CODING OF MULTIVIEW IMAGERY WITH MOTION AND DISPARITY COMPENSATED ORTHOGONAL TRANSFORMS

Markus Flierl

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

ABSTRACT

We consider the problem of multiview video compression with recently introduced motion-compensated orthogonal video transforms. It is well known that so called motion-compensated lifted wavelets may substantially deviate from orthonormality due to motion compensation, even if based on an orthogonal or near-orthogonal wavelet. Therefore, inaccurate motion compensation may result in a motion-dependent suboptimal subband decomposition. On the other hand, motion-compensated orthogonal video transforms are designed such that orthonormality is strictly maintained for any motion field. Even inaccurate motion compensation leads to an orthogonal subband decomposition. This is of particular interest for complex motion and disparity fields. For subband coding of multiview video, motion-compensated subbands will be further processed by disparity-compensated subband decompositions. In case of suboptimal subband decompositions, this cascade of processing steps may result in further inefficiencies. On the other hand, motion or disparity compensated orthogonal transforms offer the advantage that the subbands are always orthogonal, even if motion or disparity compensation is inaccurate. This paper investigates the energy compaction of motion and disparity compensated orthogonal transforms.

Index Terms— Multiview video coding, motion-compensated orthogonal video transforms

1. INTRODUCTION

Capturing dynamic scenes with multiple video cameras is getting more and more attractive with today's advances in display and camera technology. Applications for multiview video signals range from free viewpoint video [1] to free viewpoint television (FTV) [2]. All of them require efficient coding of the multiview imagery.

For coding and transmission of dynamic scenes, statistical dependencies within the captured data have to be exploited. When capturing multiview video signals, disparities between views and motion between temporally successive frames are the most important parameters of the dynamic scene. To achieve a good trade-off between scene quality and bit-rate, the correlation among all the pictures has to be exploited efficiently. Usually, this is accomplished with either predictive or subband coding schemes.

Based on the video coding standard H.264/AVC [3], the Joint Video Team (JVT) is developing a Joint Multiview Video Model (JMVM) [4] for multiview video coding. It utilizes motion and disparity compensated prediction to exploit the correlation in temporal and view direction. The JMVM is a predictive coding scheme.

For subband coding of multiview video, we have developed an efficiency analysis of motion and disparity compensated coding [5]. We have found that high-rate performance bounds may be achieved with motion and disparity compensated orthogonal transforms. For video coding, we have developed unidirectionally [6] and bidirectionally [7] motion-compensated orthogonal video transforms. In this paper, we investigate multiview video coding with motion and disparity compensated orthogonal transforms.

The paper is organized as follows: Section 2 recaps the principles of motion-compensated orthogonal video transforms. We summarize the incremental transform and the energy concentration constraint, as well as the dyadic decomposition of groups of pictures for the bidirectionally compensated orthogonal transform. Section 3 discusses a decomposition structure for multiview video coding with motion and disparity compensated orthogonal transforms. Section 4 presents the experimental results assessing the energy compaction of motion and disparity compensated orthogonal transforms.

2. MOTION AND DISPARITY COMPENSATED ORTHOGONAL TRANSFORMS

In previous work we have developed unidirectionally [6] and bidirectionally [7] motion-compensated orthogonal video transforms. We recap the principle by summarizing briefly the bidirectionally motion-compensated orthogonal transform. Disparity-compensated orthogonal transforms operate in the same fashion.

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.

