4D Wavelet-Based Multi-view Video Coding

Wenxian Yang, Yan Lu, Feng Wu, Jianfei Cai, King Ngi Ngan, and Shipeng Li

Abstract—**The conventional multi-view video coding (MVC) schemes, utilizing both neighboring temporal frames and view frames as possible references, have only shown a slight gain over those using temporal frames alone in terms of coding efficiency. The reason is that the neighboring temporal frames exhibit stronger correlation with the current frame and the view frames often fail to be selected as references. This paper proposes an elegant MVC framework using high dimensional wavelet, which rightly matches the inherent high dimension property of multi-view video. It also makes a better usage of both temporal and view correlations thanks to the hierarchical decomposition. Besides the proposed framework, this paper also investigates MVC coding from the following aspects. Firstly, a disparity-compensated view filter (DCVF) with pixel alignment is proposed, which can accommodate both global and local view disparities among view frames. The proposed DCVF and the existing motion-compensated temporal filter (MCTF) unify the view and temporal decompositions as a generic lifting transform. Secondly, an adaptive decomposition structure based on the analysis of the temporal and view correlations is proposed. A Lagrangian cost function is derived to determine the optimum decomposition structure. Thirdly, the major components of the proposed MVC coding are figured out, including macroblock type design, subband coefficient coding, rate allocation, etc. Extensive experiments are carried out on the MPEG 3DAV test sequences and the superior performance of the proposed MVC coding is demonstrated. In addition, the proposed MVC framework can easily support temporal, spatial, SNR as well as view scalabilities.**

Index Terms—**Multi-view video coding, high dimensional wavelet, lifting, global and local motion alignment.**

I. Introduction

A wide variety of future interactive multimedia applications, including free-viewpoint video or free-viewpoint television, 3D television, immersive teleconference, surveillance, etc, are emerging. The existing and emerging technologies for modeling, coding, and rendering of real world scenes for these applications have been reviewed at the convergence point of computer graphics, computer vision, and classical media [1]. It has been widely recognized that multi-view video coding (MVC) is one of the key technologies for the interactive multimedia applications [2]. Because of the huge data volume, efficient compression becomes more important in storage and delivery of multi-view video than any other data. In most cases, multi-view video is captured by multiple cameras from different angles and locations at the same time instant, and significant correlations may exist between the adjacent views.

In the past decade, various stereoscopic and multi-view video coding techniques have been proposed [3]~[15]. The key point of these techniques lies in how to efficiently utilize the correlations between the adjacent views besides the temporal and spatial correlations within a single view that have been fully studied in the traditional video coding. Some of these techniques only target at multi-view teleconferencing-type sequences [3]~[5]. Particularly, the object-based stereoscopic video coding, which relies on a 3D wire-frame model to describe the foreground object structure, is proposed in [3] and [4]. In [5], the authors proposed a 2D mesh model for multi-view teleconferencing-type sequences and further extended it to MVC. Having targeted at the narrow and specific application, these techniques are unsuitable or inapplicable to generic video contents, because it is hard to accurately segment foreground objects from a natural scene.

The MVC techniques for generic multi-view sequences usually do not take any model either 3D or 2D. Instead, they directly make use of the traditional block-based hybrid video coding, e.g., MPEG-2 multi-view profile (MVP) and MPEG-4 multiple auxiliary components (MAC) [6]~[8]. Based on the frameworks provided by these video standards, some standard-compatible MVC techniques have been proposed in [9]~[11]. Others in [12]~[15] extended the syntaxes and/or semantics of these standards for better coding efficiency. In particular more recently, several MVC techniques based on MPEG-4 AVC/H.264 have been proposed in [13]~[15]. The existing multiple reference structure in this standard can be borrowed to alternatively utilize the motion-compensated and disparity-compensated predictions. It can also benefit from the existing rate-distortion optimization (RDO) technique used in the reference software.

In a summary, most of the previous MVC techniques based on the traditional hybrid video coding have utilized the temporal and view correlations alternatively in the prediction [9]~[15]. Since the temporal correlation is usually stronger than the view one in most natural video sequences, the inter-view

Manuscript received March 17, 2005; revised October 26, 2005. This work was done in Microsoft Research Asia. The associate editor coordinating the review of this manuscript and approving it for publishing was Prof. Jens-Rainer Ohm.

W. Yang was with Nanyang Technological University, Singapore. She is now with the Chinese University of Hong Kong, Hong Kong (email: wxyang@pmail.ntu.edu.sg).

Y. Lu, F. Wu, and S. Li are with Microsoft Research Asia, Beijing, 100080 China (phone: +86-10-58963016; fax: +86-10-88097306; email: yanlu@microsoft.com, fengwu@microsoft.com, spli@microsoft.com).

J. Cai is with the Nanyang Technological University, Singapore 639798 (email: asjfcai@ntu.edu.sg).

K. N. Ngan is with the Chinese University of Hong Kong, Shatin, NT, Hong Kong (email: knngan@ee.cuhk.edu.hk).

prediction often fails in the reference selection. In [16], a sprite-based MVC scheme has been proposed, where the temporal and view correlations are simultaneously utilized by dynamically generating a sprite image as an alternative reference in the predictive coding. However, the dynamic sprite generation is not robust enough for all natural multi-view video sequences. Due to these limitations, a more efficient MVC scheme is desirable not only for coding performance, but should also be extendable, scalable, and flexible for real applications.

