogonal 3×3 transforms. The energy of three input pictures is accumulated in two temporal low-bands while the temporal high-band is zero if the input pictures are identical after motion compensation. Further, the motion-compensated orthogonal transforms can be cascaded to build a dyadic wavelet decomposition. The new bidirectionally motion-compensated orthogonal transform compares favorably with the lifted 5/3 wavelet in video coding experiments with integer-pixel motion compensation.

Index Terms— Temporal subband coding of video, orthogonal transforms, bidirectional motion compensation, motion-compensated orthogonal transforms.

1. INTRODUCTION

Well known methods for representing image sequences for coding and communication applications are standard hybrid video coding techniques as well as motion-compensated subband coding schemes. To achieve high compression efficiency, standard hybrid video encoders operate in a closed-loop fashion such that the total distortion across the reconstructed pictures equals the total distortion in the corresponding intra picture and encoded displaced frame differences. In case of channel errors, decoded reference frames differ from the optimized reference frames at the encoder and error propagation is observed. On the other hand, transform coding schemes operate in an open-loop fashion. Consider high-rate transform coding schemes in which the analysis transform produces independent transform coefficients. With uniform quantization, these schemes are optimal when utilizing an orthogonal transform [1]. Further, Parseval's relation holds for orthogonal transforms such that the total quantization distortion in the coefficient domain equals that in the image domain. In case of channel errors, the error energy in the image domain equals that in the coefficient domain. Hence, the error energy is preserved in the image domain and is not amplified by the decoder, as is the case, e.g., for predictive decoders.

During the last decade, there have been attempts to incorporate motion compensation into temporal subband coding schemes [2, 3, 4, 5] by approaching problems arising from multi-connected pixels. In [6], we propose a unidirectionally motion-compensated orthogonal transform that strictly maintains orthogonality for any motion field. The transform is factored into a sequence of incremental transforms that are strictly orthogonal. The incremental transforms maintain scale counters to keep track of the scale factors that are introduced to ensure orthogonality. The decorrelation factor of each incremental transform is determined by the scale counters and is chosen such that the transform meets an energy-concentration constraint. The experiments show that this orthogonal transform offers an improved energy compaction when compared to motion-compensated lifted Haar wavelets and closed-loop hierarchical P pictures.

In contrast to our previous work in [6] where only unidirectional motion compensation considered, this paper extends the approach to bidirectional motion compensation. The presented bidirectionally motion-compensated orthogonal transform is able to consider up to two motion fields per frame. Similar to our previous work, we factor the transform into a sequence of incremental transforms which are strictly orthogonal. The incremental transforms maintain scale counters that are compatible with the scale counters in [6]. The decorrelation factors of each incremental transform are determined such that an energy-concentration constraint is met for bidirectional motion compensation.

The paper is organized as follows: Section 2 introduces the bidirectionally motion-compensated orthogonal transform and discusses the incremental transform as well as the energy-concentration constraint. Section 3 proposes a method to incorporate this transform into a dyadic decomposition for groups of pictures. Section 4 presents the experimental results.

2. BIDIRECTIONALLY MOTION-COMPENSATED ORTHOGONAL TRANSFORM

This section discusses how the transform is factored into incremental transforms. We outline the construction of the incremental transform and the incorporation of the energy-concentration constraint.

Let $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ be three vectors representing consecutive pictures of an image sequence. The transform T maps these vectors according to

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \mathbf{y}_3 \end{pmatrix} = T \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{pmatrix} \tag{1}$$

into three vectors $\mathbf{y}_1$, $\mathbf{y}_2$, and $\mathbf{y}_3$ which represent the first temporal low-band, the high-band, and the second temporal low-band, respectively. We factor the transform T into a sequence of k incremental

transforms T_κ such that

$$T = T_k T_{k-1} \cdots T_\kappa \cdots T_2 T_1, \tag{2}$$

where each incremental transform T_κ is orthogonal by itself, i.e., $T_\kappa T_\kappa^T = I$ holds for all $\kappa = 1, 2, \cdots, k$. This guarantees that the transform T is also orthogonal.

