

Multi-View Video Compression

— *Exploiting Inter-Image Similarities* —

Markus Flierl and Bernd Girod

The authors are with the Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305. This work has been supported by the Max Planck Center for Visual Computing and Communication. Contact e-mail: mflierl@IEEE.org

Multi-View Video Compression

— *Exploiting Inter-Image Similarities* —

Advances in display and camera technology enable new applications for 3D scene communication. 3DTV is among the most important of these applications; it strives to create a realistic 3D depth impression of the observed scene [1]. Typically, multiple video cameras are used to simultaneously acquire various viewpoints of the scene. The resulting data are often referred to as multi-view video. As the potential degree of 3D realism improves with the camera density around the scene, a vast amount of multi-view video data needs to be stored or transmitted for 3DTV. Multi-view video data is also expected to consume a large portion of the bandwidth available in the Internet of the future. This will include point-to-point communication as well as multicast scenarios. Multimedia distribution via sophisticated content delivery networks and flexible peer-to-peer networks will enable multi-view-video-on-demand as well as live broadcasts.

Due to the vast raw bit-rate of multi-view video, efficient compression techniques are essential for 3D scene communication [2]. As the video data originate from the same scene, the inherent similarities of the multi-view imagery are exploited for efficient compression. These similarities can be classified into two types. First, inter-view similarity is observed between adjacent camera views. Second, temporal similarity is noticed between temporally successive images of each video. Temporal similarities can be exploited with motion compensation techniques that are well known from single-view video compression. Extending that idea, disparity compensation techniques make use of inter-view similarities for multi-view video compression.

When designing compression schemes for multi-view video data, several constraints shape their architecture. In a communication scenario, multi-view video representations should be robustness against unreliable transmission. Further, it is desirable that these representations are highly flexible such that subsets of the original data can be accessed easily at various levels of image quality; the level of user interactivity that can be supported by a particular multi-view video representation will be an important consideration for on-demand applications. Finally, the overall trade-off between the quality of the reconstructed views and the bit-rate of its representation will be of high interest when processing the vast amount of data.

In the rest of the paper, we will first discuss the importance of exploiting inter-image similarities in multi-view video compression. We then introduce the basic approaches to multi-view video compression. One class of algorithms extends predictive coding as currently used in video compression standards

to multiple views. Another class of algorithms uses adaptive subband decomposition within and across video sequences from different cameras. We conclude the article by discussing the relative advantages and disadvantages of these compression approaches when faced with additional constraints that often arise in practical systems.

I. MULTI-VIEW VIDEO IMAGERY

Dynamic depth impressions of natural scenes can be created with multi-view video imagery. The imagery is generated by multiple video cameras that capture various viewpoints of the scene. The video camera arrangement is chosen according to the desired 3D scene representation. For example, a linear camera array is the simplest arrangement and offers parallax in one spatial dimension only. Planar camera arrays provide a broader depth impression but require a substantially larger number of cameras.

As the multi-view video imagery captures the same dynamic 3D scene, there exist inherent similarities among the images. We classify these similarities into two types. First, inter-view similarity is observed between adjacent camera views. Second, temporal similarity is noticed between temporally successive images of each video. This classification corresponds to the natural arrangement of multi-view video images into a *Matrix Of Pictures* (MOP) [3]. Each row holds temporally successive pictures of one view, and each column consists of spatially neighboring views captured at the same time instant. In case we deviate from linear camera arrays, all view sequences are still arranged into the rows of the MOP. Here, the idea is to distinguish between inter-view similarity and temporal similarity only. Therefore, further sub-classification of inter-view similarities is not intended.

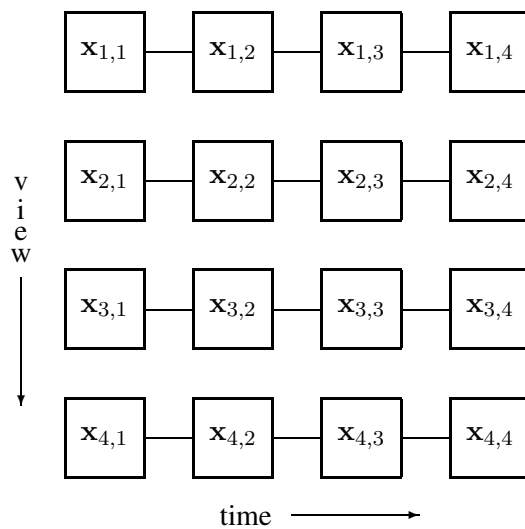


Fig. 1. Matrix of pictures (MOP) for $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures.

Fig. 1 depicts a matrix of pictures for $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. $N = 4$ views form a group of views (GOV), and $K = 4$ temporally successive pictures a temporal group of pictures (GOP). For example, the images of the first view sequence are denoted by $x_{1,k}$, with $k = 1, 2, \dots, K$. We choose MOPs with NK images to discuss the compression efficiency of coding schemes that process NK images jointly. Joint compression aims to exploit all similarities among these images. Later, we will discuss the impact of the MOP size (N, K) on the compression performance and the trade-off between the size N of the group of views and the size K of the temporal group of pictures.

II. EXPLOITING SIMILARITIES IN TIME AND AMONG VIEWS

Exploiting similarities among the multi-view video images is the key to efficient compression. When considering temporally successive images of one view sequence, i.e. one row of the MOP, the same view-point is captured at different time instances. Usually, the same objects appear in successive images but possibly at different pixel locations. If so, objects are in motion and practical compression schemes utilize *motion compensation* techniques to exploit this temporal similarities.

On the other hand, spatially neighboring views captured at the same time instant, i.e., images in one column of the MOP, show the same objects from different view-points. Similar to the previous case, the same objects appear in neighboring views but at different pixel locations. Here, the objects in each image are subject to parallax and practical compression schemes use *disparity compensation* techniques to exploit these inter-view similarities.

A. Temporal Similarities

Consider temporally successive images of one view sequence, i.e., one row of the MOP. If objects in the scene are subject to motion, the same objects appear in successive images but at different pixel locations. To exploit these temporal similarities, sophisticated motion compensation techniques have been developed in the past. Frequently used are so-called *block matching* techniques where a motion vector establishes a correspondence between two similar blocks of pixels chosen from two successive images [4]. Practical compression schemes signal this motion vectors to the decoder as part of the bit-stream. *Variable block size* techniques improve the adaptation of the block motion field to the actual shape of the object [5]. Lately, so-called *multi-frame* techniques have been developed. Classic block matching techniques use a single preceding image when choosing a reference for the correspondence. Multi-frame techniques, on the other hand, permit choosing the reference from several previously transmitted images;

a different image could be selected for each block [6]. Finally, *superposition* techniques are also used widely. Here, more than one correspondence per block of pixels is specified and signaled as part of the bit-stream [7]. A linear combination of the blocks resulting from multiple correspondences is used to better match the temporal similarities. A special example is the so-called *bidirectionally predicted picture* where blocks resulting from two correspondences are combined [8]. One correspondence uses a temporally preceding reference, the other uses a temporally succeeding reference. The generalized version is the so-called *bi-predictive picture* [9]. Here, two correspondences are chosen from an arbitrary set of available reference images.