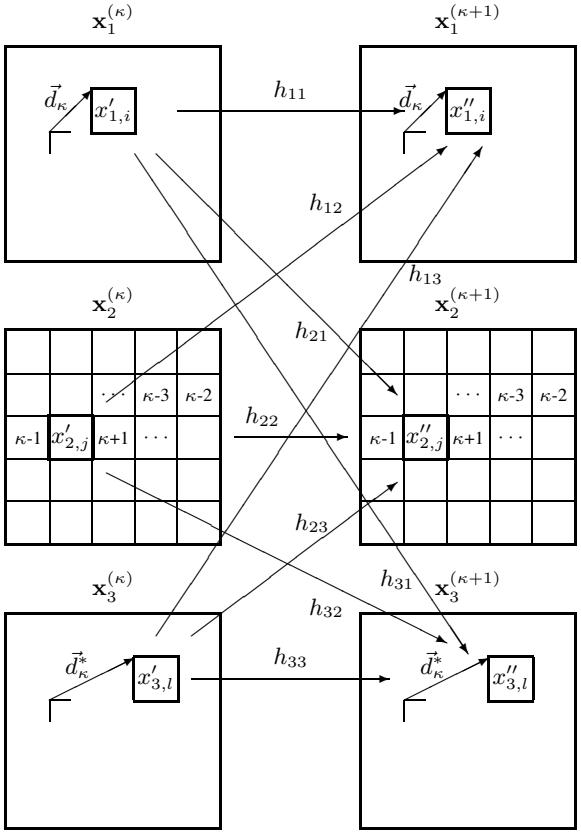

Fig. 1. The incremental transform T_κ for the three frames $\mathbf{x}_1^{(\kappa)}$, $\mathbf{x}_2^{(\kappa)}$, and $\mathbf{x}_3^{(\kappa)}$ which strictly maintains orthogonality for any bidirectional motion field $(\vec{d}_\kappa, \vec{d}^*_\kappa)$. T_κ minimizes the energy in $x_{2,j}$.

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.

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

$$H = \begin{pmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{pmatrix} = \begin{pmatrix} \cos(\psi) & 0 & \sin(\psi) \\ 0 & 1 & 0 \\ -\sin(\psi) & 0 & \cos(\psi) \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos(\theta) & -\sin(\theta) \\ 0 & \sin(\theta) & \cos(\theta) \end{pmatrix} \begin{pmatrix} \cos(\phi) & 0 & \sin(\phi) \\ 0 & 1 & 0 \\ -\sin(\phi) & 0 & \cos(\phi) \end{pmatrix} \tag{4}$$

with the Euler angles ψ, θ, and ϕ.

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:

$$\begin{pmatrix} u_1 x_{1,i} \\ 0 \\ u_3 x_{1,i} \end{pmatrix} = H_3 H_2 H_1 \begin{pmatrix} v_1 x_{1,i} \\ v_2 x_{1,i} \\ v_3 x_{1,i} \end{pmatrix} \tag{5}$$

Energy conservation requires that

$$u_1^2 + u_3^2 = v_1^2 + v_2^2 + v_3^2. \tag{6}$$

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.

$$\tan(\phi) = -\frac{v_1}{v_3} \tag{7}$$

The Euler angle θ in H_2 is chosen such that it meets the zero-energy constraint for the high-band in (5).

$$\tan(\theta) = \frac{v_2}{\sqrt{v_1^2 + v_3^2}} \tag{8}$$

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.

$$\tan(\psi) = \frac{u_1}{u_3} \tag{9}$$

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.

$$u_1 = \sqrt{v_1^2 + \frac{v_2^2}{2}} \quad \text{and} \quad u_3 = \sqrt{v_3^2 + \frac{v_2^2}{2}} \tag{10}$$

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

$$v = \sqrt{n+1} \quad \text{and} \quad u = \sqrt{m+1}. \tag{11}$$

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 (10) as follows:

$$m_1 = n_1 + \frac{n_2+1}{2} \quad \text{and} \quad m_3 = n_3 + \frac{n_2+1}{2} \tag{12}$$

2.3. Dyadic Transform for Groups of Pictures

The bidirectional transform is defined for three input pictures but generates two temporal low-bands. In combination with the unidirectional 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.

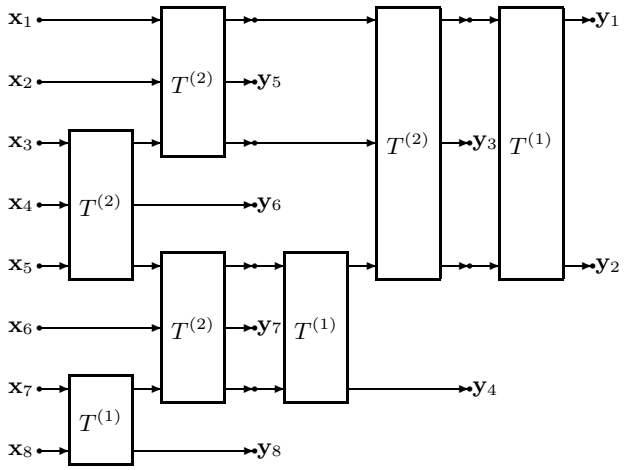

Fig. 2. Decomposition of a group of 8 pictures with orthogonal transforms $T^{(1)}$ and $T^{(2)}$.

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 compensated orthogonal transform as presented in [6]. $T^{(2)}$ denotes a bidirectionally compensated orthogonal transform as presented in [7]. Note that this architecture permits also block-wise decisions between unidirectional and bidirectional compensation. This adaptivity is used by the following multiview video decomposition.