2.1. Incremental Transform

Let $\mathbf{x}_1^{(\kappa)}$, $\mathbf{x}_2^{(\kappa)}$, and $\mathbf{x}_3^{(\kappa)}$ be three vectors representing consecutive pictures of an image sequence if $\kappa = 1$, or three output vectors of the incremental transform $T_{\kappa-1}$ if $\kappa > 1$. The incremental transform T_κ maps these vectors according to

$$\begin{pmatrix} \mathbf{x}_1^{(\kappa+1)} \\ \mathbf{x}_2^{(\kappa+1)} \\ \mathbf{x}_3^{(\kappa+1)} \end{pmatrix} = T_\kappa \begin{pmatrix} \mathbf{x}_1^{(\kappa)} \\ \mathbf{x}_2^{(\kappa)} \\ \mathbf{x}_3^{(\kappa)} \end{pmatrix} \tag{3}$$

into three vectors $\mathbf{x}_1^{(\kappa+1)}$, $\mathbf{x}_2^{(\kappa+1)}$, and $\mathbf{x}_3^{(\kappa+1)}$ which will be further transformed into the first temporal low-band, high-band, and second temporal low-band, respectively.

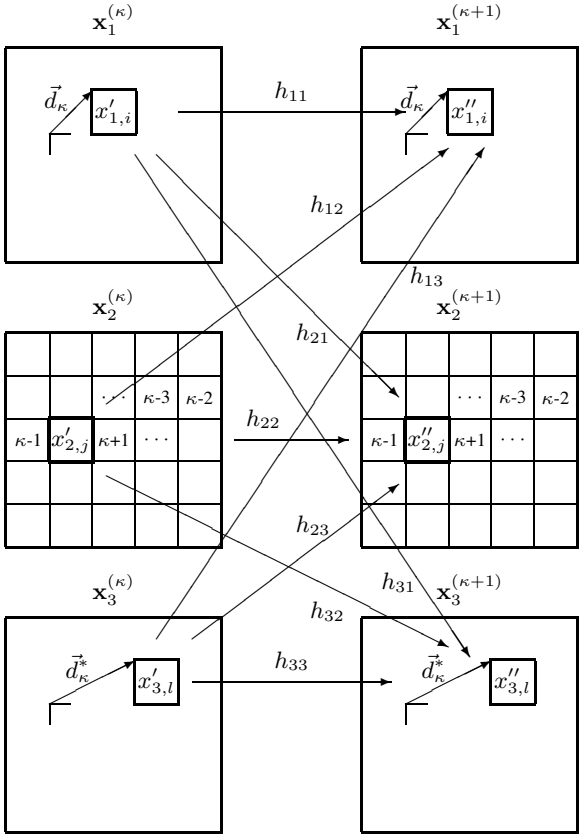

Fig. 1. The incremental transform T_κ for the three frames $\mathbf{x}_1^{(\kappa)}$, $\mathbf{x}_2^{(\kappa)}$, and $\mathbf{x}_3^{(\kappa)}$ which strictly maintains orthogonality for any bidirectional motion field $(\vec{d}_\kappa, \vec{d}^*_\kappa)$. T_κ minimizes the energy in $x_{2,j}$.

Fig. 1 depicts the process accomplished by the incremental transform T_κ with its input and output images as defined above. The incremental transform removes the energy of the j-th pixel $x'_{2,j}$ in the image $\mathbf{x}_2^{(\kappa)}$ with the help of both the i-th pixel $x'_{1,i}$ in the image $\mathbf{x}_1^{(\kappa)}$ which is linked by the motion vector $\vec{d}_\kappa$ and the l-th pixel $x'_{3,l}$ in the image $\mathbf{x}_3^{(\kappa)}$ which is linked by the motion vector $\vec{d}^*_\kappa$ (or the j-th block with the help of both the i-th and the l-th block if all the pixels of the i-th and l-th block have the motion vectors $\vec{d}_\kappa$ and $\vec{d}^*_\kappa$, respectively). The energy-removed pixel value $x''_{2,j}$ is obtained by a linear combination of the pixel values $x'_{1,i}$, $x'_{2,j}$, and $x'_{3,l}$ with scalar weights h_{21}, h_{22}, and h_{23}. The energy-concentrated pixel value $x''_{1,i}$ is also obtained by a linear combination of the pixel values $x'_{1,i}$, $x'_{2,j}$, and $x'_{3,l}$ but with scalar weights h_{11}, h_{12}, and h_{13}. The energy-concentrated pixel value $x''_{3,l}$ is calculated accordingly. All other pixels are simply kept untouched.