B. Inter-View Similarities

Consider spatially neighboring views captured at the same time instant, i.e., images in one column of the MOP. Objects in each image are subject to parallax and appear at different pixel locations. To exploit this inter-view similarities, disparity compensation techniques are used.

The simplest approach to disparity compensation are block matching techniques similar to those used for motion compensation [10]. These techniques offer the advantage of not requiring knowledge of the geometry of the underlying 3D objects. However, if the cameras are sparsely distributed, the block-based translatory disparity model fails to compensate accurately.

More advanced approaches to disparity compensation are depth-image-based rendering algorithms [11]. They synthesize an image as seen from a given view-point by using the reference texture and depth image as input data. These techniques offer the advantage that the given view-point image is compensated more accurately even when the cameras are sparsely distributed. However, these techniques rely on accurate depth images, which are difficult to estimate.

Finally, hybrid techniques that combine the advantages of both approaches may also be considered. For example, if the accuracy of a depth image is not sufficient for accurate depth-image-based rendering, block-based compensation techniques may be used on top for selective refinement [12].

C. Performance Bounds

The rate-distortion efficiency of multi-view video coding is of great interest. For single-view video coding, theoretical performance bounds have been established for motion-compensated prediction [13] as well as motion-compensated subband coding [14]. Obviously, the simplest approach to multi-view video coding is to encode the individual video sequences independently. But for the most efficient compression of multi-view video data, the similarities among the views must also be taken into account. Therefore,

the work in [3] proposes a mathematical model that captures both inter-view correlation and temporal correlation. It is based on the high-rate model for motion-compensated subband coding of video [14].

The model captures the effects of motion compensation accuracy and disparity compensation accuracy. For that, it does not consider a particular compensation technique. Instead, it assumes perfect compensation up to a given motion inaccuracy and disparity inaccuracy. With that, rate-distortion bounds for both perfect and inaccurate compensation can be determined. Moreover, the model captures also the encoding of N views, each with K temporally successive pictures and its impact on the overall coding performance. In short, it models NK disparity and motion compensated pictures. These pictures are then decorrelated by the Karhunen-Loève Transform (KLT) for optimal encoding and for achieving rate-distortion bounds. At this point, we are not interested in bounds for a particular coding scheme. Rather, we are interested in compression bounds for multi-view video imagery given parameters like the size of the MOP (N, K) or the inaccuracy of disparity compensation. At high rates, good coding bounds can be determined by optimal transform coding with the KLT. This will help us to understand the fundamental trade-offs that are inherent to multi-view video coding. **Box 1** describes the signal model in more detail.

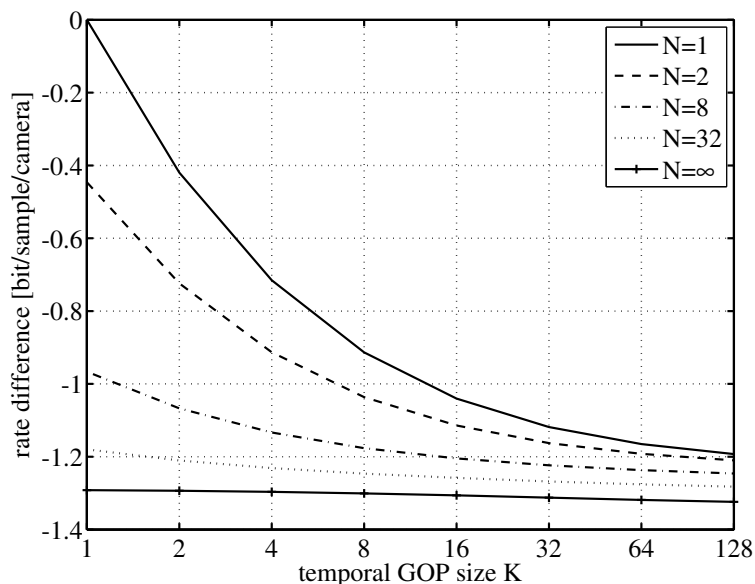


Fig. 2. Rate reduction due to exploiting the similarities among NK pictures at high image quality. Rate differences are calculated with a statistical signal model relative to intra coding of all images at the same quality and are negative as bit-rate is saved by joint coding. Rate differences are given for various temporal GOP sizes K and groups of views N .

Fig. 2 depicts typical rate reductions that can be achieved by exploiting the similarities among NK pictures at high image quality. The rate differences are obtained with the mathematical model in [3] and are calculated with respect to intra coding of all images at the same quality. For example, take the group

of views of size $N = 1$, meaning that each video signal is encoded independently. By increasing the temporal GOP size K , i.e., jointly encoding K motion-compensated pictures, the bit-rate decreases when compared to intra coding of the MOP. This observation holds also for larger groups of views N , where N disparity-compensated pictures are encoded jointly. But note that the relative decrease in bit-rate gets smaller for growing groups of views N . This result suggests a possible trade-off between the size of the group of views N and the size of the temporal GOP K , when considering the bit-rate savings only. Using the numerical values in **Fig. 2** as an example, jointly encoding of MOPs with $N = 8$ views and $K = 4$ temporal images yields on average similar rate savings than MOPs with $N = 2$ views and $K = 32$ temporal images. But note that actual quantitative values depend strongly on the type of multi-view video data, in particular, the motion in the scene, the accuracy of disparity compensation, and the noise level in the sequences.

Finally, the accuracy of disparity compensation affects the overall bit-rate savings significantly. In practice, neither block matching techniques nor depth-image-based rendering algorithms can perform perfect disparity compensation. Occlusions and varying lighting conditions among the views are challenging. In cases in which we are able to improve the accuracy of compensation, we will benefit in terms of overall bit-rate savings [3].

III. COMPRESSION SCHEMES

The vast amount of multi-view data is a huge challenge not only for capturing and processing but also for compression. Efficient compression exploits statistical dependencies within the multi-view video imagery. Usually, practical schemes accomplish this either with predictive coding or with subband coding. In both cases, motion compensation and disparity compensation are employed to make better use of statistical dependencies.

Note that predictive coding and subband coding have different constraints for efficient compression. Predictive coding is accomplished by processing images sequentially. Hence, the order in which the images are processed is of importance. Moreover, coding decisions made in the beginning of the sequence will affect subsequent coding decisions. On the other hand, subband coding does not require sequential processing of images. All images to be encoded are subject to a subband decomposition which is followed by independent encoding of its coefficients. Hence, coding decisions made at the second stage do not affect the subband decomposition in the first stage.

In the following, we consider these practical schemes for multi-view video compression and discuss them in more detail.

A. Predictive Coding

Predictive coding schemes encode multi-view video imagery sequentially. Two basic types of coded pictures are possible: *intra* and *inter* pictures. Intra pictures are coded independently of any other image. Inter pictures, on the other hand, depend on one or more reference pictures that have been encoded previously. By design, an intra picture does not exploit the similarities among the multi-view images. But an inter picture is able to make use of these similarities by choosing one or more reference pictures and generating a motion and/or disparity compensated image for efficient predictive coding. The basic ideas of motion-compensated predictive coding are summarized in **Box 2**.