3. DECOMPOSITION OF MULTIVIEW VIDEO

In the following, we discuss a decomposition of multiview video signals with motion and disparity compensated orthogonal transforms. We arrange the multiview video data into a *Matrix of Pictures* (MOP). Each MOP consists of N image sequences, each with K temporally successive pictures. With that, we consider the correlation among all the pictures within a MOP. We utilize our adaptive orthogonal transforms such that the MOP is encoded jointly.

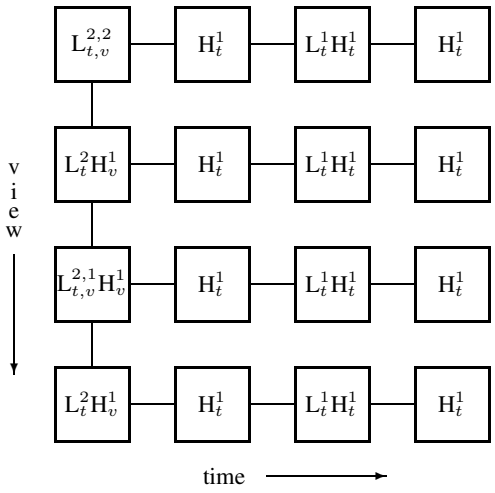

Fig. 3. Matrix of pictures (MOP) for $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. The coding structure is also shown. The temporal decomposition of each view is followed by one view decomposition only.

We explain our decomposition of the multiview video signal with the example in **Fig. 3**. It depicts a MOP of $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. Each MOP is encoded with one low-band picture and $NK - 1$ high-band pictures. First, a 2-level multiresolution decomposition of each view sequence in temporal direction is accomplished with motion-compensated orthogonal transforms. The first frame of each view is represented by the temporal low-band L^2_t, the remaining frames of each view by temporal high-bands H^1_t. Second, a 2-level multiresolution decomposition of the temporal low-bands L^2_t in view direction is accomplished with disparity-compensated orthogonal transforms. After the decomposition of N temporal low-bands, we obtain the MOP low-band $\mathrm{L}^2_t\mathrm{L}^2_v$ and the remaining $N - 1$ view high-bands H^1_v. We will use only the disparity fields among the views at the first time instant in the MOP. Therefore, we do not further decompose the temporal high-bands H^1_t in view direction.

4. EXPERIMENTAL RESULTS

Given our decomposition structure, we investigate the energy compaction of motion and disparity compensated orthogonal transforms for multiview video signals. For that, we use the multiview video data set *Breakdancers* with 4 views, 32 temporal frames, 15 fps, and a spatial resolution of 256×192. For simplicity, we have reduced the original spatial resolution of the data set with the MPEG downsampling filters. For the experiment, we choose a MOP with a group of $N = 4$ views and temporal GOP size $K = 2$. With above decomposition structure, we compare our adaptive orthogonal transforms

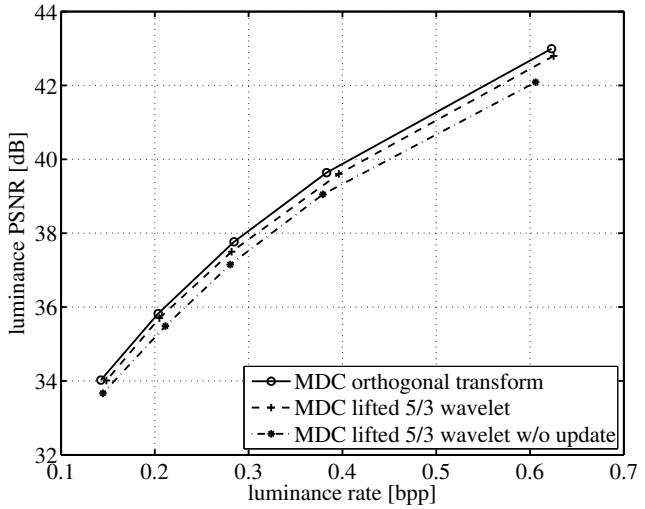

Fig. 4. PSNR over rate for the luminance signal of the data set *Breakdancers* with $N = 1$ view and temporal GOP size $K = 2$.