To summarize, the incremental transform T_κ touches only pixels that are linked by the same motion vector pair $(\vec{d}_\kappa, \vec{d}^*_\kappa)$. Of these, T_κ performs only a linear combination with three pixels that are connected by this motion vector pair. All other pixels are kept untouched. This is reflected in the following matrix notation:

$$T_\kappa = \begin{pmatrix} \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots \\ \cdots & 0 & h_{11} & \cdots & 0 & h_{12} & \cdots & 0 & h_{13} & \cdots \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots & 0 & 0 & \cdots \\ \cdots & 0 & h_{21} & \cdots & 0 & h_{22} & \cdots & 0 & h_{23} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ \cdots & 0 & 0 & \cdots & 0 & 0 & \cdots & 1 & 0 & \cdots \\ \cdots & 0 & h_{31} & \cdots & 0 & h_{32} & \cdots & 0 & h_{33} & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix} \tag{4}$$

The diagonal elements equal to 1 represent the untouched pixels and the elements $h_{\mu\nu}$ represent the pixels subject to linear operations. All other entries are zero.

Now, the scalar weights $h_{\mu\nu}$ are arranged into the 3×3 matrix H. The incremental transform T_κ is orthogonal if H is also orthogonal. We construct an orthogonal H with the help of Euler's rotation theorem which states that any rotation can be given as a composition of rotations about three axes, i.e. $H = H_3 H_2 H_1$, where H_r denotes a rotation about one axes. We choose the composition

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} = \begin{pmatrix} \cos(\psi) & 0 & \sin(\psi) \\ 0 & 1 & 0 \\ -\sin(\psi) & 0 & \cos(\psi) \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} \cos(\phi) & 0 & \sin(\phi) \\ 0 & 1 & 0 \\ -\sin(\phi) & 0 & \cos(\phi) \end{pmatrix} \tag{5}$$

with the Euler angles ψ, θ, and ϕ. The Euler angles will be determined in the next subsection which discusses the energy concentration constraint.

Note that, to carry out the full transform T, each pixel in $\mathbf{x}_2$ is touched only once whereas the pixels in $\mathbf{x}_1$ and $\mathbf{x}_3$ may be touched multiple times or never. Further, the order in which the incremental transforms T_κ are applied does not affect the orthogonality of T, but it may affect the energy concentration of the transform T.

2.2. Energy Concentration Constraint

The three Euler angles for each pixel touched by the incremental transform have to be chosen such that the energy in image $\mathbf{x}_2$ is minimized. Consider the pixel triplet $x_{1,i}$, $x_{2,j}$, and $x_{3,l}$ to be processed

by the incremental transform T_κ. To determine the Euler angles for the pixel $x_{2,j}$, we assume that the pixel $x_{2,j}$ is connected to the pixels $x_{1,i}$ and $x_{3,l}$ such that $x_{2,j} = x_{1,i} = x_{3,l}$. Consequently, the resulting high-band pixel $x''_{2,j}$ shall be zero. Note that the pixels $x_{1,i}$ and $x_{3,l}$ may have been processed previously by T_τ, where $\tau < \kappa$. Therefore, let v_1 and v_3 be the *scale factors* for the pixels $x_{1,i}$ and $x_{3,l}$, respectively, such that $x'_{1,i} = v_1 x_{1,i}$ and $x'_{3,l} = v_3 x_{3,l}$. The pixel $x_{2,j}$ is used only once during the transform process T and no scale factor needs to be considered. But in general, when considering subsequent dyadic decompositions with T, scale factors are passed on to higher decomposition levels and, consequently, they need to be considered, i.e., $x'_{2,j} = v_2 x_{2,j}$. Obviously, for the first decomposition level, $v_2 = 1$. Let u_1 and u_3 be the scale factors for the pixels $x_{1,i}$ and $x_{3,l}$, respectively, after they have been processed by T_κ. Now, the pixels $x'_{1,i}$, $x'_{2,j}$, and $x'_{3,l}$ are processed by T_κ as follows:

$$\begin{pmatrix} u_1 x_{1,i} \\ 0 \\ u_3 x_{1,i} \end{pmatrix} = H_3 H_2 H_1 \begin{pmatrix} v_1 x_{1,i} \\ v_2 x_{1,i} \\ v_3 x_{1,i} \end{pmatrix} \tag{6}$$

Energy conservation requires that

$$u_1^2 + u_3^2 = v_1^2 + v_2^2 + v_3^2. \tag{7}$$

The Euler angle ϕ in H_1 is chosen such that the two hypotheses $x'_{1,i}$ and $x'_{3,l}$ are weighted equally after being attenuated by their scale factors v_1 and v_3.

$$\tan(\phi) = -\frac{v_1}{v_3} \tag{8}$$

The Euler angle θ in H_2 is chosen such that it meets the zero-energy constraint for the high-band in (6).

$$\tan(\theta) = \frac{v_2}{\sqrt{v_1^2 + v_3^2}} \tag{9}$$

Finally, the Euler angle ψ in H_3 is chosen such that the pixels $x_{1,i}$ and $x_{3,l}$, after the incremental transform T_κ, have scalar weights u_1 and u_3, respectively.

$$\tan(\psi) = \frac{u_1}{u_3} \tag{10}$$

But note that we are free to choose this ratio. We have chosen the Euler angle ϕ such that the previous frame and the future frame have equal contribution after rescaling with v_1 and v_3. Consequently, we choose the scale factors u_1 and u_3 such that they increase equally.

$$u_1 = \sqrt{v_1^2 + \frac{v_2^2}{2}} \quad \text{and} \quad u_3 = \sqrt{v_3^2 + \frac{v_2^2}{2}} \tag{11}$$

Similar to the work in [6], we utilize *scale counters* to keep track of the scale factors. Scale counters simply count how often a pixel is used as reference for motion compensation. Before any transform is applied, the scale counter for each pixel is $n = 0$ and the scale factor is $v = 1$. For arbitrary scale counter n and m, the scale factors are

$$v = \sqrt{n+1} \quad \text{and} \quad u = \sqrt{m+1}. \tag{12}$$

After applying the incremental transform, the scale counter have to be updated for the modified pixels. For the unidirectionally motion-compensated orthogonal transform in [6], the updated scale counter for low-band pixels is given by $m = n_1 + n_2 + 1$, where n_1 and n_2 are the scale counters of the utilized input pixel pairs. For the bidirectionally motion-compensated orthogonal transform, the updated scale counters for low-band pixels result from (11) as follows:

$$m_1 = n_1 + \frac{n_2 + 1}{2} \quad \text{and} \quad m_3 = n_3 + \frac{n_2 + 1}{2} \tag{13}$$

For example, consider the transform in the first decomposition level where $n_2 = 0$. The unidirectionally motion-compensated transform increases the scale counter by 1 for each used reference pixel, whereas the bidirectionally motion-compensated transform increases the counter by 0.5 for each of the two used reference pixels.

3. DYADIC TRANSFORM FOR GROUPS OF PICTURES

The orthogonal transform in Section 2 is defined for three input pictures but generates two temporal low-bands. In combination with the orthogonal transform in [6], we are able to define an orthogonal transform with only one temporal low-band for groups of pictures whose number of pictures is larger than two and a power of two.