When choosing the encoding order of images, various constraints should be considered. For example, high coding efficiency as well as good temporal multiresolution properties may be desirable. Interestingly, both goals can be combined very well. Similar to a temporal multiresolution decomposition, a coarse resolution layer of temporally distant images is successively refined by inserting inter coded pictures at half temporal distance. Note that these inter coded pictures use the coded images of the coarser resolution layer as references. This method of *hierarchical encoding* offers not only a temporal multiresolution representation but also high coding efficiency. For hierarchical encoding, the *bi-predictive picture* [9] is very useful. It is a special inter picture that chooses up to two reference pictures for generating a motion and/or disparity compensated image. Its coding efficiency is superior to that of the “basic” inter picture (predictive picture), which chooses only one reference picture for compensation.

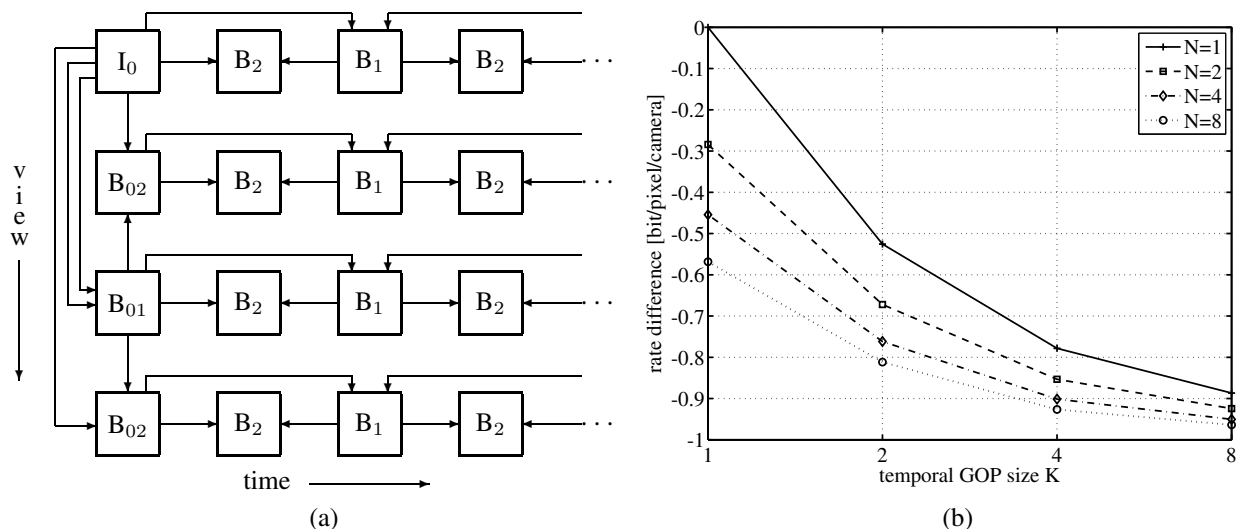


Fig. 3. Hierarchical encoding of a matrix of pictures (MOP) with bi-predictive pictures. (a) MOP with $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. (b) Rate difference for the multi-view video *Ballroom* at an average image quality of 40 dB PSNR achieved by exploiting the similarities within each MOP of size (N, K) .

Fig. 3(a) depicts a possible hierarchical encoding of a MOP with $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. Each MOP is encoded with one intra picture and $NK - 1$ bi-predictive pictures. First, each MOP is decomposed in view direction at the first time instant only. That is, the sequences have view decompositions at every K -th time instant. The intra picture I_0 in each MOP represents the lowest view resolution. The next view resolution level is attained by including the bi-predictive picture B_{01} . The highest view resolution is achieved with the bi-predictive pictures B_{02} . Second, the reconstructed N view images at every K -th time instant are now used as reference for multiresolution decompositions with bi-predictive pictures in temporal direction. The decomposition in view direction at every K -th time instant represents already the lowest temporal resolution level. The next temporal resolution level is attained by including the bi-predictive pictures B_1 . The highest temporal resolution is achieved with the bi-predictive pictures B_2 . Thus, hierarchical encoding of each MOP with bi-predictive pictures generates a representation with multiple resolutions in time and view direction [15].

Multi-view video coding is currently investigated by the *Joint Video Team* (JVT). The JVT is developing a *Joint Multiview Video Model* (JMVM) [16] which is based on the video coding standard ITU-T Recommendation H.264 – ISO/IEC 14496-10 AVC [17]. The current JMVM proposes illumination change-adaptive motion compensation and prediction structures with hierarchical bi-predictive pictures. The JMVM uses the block-based coding techniques of H.264/AVC to exploit both temporal similarities and view similarities. The coding structure is investigated in [18], [19]. The standard codec H.264/AVC is a hybrid video codec and incorporates an intra-frame codec and a motion-compensated inter-frame predictor. When encoding image sequences, sophisticated coder control techniques choose from multiple intra- and inter-picture modes to optimize rate-distortion efficiency. An important parameter is the number of previously decoded pictures stored in the reference frame buffer. Both, rate-distortion efficiency and computational complexity grow with the number of stored reference pictures. **Fig. 3(b)** shows experimental results obtained with hierarchical bi-predictive pictures for the multi-view video *Ballroom*. It depicts achievable rate differences to intra coding by exploiting the similarities within each MOP of size (N, K) . The rate difference is measured at an average image quality of 40 dB PSNR relative to the intra coding rate of 1.4 bit per pixel per camera.

In summary, predictive coding schemes are technologically well advanced and offer good image quality at low bit-rates, in particular with the advent of the latest standard H.264/AVC. Though, such schemes are burdened by the inherent constraint of sequential coding. Recall that coding decisions made in the beginning of the sequence will affect subsequent coding decisions. This affects overall coding efficiency and produces multi-view video representations of limited flexibility.

B. Subband Coding

All images to be encoded by a subband coding scheme are subject to a subband decomposition which is followed by quantization and entropy-coding of its coefficients. Such schemes do not require sequential processing of images and, hence, offer more flexible multi-view video representations. Like in predictive coding, the subband decomposition makes use of similarities among the multi-view video images. As similarities are exploited by motion and disparity compensation, adaptive subband decompositions are of interest [20]–[22].