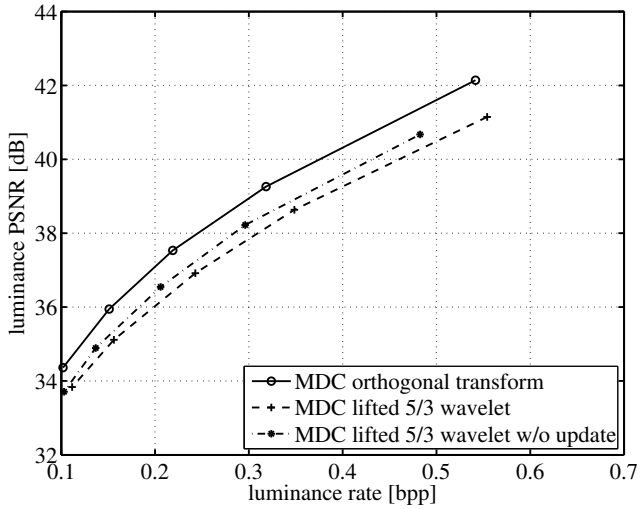

Fig. 5. PSNR over rate for the luminance signal of the data set *Breakdancers* with $N = 4$ views and temporal GOP size $K = 2$.

to adaptive lifted 5/3 wavelets with and without update step.

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 and disparity 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 and disparity vectors.

Motion and disparity compensation is limited to 8×8 blocks and integer-pel accuracy. An extension to sub-pel accuracy is possible [8]. Conditional motion and disparity estimation is used for bidirectional estimation. The same block motion/disparity fields are used for both orthogonal transform and 5/3 wavelet. For simplicity, the resulting temporal subbands are coded with JPEG 2000. The high-bands are coded directly, whereas the low-band of the MOP is rescaled by (11) before encoding. Lagrangian costs are used for optimal rate allocation. Note that the scale factors of the low-band are considered in the distortion term.

Figs. 4 and 5 depict the rate-distortion performance of the luminance signal of the test data set with $N = 1$ view and $N = 4$ views, respectively. Results for the motion and disparity compensated (MDC) orthogonal transform as well as the MDC lifted 5/3 wavelet with and without update step are given. As we assess energy compaction of subband decompositions, no intra modes have been used. Further, the bit-rate for motion and disparity information is not considered in the results as the same block motion/disparity fields are used for all three decompositions. The results show that the MDC orthogonal transform compares favorably with the MDC lifted 5/3 wavelet with and without update step. Moreover, the energy compaction of the MDC orthonormal transform improves relative to that of the MDC lifted wavelets when increasing the number of decomposition levels; see Fig. 5 vs. Fig. 4.

5. CONCLUSIONS

This paper investigates multiview video subband decompositions based on motion and disparity compensated orthogonal transforms. The orthogonal transforms offer the advantage that their subbands are strictly orthogonal, even if motion or disparity compensation is inaccurate. This is in contrast to adaptive lifted wavelets whose subband decomposition properties are motion- and disparity-dependent. The results show that improved energy compaction can be achieved with strictly orthogonal subbands, in particular, if the number of decomposition levels increases. Future work will develop a complete coding scheme and provide comparisons to standardized schemes.

6. REFERENCES

[1] A. Smolic and P. Kauff, "Interactive 3-D video representation and coding technologies," *Proceedings of the IEEE*, vol. 93, no. 1, pp. 98–110, Jan. 2005.

[2] M. Tanimoto, "FTV (free viewpoint television) creating ray-based image engineering," in *Proceedings of the IEEE International Conference on Image Processing*, Genova, Italy, Sept. 2005.

[3] *ITU-T Rec. H.264 – ISO/IEC 14496-10 AVC : Advanced Video Coding for Generic Audiovisual Services*, ITU-T and ISO/IEC Joint Video Team, 2005.

[4] A. Vetro, Y. Su, H. Kimata, and A. Smolic, "Joint multiview video model JMVM 2.0," ITU-T and ISO/IEC Joint Video Team, Document JVT-U207, Nov. 2006, http:// ftp3. itu. int/ av-arch/ jvt-site/ 2006_10_Hangzhou/ JVT-U207.zip.

[5] M. Flierl, A. Mavlankar, and B. Girod, "Motion and disparity compensated coding for video camera arrays," in *Proceedings of the Picture Coding Symposium*, Beijing, China, Apr. 2006, invited paper.

[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.

[7] ——, "A new bidirectionally motion-compensated orthogonal transform for video coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, Apr. 2007.

[8] ——, "A double motion-compensated orthogonal transform with energy concentration constraint," in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, vol. 6508, San Jose, CA, Jan. 2007.