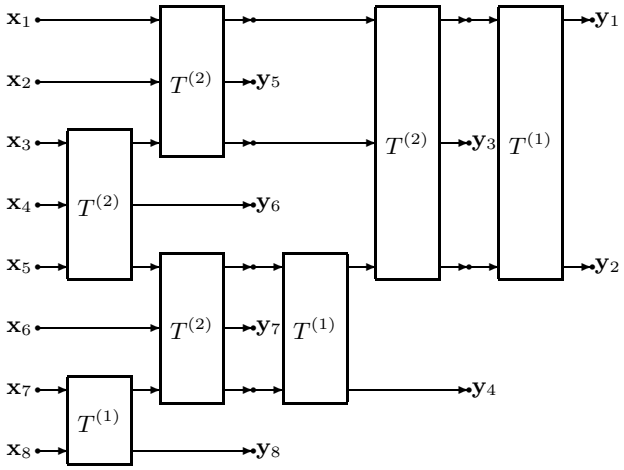

Fig. 2. Decomposition of a group of 8 pictures with orthogonal transforms $T^{(1)}$ and $T^{(2)}$.

Fig. 2 depicts a decomposition of a group of 8 pictures $\mathbf{x}_\rho$ into one temporal low-band $\mathbf{y}_1$ and 7 high-bands $\mathbf{y}_\rho$, $\rho = 2, 3, \ldots, 8$. $T^{(1)}$ denotes a unidirectionally motion-compensated orthogonal transform as presented in [6]. $T^{(2)}$ denotes a bidirectionally motion-compensated orthogonal transform as introduced in Section 2. Note that this architecture permits also block-wise decisions between unidirectional and bidirectional motion compensation. This adaptivity is used in the following experiments.

4. EXPERIMENTAL RESULTS

Experimental results assessing the energy compaction are obtained for the CIF sequences *Foreman*, *Bus*, and *Soccer*. Our coding scheme with the bidirectionally motion-compensated orthogonal transform is compared to schemes which use a motion-compensated lifted 5/3 wavelet with and without update step. In addition, the performance of the unidirectionally motion-compensated orthogonal transform [6] is reported.

For the coding process with the orthogonal transforms, a scale counter n is maintained for every pixel of each picture. The scale counters are an immediate results of the utilized motion vectors and are only required for the processing at encoder and decoder. The scale counters do not have to be encoded as they can be recovered from the motion vectors.

All schemes operate with a GOP size of 16 frames as well as with integer-pixel accurate motion compensation. The block size for motion compensation is limited to 8×8. Conditional motion estimation is used for bidirectional motion estimation. The same block

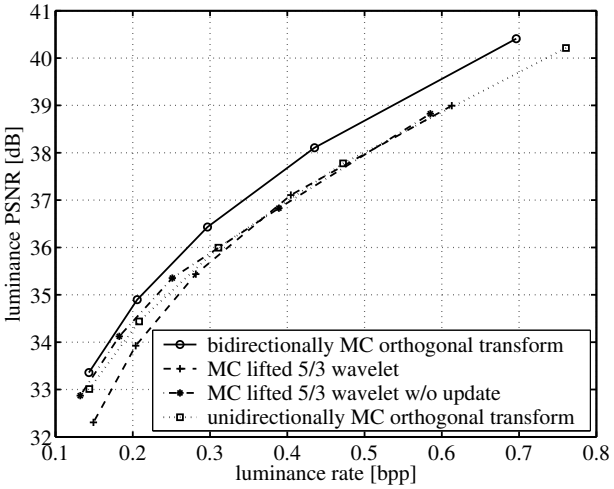

Fig. 3. PSNR over the rate for the luminance signal of the CIF sequence *Foreman* at 30 fps with 288 frames.

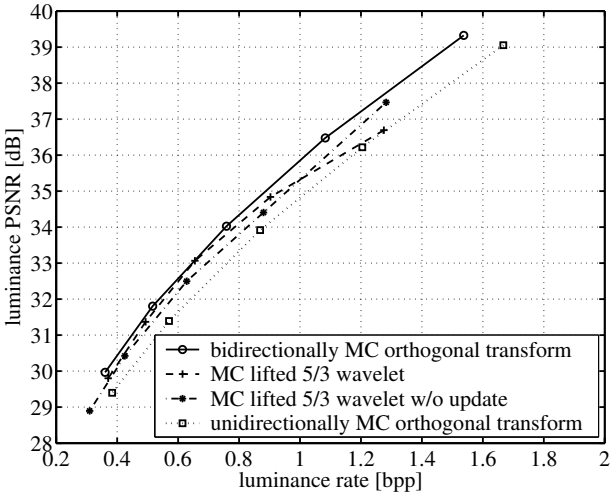

Fig. 4. PSNR over the rate for the luminance signal of the CIF sequence *Bus* at 30 fps with 64 frames.