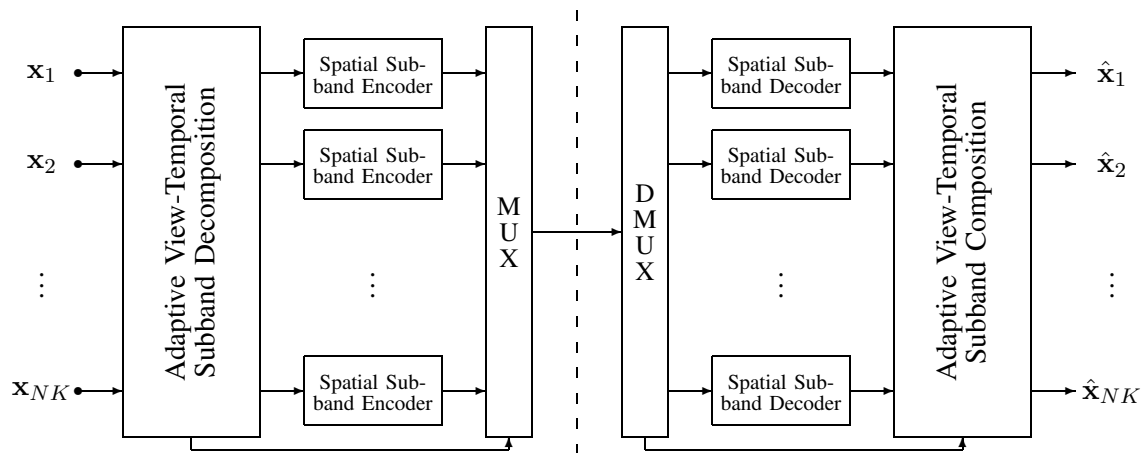


Fig. 4. Motion and disparity adaptive subband coding. The multi-view video is represented by a motion and disparity compensated subband decomposition. The resulting view-temporal subbands are encoded and multiplexed with motion and disparity side information into one bit-stream.

A typical motion and disparity adaptive subband coding scheme is shown in **Fig. 4**. NK images of the multi-view video data are transformed into NK subband images by a motion and disparity compensated subband decomposition. Only one subband image, the so-called low-band image, accumulates the major energy of all images in the MOP. The remaining $NK - 1$ subband images, so-called high-band images, carry only minor energy components that could not be concentrated into the low-band image. The decomposition is followed by spatial encoding of the view-temporal subband coefficients. The output bit-stream of the encoder includes both the compressed representation of the subband coefficients as well as the motion and the disparity information. The corresponding decoder simply inverts the processing steps of the encoder.

When choosing an adaptive transform for multi-view video subband coding, various constraints should be considered. For example, given the unquantized subband coefficients of the forward transform, the inverse adaptive transform at the decoder should be able to reconstruct the input imagery perfectly. In

addition, good view-temporal multiresolution properties are desirable. Both goals can be combined very well with so-called *motion and disparity compensated lifted wavelets* [23], [24]. Wavelets implemented with the lifting architecture are reversible, even if the operations in the lifting steps are non-linear like motion and disparity compensation. Moreover, multiresolution representations are easily obtained with wavelet transforms. Similar to predictive coding where predictive and bi-predictive pictures exploit the similarities among the images, two basic types of motion-compensated lifted wavelets are popular. The basic adaptive wavelet is the motion-compensated lifted Haar wavelet where high-bands are generated from one motion-compensated image only. The advanced adaptive wavelet is the motion-compensated lifted 5/3 wavelet where high-bands are generated by a linear combination of two motion-compensated images. Better energy concentration is achieved with the adaptive 5/3 wavelet, which is also more complex than the adaptive Haar wavelet. **Box 3** outlines the basic concepts of motion-compensated lifted wavelets.

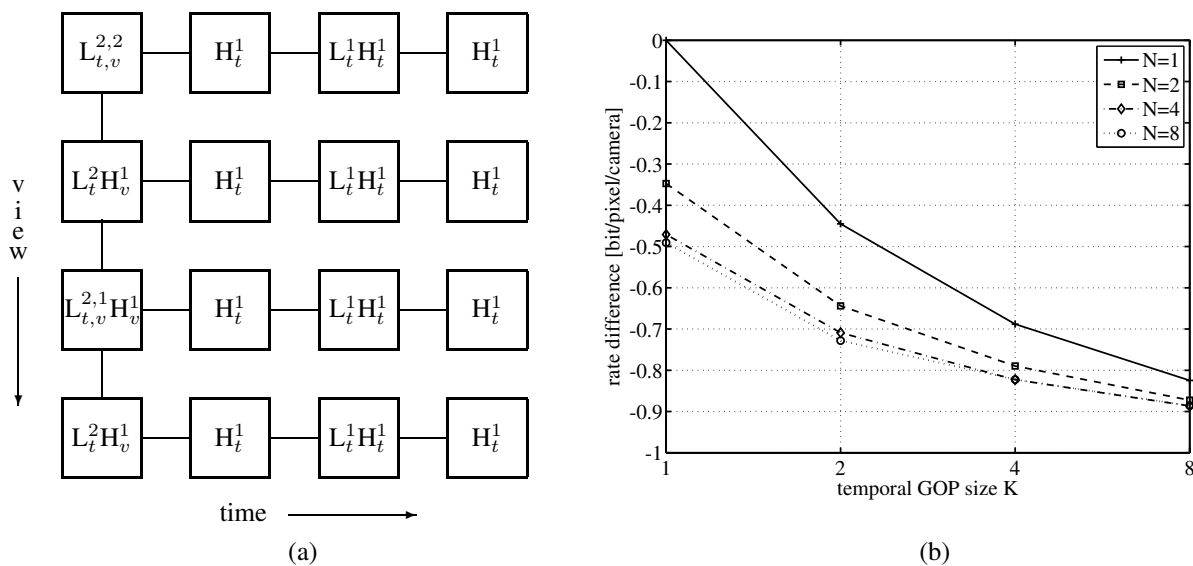


Fig. 5. Hierarchical subband decomposition of a matrix of pictures (MOP). (a) MOP with $N = 4$ image sequences, each comprising of $K = 4$ temporally successive pictures. (b) Rate difference for the multi-view video *Ballroom* at an average image quality of 40 dB PSNR achieved by exploiting the similarities within each MOP of size (N, K) .

Fig. 5(a) shows a possible view-temporal multiresolution decomposition of a MOP with $N = 4$ 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 wavelets. The first frame of each view is represented by the temporal low-band L_t^2 , the remaining frames of each view

high-bands H_t^1 . Second, a 2-level multiresolution decomposition of the temporal low-bands L_t^2 in view direction is accomplished with disparity-compensated wavelets. After the decomposition of N temporal low-bands, we obtain the MOP low-band $L_t^2 L_v^2$ and the remaining $N - 1$ view high-bands H_v^1 . This decomposition uses only the disparity information among the views at the first time instant in the MOP. **Fig. 5(b)** gives experimental results for the multi-view video *Ballroom* obtained with an adaptive subband decomposition using lifted wavelets. The results are based on a version of the *Joint Scalable Video Model* (JSVM) [25] which supports adaptive lifted wavelets. The plot depicts achievable rate differences to intra coding by exploiting the similarities within each MOP of size (N, K) . Note that the rate difference is measured at an average image quality of 40 dB PSNR relative to the intra coding rate of 1.4 bit per pixel per camera.

In summary, subband coding schemes offer more flexible representations for multi-view imagery. For static light fields, this has been demonstrated in [26], where disparity-compensated wavelets have been investigated. Further examples for multi-view wavelet video coding are given in [27]. Though, decompositions resulting from motion and disparity compensated lifted wavelets usually suffer from a compensation mismatch in predict and update step, especially for multi-connected motion and disparity fields. This compensation mismatch alters properties that are offered by the corresponding non-adaptive wavelet transforms. For example, the non-adaptive lifted Haar wavelet is strictly orthonormal, whereas the motion-compensated lifted Haar wavelet loses the property of orthonormality if multi-connected motion fields are compensated [28].

The development of view-temporal subband decompositions that maintain their orthogonality with arbitrary motion and disparity compensation is still a challenging research problem. First attempts at a solution have been reported recently for unidirectional motion compensation [28], sub-pel accurate motion compensation [29], [30], and bidirectional motion compensation [31]. **(Box 4)**.

IV. COMPRESSION WITH ADDITIONAL CONSTRAINTS

Compression engines are usually part of information or communication systems that impose additional constraints on the compression scheme itself. Basic constraints are delay and memory requirements. Interactive applications like free-viewpoint video [2] impose random access requirements which permit access to individual image sequences in the compressed multi-view video representation. On the other hand, communication systems require compressed representations to be robust to transmission errors and might benefit from rate scalability. In the following, we revisit above compression schemes while considering practical system constraints.

A. Delay and Memory Constraints

Delay is caused by the wait time that elapses when a coding scheme collects additional images from the source which are required for encoding. Sequential encoding with predictive schemes permits flexible encoding orders. This wait time can be reduced to zero with forward prediction only. For that case, bi-directional prediction in temporal direction cannot be used and, hence, lower coding efficiency is observed. Higher coding gains can be achieved by permitting delay. Delay constraints are different for subband coding schemes. In general, all images of a MOP have to be considered to determine the low-band image of the subband decomposition. Hence, the minimum delay for a MOP of size (N, K) is the wait time necessary to collect additional $K - 1$ temporally successive images.

Memory requirements specify the size of the memory that is necessary to facilitate encoding or decoding. For predictive schemes, the size of the multi-frame reference buffer determines the memory requirement. At least one reference image needs to be in memory for predictive coding. And large reference frame buffers are likely to improve compression efficiency. Memory requirements are also different for subband coding schemes. In general, subband decompositions require all input images associated with a MOP to reside in the memory of the encoder. Therefore, the memory requirement increases with the size of the MOP and, hence, the desired compression efficiency.

B. Random Access

Applications like interactive light field streaming [32] or free-viewpoint video [2] impose random access requirements on multi-view imagery. Random access refers to the accessibility of individual images or image sequences in compressed representations. For predictive coding schemes, access to individual images depends highly on actual prediction dependencies. Note that sequential encoding requires all intermediate reference pictures to be decoded in order. Hence, hierarchical encoding orders facilitate more flexible access to individual images than linear encoding orders. For subband coding schemes, random access is facilitated by multi-resolution subband decompositions, Again, hierarchical representations allow flexible access to individual images. Moreover, subband schemes offer the opportunity to trade off between the burden of access and the quality of retrieved images.

C. Flexible Representations and Robustness

Practical 3DTV systems require multi-view video representations to be robust against unreliable transmission. Scalable representations allow flexible adaptations to network and channel conditions. For

example, view scalability and temporal scalability facilitate transmission of subsets of the original multi-view video data. This is achieved by using hierarchical encoding structures for both predictive and subband coding schemes. Quality scalability facilitates transmission of multi-view video at various image quality levels. For efficient predictive coding, reference pictures at encoder and decoder have to match exactly. If decoding at various quality levels is desired, the encoder has to encode all desired quality levels to match exactly the necessary reference pictures. Subband coding schemes, on the other hand, process the quantization noise differently and allow for efficient quality scalability.

Finally, decoders for robust representations should minimize the impact of transmission errors on the reconstructed multi-view video. Note that predictive encoders operate in a closed-loop fashion. The total quantization error energy across intra picture and displaced frame differences equals that in the corresponding reconstructed pictures. In case of transmission errors, decoded reference frames differ from the optimized reference frames at the encoder and errors propagate from frame to frame, resulting in an often very large amplification of the transmission error energy. On the other hand, subband coders operate in an open-loop fashion. In particular, energy conservation holds for orthogonal transforms such that the total quantization error energy in the coefficient domain equals that in the image domain. In case of transmission errors, the same relation holds. Hence, the error energy is preserved rather than amplified by the decoder, as is the case for predictive decoders.

V. FUTURE CHALLENGES

Both predictive coding schemes and subband coding schemes have the potential to exploit the inter-image similarities of multi-view video. Predictive coding schemes are technologically well advanced and offer good image quality at low bit-rates. Though, such schemes are burdened by the inherent constraint of sequential coding. Subband coding approaches provide desirable properties for the compressed representation. But these techniques are not at the same level of maturity as predictive coding schemes. The vast amount of data that comes with multi-view video renders highly structured representations more desirable. Additional constraints on adaptive subband decompositions may be necessary. It is a future challenge to make subband coding competitive with predictive schemes, while maintaining all the desirable properties that come with such decompositions.

ACKNOWLEDGMENTS

This work has been supported by the Max Planck Center for Visual Computing and Communication at Stanford University.

APPENDIX I

 BOX 1: STATISTICAL SIGNAL MODEL FOR
 MULTI-VIEW VIDEO

The model generates NK disparity and motion compensated pictures $\{s_i, i = 1, 2, \dots, NK\}$ from one *root picture* \mathbf{v} in two steps. First, the *root image sequence* $\{c_k, k = 1, 2, \dots, K\}$ with K motion-compensated pictures is generated from the root image \mathbf{v} . For this, the root picture is shifted by *displacement error vectors* Δ_{1k} and distorted by additive *residual video noise* \mathbf{n}_k . Second, N view sequences with NK disparity and motion compensated pictures are generated from the root image sequence. Here, the pictures of the root image sequence are shifted by *disparity error vectors* $\Theta_{1\nu}$, $\nu = 2, 3, \dots, N$, and distorted by additive *residual multi-view noise* \mathbf{z}_i , $i = 1, 2, \dots, NK$. Note that the first picture of the root image sequence is the reference image. The remaining $K - 1$ pictures are motion compensated with respect to the reference image up to the specified displacement error. The concept of reference is also used for the N view sequences. $N - 1$ view sequences are disparity compensated with respect to the reference view sequence, i.e., the first view sequence, up to the specified disparity error. The complete signal model is depicted in **Fig. 6**. Note that all K temporal pictures of the ν -th view are shifted by the same disparity error vector $\Theta_{1\nu}$. We assume that the position of each camera is constant in time. Hence, we observe the same disparity error vector at each

time instant.