motion fields are used for both orthogonal transform and 5/3 wavelet. For simplicity, the resulting temporal subbands are coded with JPEG 2000. The temporal high-bands are coded directly, whereas the temporal low-band is rescaled with (12) before encoding. Lagrangian costs are used for optimal rate allocation. Note that the scale factors of the temporal low-band are considered in the distortion term.

Figs. 3, 4, and 5 depict the rate distortion performances for the luminance signals of the test sequences. Results for the bidirectionally motion-compensated (MC) orthogonal transform, the MC lifted 5/3 wavelet with and without update step, as well as for the unidirectionally MC orthogonal transform [6] are given. In order to assess the energy compaction, no intra modes have been used for all temporal coding schemes. For all test sequences, the bidirectionally MC orthogonal transform outperforms the unidirectionally MC orthogonal transform. Further, the bidirectionally MC orthogonal transform compares favorably with the MC lifted 5/3 wavelet with and without update step.

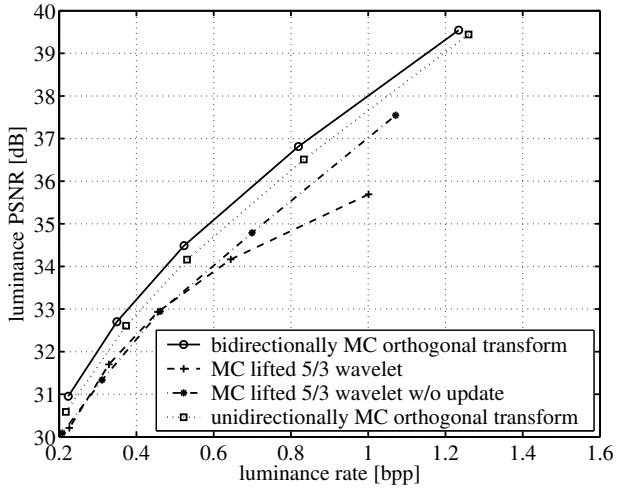

Fig. 5. PSNR over the rate for the luminance signal of the CIF sequence *Soccer* at 30 fps with 64 frames.

5. CONCLUSIONS

The paper presents a bidirectionally motion-compensated orthogonal transform which strictly maintains orthogonality for any bidirectional motion field. In terms of energy compaction, it outperforms the unidirectionally motion-compensated orthogonal transform and provides benefits over the motion-compensated lifted 5/3 wavelet. The current implementation considers integer-pixel accurate motion compensation. Applying the incremental orthogonal 3×3 transform to pairs of input frames will yield more accurate motion compensation. This is currently under investigation. In summary, the orthogonality principle can be successfully combined with efficient bidirectional motion compensation.

6. REFERENCES

[1] V.K. Goyal, "Theoretical foundations of transform coding," *IEEE Signal Processing Magazine*, vol. 18, no. 5, pp. 9–21, Sept. 2001.

[2] J.-R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559–571, Sept. 1994.

[3] S.-J. Choi and J.W. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.

[4] B. Pesquet-Popescu and V. Bottreau, "Three-dimensional lifting schemes for motion compensated video compression," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Salt Lake City, UT, May 2001, vol. 3, pp. 1793–1796.

[5] A. Secker and D. Taubman, "Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting," in *Proceedings of the IEEE International Conference on Image Processing*, Thessaloniki, Greece, Oct. 2001, vol. 2, pp. 1029–1032.

[6] M. Flierl and B. Girod, "A motion-compensated orthogonal transform with energy-concentration constraint," in *Proceedings of the IEEE Workshop on Multimedia Signal Processing*, Victoria, BC, Oct. 2006.