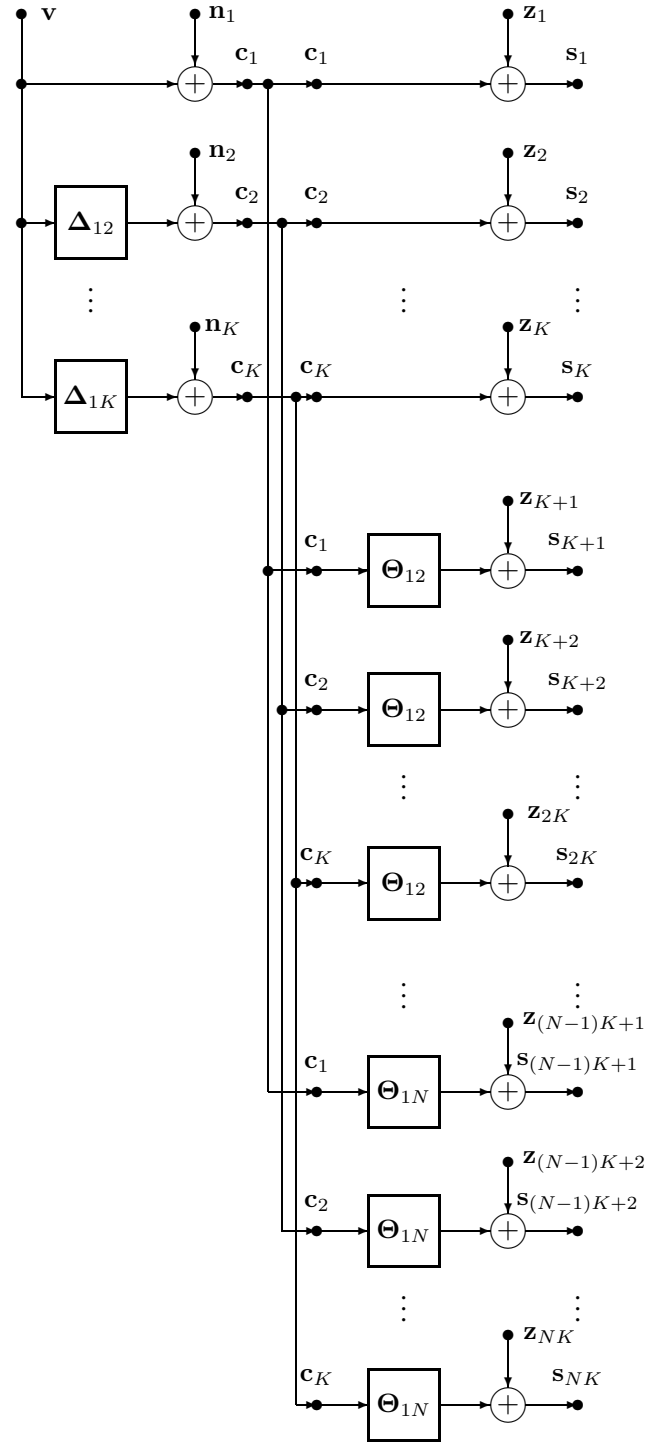


Fig. 6. Signal model for N image sequences each comprising of a group of K temporally successive pictures.

With additional assumptions as stated in [3], the power spectral density matrix of NK motion and disparity compensated pictures is

$$\Phi_{ss}(\omega) = \Gamma(\omega) \otimes \Phi_{cc}(\omega) + \Phi_{zz}(\omega), \quad (1)$$

where $\Gamma(\omega)$ is the $N \times N$ characteristic matrix of $N - 1$ disparity errors, $\Phi_{cc}(\omega)$ the $K \times K$ power spectral density matrix of the root image sequence [14], and $\Phi_{zz}(\omega)$ the $NK \times NK$ power spectral density matrix of the residual multi-view noise. \otimes denotes the Kronecker product and ω is the vector of spatial frequencies in horizontal and vertical direction.

The key parameters of the model specify displacement error and disparity error distributions as well as residual video noise and residual multi-view noise. The variances of displacement error and disparity error capture motion inaccuracy and disparity inaccuracy, respectively. For example, very accurate motion compensation is modeled by a very small displacement error variance. The residual video noise captures signal components that cannot be removed even by very accurate motion compensation, e.g., detail visible in one frame, but not in the other. The residual multi-view noise captures signal components that cannot be removed by very accurate disparity compensation between views, e.g., camera noise. Further details on the model are given in [3].

APPENDIX II

BOX 2: MOTION-COMPENSATED PREDICTIVE CODING

Motion-compensated predictive coding of image sequences is accomplished with intra and inter pictures as depicted in **Fig. 7**. **(a)** The input image \mathbf{x}_k is independently encoded into the intra picture I_k . The intra decoder is used to independently reconstruct the image $\hat{\mathbf{x}}_k$. **(b)** The input image \mathbf{x}_k is predicted by the motion-compensated (MC) reference image $\hat{\mathbf{x}}_r$. The prediction error, also called displaced frame difference (DFD), is encoded and constitutes in combination with the motion information the inter picture P_k . The inter-picture decoder reverses this process but requires the same reference image $\hat{\mathbf{x}}_r$ to be present at the decoder side. If the reference picture differs at encoder and decoder side, e.g. due to network errors, the decoder is not able to reconstruct the same image $\hat{\mathbf{x}}_k$ that the encoder has encoded. Note that reference pictures can be either reconstructed intra pictures or other reconstructed inter pictures.

Fig. 7(b) shows the “basic” inter picture (predictive picture) which chooses only one reference picture for compensation. More advanced are bi-predictive pictures that use a linear combination of two motion-compensated reference pictures. Bidirectionally motion-compensated prediction is a special case of bi-predictive pictures and is widely employed in standards like MPEG-1, MPEG-2, and H.263. The general concept of bi-predictive pictures [9] has been implemented with the standard H.264/AVC [17].

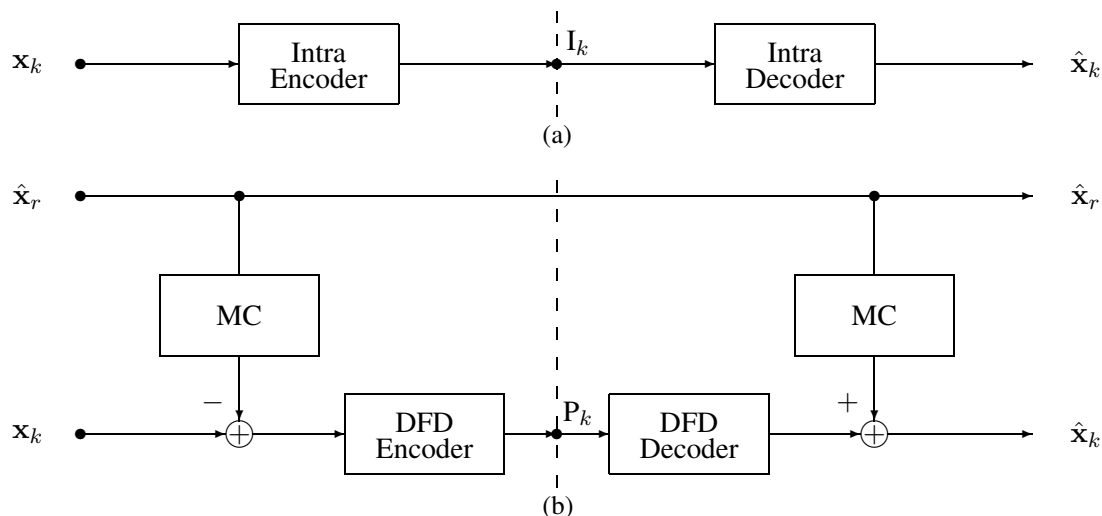


Fig. 7. Motion-compensated predictive coding with (a) intra pictures and (b) inter pictures.

Inter pictures have been studied extensively and theoretical performance bounds have been established. A high-rate model for predictive pictures is presented in [13]. This work has been extended to accommodate fractional-pel accuracy [33], multihypothesis prediction [34], and complementary hypotheses [35].

APPENDIX III

BOX 3: MOTION-COMPENSATED LIFTED WAVELETS

Motion-compensated lifted wavelets benefit from the fact that any wavelet implemented with the lifting architecture is reversible and, hence, biorthogonal [36]. The lifting architecture has ladder structure where predict and update steps modify even and odd samples of the signal to generate low-band and high-band samples, respectively. The operations performed in the lifting steps do not affect the biorthogonality of the wavelet, hence, non-linear operations like motion compensation can be introduced to design motion-adaptive lifted wavelets.

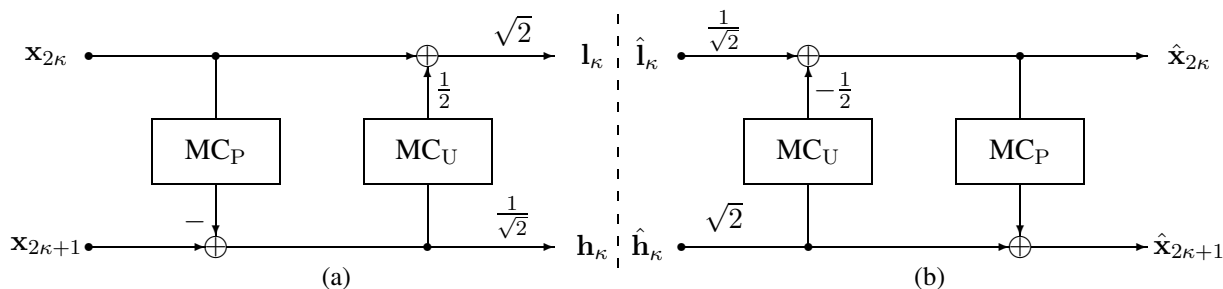


Fig. 8. Haar transform with motion-compensated lifting steps. (a) The encoder uses the forward transform, (b) the decoder the backward transform.

Fig. 8 shows the motion-compensated lifted Haar wavelet with **(a)** analysis and **(b)** synthesis [23], [24]. In the analysis, even images of a sequence $\mathbf{x}_{2\kappa}$ are motion compensated in the predict step (MC_P) to generate temporal high-band images \mathbf{h}_κ from odd images $\mathbf{x}_{2\kappa+1}$. Temporal low-band images \mathbf{l}_κ are derived from even images by adding the motion-compensated update (MC_U) of the scaled high-band images \mathbf{h}_κ . The synthesis simply reverses the sequence of lifting steps that are used in the analysis. To maintain reversibility, operations in the lifting steps need not to be invertible. This is advantageous as motion compensation is generally not invertible due to unconnected and multi-connected pixels. But note that the non-adaptive lifted Haar wavelet is strictly orthonormal, whereas the motion-compensated version loses this property if unconnected and multi-connected pixels are compensated.

Motion-compensated lifted wavelets have been investigated for subband coding of video. Theoretical performance bounds have been derived for additive motion [7], [14] as well as for complementary motion-compensated signals [37].

APPENDIX IV

BOX 4: MOTION-COMPENSATED ORTHOGONAL TRANSFORMS

Motion-compensated orthogonal transforms (MCOT) maintain strict orthogonality with arbitrary motion compensation. For the following discussion, we choose a unidirectionally motion-compensated orthogonal transform as depicted in **Fig. 9(a)**.

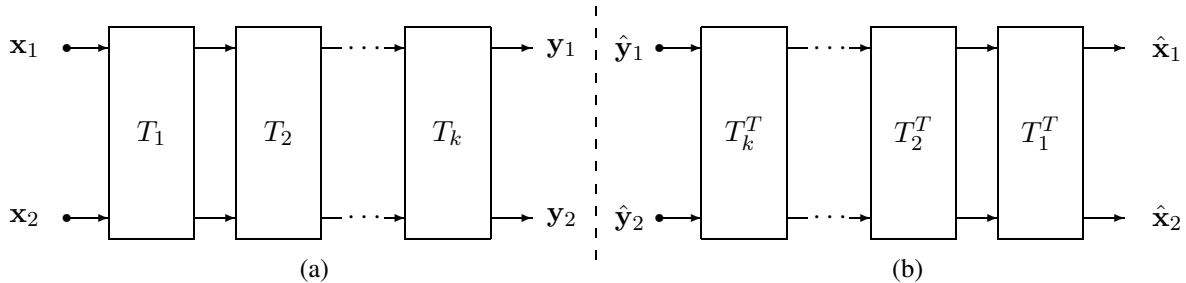


Fig. 9. Unidirectionally motion-compensated orthogonal transform. (a) The encoder uses the forward transform, (b) the decoder the backward transform. Each incremental transform T_κ , $\kappa = 1, 2, \dots, k$, is orthogonal, i.e., $T_\kappa T_\kappa^T = I$.

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

$$\begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix} = T \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{pmatrix} \quad (2)$$

into two vectors \mathbf{y}_1 and \mathbf{y}_2 which represent the temporal low-band and the temporal high-band, respectively. The transform T is factored into a sequence of k incremental transforms T_κ such that

$$T = T_k T_{k-1} \cdots T_\kappa \cdots T_2 T_1, \quad (3)$$

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

The incremental transform T_κ is nearly an identity matrix. The diagonal elements equal to 1 represent the untouched pixels in step κ . If one pixel in \mathbf{x}_2 is unidirectionally motion compensated in step κ , the incremental transform T_κ has two diagonal elements that are not equal to 1. These two diagonal elements and their corresponding two off-diagonal elements are equal to the four elements of a 2D rotation matrix. These two diagonal elements also indicate the two pixels that are connected by the associated motion vector and are subject to linear operations.

Further, if unidirectional motion compensation is not suitable for a pixel or block in \mathbf{x}_2 , the corresponding incremental transform in step κ is set to $T_\kappa = I$, where I denotes the identity matrix. This is called the *intra mode* for a pixel or block in picture \mathbf{x}_2 . Note that a pixel or block in picture \mathbf{x}_2 is modified by at most one incremental transform. Therefore, the type of incremental transform can be chosen freely in each step κ to match the motion of the affected pixels in \mathbf{x}_2 without destroying the property of orthonormality.

The unidirectionally motion-compensated incremental transform is just one example. There are also double motion-compensated [30] and bidirectionally motion-compensated transforms [31]. Each type of incremental transform has its own energy concentration constraint which efficiently removes energy in high-band pixels while considering motion compensation.

Any combination of these transforms can be used for dyadic decompositions while maintaining strict orthonormality. When used for multi-view video in view direction, motion compensation is replaced by disparity compensation while maintaining the principles of the transform. Hence, adaptive view-temporal subband decompositions that are strictly orthogonal can be generated from multi-view video data.

Finally, adaptive orthogonal transforms do not suffer from compensation mismatch in predict and update step that can be observed with block-compensated lifted wavelets. For example, **Fig. 10** compares decoded frames of the multi-view video *Breakdancers*. The complex motion of the dancer causes the lifted 5/3 wavelet to produce annoying noise artifacts that are not observed with the bidirectionally compensated orthogonal transform. Note that both schemes use the same block motion/disparity field as well as the same view-temporal decomposition structure as depicted in **Fig. 5(a)**.



Fig. 10. Decoded pictures of the multi-view video *Breakdancers*. The subband coding scheme uses (a) the motion and disparity compensated lifted 5/3 wavelet or (b) the bidirectionally motion and disparity compensated orthogonal transform. In both cases, the same 8×8 block motion/disparity field is used. View-temporal subbands are encoded with JPEG 2000.

REFERENCES

- [1] 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.
- [2] 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.
- [3] M. Flierl, A. Mavlanckar, and B. Girod, "Motion and disparity compensated coding for multi-view video," *IEEE Transactions on Circuits and Systems for Video Technology*, 2007, invited paper, to appear.
- [4] J. Jain and A. Jain, "Displacement measurement and its application in interframe image coding," *IEEE Transactions on Communications*, vol. 29, no. 12, pp. 1799–1808, Dec. 1981.
- [5] P. Strobach, "Tree-structured scene adaptive coder," *IEEE Transactions on Communications*, vol. 38, no. 4, pp. 477–486, Apr. 1990.
- [6] T. Wiegand, X. Zhang, and B. Girod, "Long-term memory motion-compensated prediction," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 1, pp. 70–84, Feb. 1999.
- [7] M. Flierl and B. Girod, *Video Coding with Superimposed Motion-Compensated Signals: Applications to H.264 and Beyond*. Boston – Dordrecht – London: Kluwer Academic Publishers (now Springer), 2004.
- [8] A. Puri, R. Aravind, B. Haskell, and R. Leonardi, "Video coding with motion-compensated interpolation for CD-ROM applications," *Signal Processing: Image Communication*, vol. 2, no. 2, pp. 127–144, Aug. 1990.
- [9] M. Flierl and B. Girod, "Generalized B pictures and the draft H.264/AVC video compression standard," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 587–597, July 2003, invited paper.
- [10] M. Lukacs, "Predictive coding of multi-viewpoint image sets," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, Tokyo, Japan, Apr. 1986.
- [11] L. McMillan and G. Bishop, "Plenoptic modeling: An image-based rendering system," in *Proceedings of the ACM SIGGRAPH*, Los Angeles, CA, Aug. 1995, pp. 39–46.

- [12] E. Martinian, A. Behrens, J. Xin, and A. Vetro, "View synthesis for multiview video compression," in *Proceedings of the Picture Coding Symposium*, Beijing, China, Apr. 2006.
- [13] B. Girod, "The efficiency of motion-compensating prediction for hybrid coding of video sequences," *IEEE Journal on Selected Areas in Communications*, vol. SAC-5, no. 7, pp. 1140–1154, Aug. 1987.
- [14] M. Flierl and B. Girod, "Video coding with motion-compensated lifted wavelet transforms," *Signal Processing: Image Communication*, vol. 19, no. 7, pp. 561–575, Aug. 2004.
- [15] 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.
- [16] 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.
- [17] *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.
- [18] K. Mueller, P. Merkle, H. Schwarz, T. Hinz, A. Smolic, T. Oelbaum, and T. Wiegand, "Multi-view video coding based on H.264/AVC using hierarchical B-frames," in *Proceedings of the Picture Coding Symposium*, Beijing, China, Apr. 2006.
- [19] P. Merkle, K. Mueller, A. Smolic, and T. Wiegand, "Efficient compression of multi-view video exploiting inter-view dependencies based on H.264/MPEG4-AVC," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, Toronto, Canada, July 2006.
- [20] T. Kronander, "Motion compensated 3-dimensional wave-form image coding," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, vol. 3, Glasgow, Scotland, May 1989, pp. 1921–1924.
- [21] J.-R. Ohm, "Three-dimensional subband coding with motion compensation," *IEEE Transactions on Image Processing*, vol. 3, no. 5, pp. 559–571, Sept. 1994.
- [22] S.-J. Choi and J. Woods, "Motion-compensated 3-d subband coding of video," *IEEE Transactions on Image Processing*, vol. 8, no. 2, pp. 155–167, Feb. 1999.
- [23] 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*, vol. 3, Salt Lake City, UT, May 2001, pp. 1793–1796.
- [24] 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*, vol. 2, Thessaloniki, Greece, Oct. 2001, pp. 1029–1032.
- [25] J. Reichel, H. Schwarz, and M. Wien, "Joint scalable video model JSVM 4.0," ITU-T and ISO/IEC Joint Video Team, Document JVT-Q202, Nov. 2005, http://ftp3.itu.int/av-arch/jvt-site/2005_10_Nice/JVT-Q202.zip.
- [26] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, "Light field compression using disparity-compensated lifting and shape adaptation," *IEEE Transactions on Image Processing*, vol. 15, no. 4, pp. 793–806, Apr. 2006.
- [27] W. Yang, F. Wu, Y. Lu, J. Cai, K. N. Ngan, and S. Li, "4-d wavelet-based multiview video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 16, no. 11, pp. 1385–1396, Nov. 2006.
- [28] 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.
- [29] —, "Half-pel accurate motion-compensated orthogonal video transforms," in *Proceedings of the Data Compression Conference*, Snowbird, UT, Mar. 2007.

- [30] —, “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.
- [31] —, “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.
- [32] P. Ramanathan, M. Kalman, and B. Girod, “Rate-distortion optimized interactive light field streaming,” *IEEE Transactions on Multimedia*, vol. 9, no. 4, pp. 813–825, June 2007.
- [33] B. Girod, “Motion-compensating prediction with fractional-pel accuracy,” *IEEE Transactions on Communications*, vol. 41, no. 4, pp. 604–612, Apr. 1993.
- [34] —, “Efficiency analysis of multihypothesis motion-compensated prediction for video coding,” *IEEE Transactions on Image Processing*, vol. 9, no. 2, pp. 173–183, Feb. 2000.
- [35] M. Flierl and B. Girod, “Multihypothesis motion estimation for video coding,” in *Proceedings of the Data Compression Conference*, Snowbird, UT, Mar. 2001, pp. 341–350.
- [36] W. Sweldens, “The lifting scheme: A construction of second generation wavelets,” *SIAM Journal on Mathematical Analysis*, vol. 29, no. 2, pp. 511–546, 1998.
- [37] M. Flierl, P. Vanderghenst, and B. Girod, “Video coding with lifted wavelet transforms and complementary motion-compensated signals,” in *Proceedings of the SPIE Conference on Visual Communications and Image Processing*, vol. 5308, San Jose, CA, Jan. 2004, pp. 497–508.