Recently, a large amount of efforts have been invested in 3D wavelet scalable video coding (SVC) [17]~[19], where the whole video is decomposed into various temporal-spatial subbands through motion-compensated temporal filtering (MCTF) followed by spatial discrete wavelet transform (DWT). The performance of 3D wavelet-based video coding schemes is comparable to that of the state-of-the-art non-scalable video coding schemes, e.g., MPEG-4 AVC/H.264 [20]. Meanwhile, the temporal, spatial and SNR scalabilities can be easily supported with the subband decomposition and bit-plane coding. Moreover, wavelet transform owns the intrinsic high multi-resolution structure, and can be easily extended to the high dimensional signals with the lifting-based implementation.

Therefore, we propose an elegant MVC framework using the high dimensional wavelet transform in this paper. Although the wavelet coding provides a more natural and flexible way for MVC, it is not a straightforward extension to use it in MVC. The first problem that should be carefully addressed is the wavelet decomposition structure. Multi-view video is inherently high dimensional, with one dimension along the view direction, one dimension along the temporal direction, and two dimensions along the spatial directions. Although the existing MCTF can be directly used for view decomposition, it may be inefficient in exploiting the view correlation because the view disparity is inherently different from the temporal motion. The basic idea of the disparity-compensated lifting transform has been proposed for the light field coding in [21]~[23]. To further extend it to MVC, we propose a disparity-compensated view filter (DCVF) based on a generic lifting structure. The proposed DCVF, the traditional MCTF and the 2D spatial DWT can be jointly utilized to decompose the multi-view video into 4D wavelet coefficients.

Another problem to be carefully addressed here is how to efficiently compress the 4D wavelet coefficients. The basic coding techniques in the proposed MVC framework are derived from the 3-D embedded subband coding with optimal truncation (3D-ESCOT) algorithm in [24] and the later barbell lifting SVC method in [25]. Above all, we design some new macroblock (MB) types and borrow some others from the original MCTF for the proposed DCVF. Since the 4D wavelet coefficients can be generated with any possible decomposition structure, we propose an efficient subband scanning technique to re-organize the 4D wavelet coefficients into 3D data according to the decomposition structure adaptively determined. Thus, the entropy coding in the 3D-ESCOT can be used directly. Moreover, a rate-distortion optimized truncation technique is employed to assemble the compressed bits from different subbands into layers.

The rest of this paper is organized as follows. Section II discusses the generic lifting transform and the proposed DCVF in detail. In Section III, we analyze and measure the correlation in terms of each temporal or view decomposition, and derive a cost function to decide the optimal decomposition structure. In Section IV, we present the proposed MVC method in detail, including the MB type design, subband coding, rate allocation, etc. The experimental results are given in Section V. Finally, Section VI concludes this paper.

II. Lifting-based Disparity Compensated View Filtering

The lifting technique is an efficient implementation of wavelet transform with low memory and computational complexity, which has been incorporated with motion compensation and formed the MCTF by some researchers in the literature [19][26][27]. In this section, we will introduce the generic lifting transform, followed by the proposed DCVF for the view decomposition.

A. Generic Lifting Transform

In general, the lifting-based wavelet transform is composed of two steps: prediction and updating. The former updates the odd frames of a video sequence to high-pass frames, and the latter updates the even frames to low-pass frames. Borrowing the prior experiences on MCTF, we directly use the 5/3 bi-orthogonal filter as an example to discuss the generic lifting transform. Obviously, the generic lifting transform can also use Haar and the 9/7 filter. Let $\boldsymbol{I}_{2k}$ and $\boldsymbol{I}_{2k+1}$ represent the even and odd frames in a video sequence, respectively. Without taking pixel alignment into account, the lifting steps can be formulated by:

$$\begin{cases} \boldsymbol{H}_k[\boldsymbol{p}] = \boldsymbol{I}_{2k+1}[\boldsymbol{p}] - \frac{1}{2}(\boldsymbol{I}_{2k}[\boldsymbol{p}] + \boldsymbol{I}_{2k+2}[\boldsymbol{p}]) \\ \boldsymbol{L}_k[\boldsymbol{p}] = \boldsymbol{I}_{2k}[\boldsymbol{p}] + \frac{1}{4}(\boldsymbol{H}_{k-1}[\boldsymbol{p}] + \boldsymbol{H}_k[\boldsymbol{p}]) \end{cases}, \qquad (1)$$

where $\boldsymbol{p}$ is the pixel location. $\boldsymbol{H}_k$ indicates the high-pass frame and $\boldsymbol{L}_k$ indicates the low-pass frame. The inverse transform can be easily derived from (1).

The generic lifting transform that is applicable to a video sequence requires pixel mapping among adjacent frames, which can be obtained from a motion model. Let $\boldsymbol{M}_{k1 \to k2}$ represents the mapping from one frame $\boldsymbol{I}_{k1}$ to another frame $\boldsymbol{I}_{k2}$, which is usually related to a transform model, e.g. the block-based translational motion model, to compensate for the difference between two frames. Thus, the lifting steps can be further formulated by:

$$\begin{cases} \boldsymbol{H}_k[\boldsymbol{p}] = \boldsymbol{I}_{2k+1}[\boldsymbol{p}] - \frac{1}{2}(\boldsymbol{M}_{2k \to 2k+1}(\boldsymbol{I}_{2k})[\boldsymbol{p}] + \boldsymbol{M}_{2k+2 \to 2k+1}(\boldsymbol{I}_{2k+2})[\boldsymbol{p}]) \\ \boldsymbol{L}_k[\boldsymbol{p}] = \boldsymbol{I}_{2k}[\boldsymbol{p}] + \frac{1}{4}(\boldsymbol{M}_{2k-1 \to 2k}(\boldsymbol{H}_{k-1})[\boldsymbol{p}] + \boldsymbol{M}_{2k+1 \to 2k}(\boldsymbol{H}_k)[\boldsymbol{p}]) \end{cases} . \qquad (2)$$

The high-pass frames corresponding to the predictive residuals will approach zero when the pixels are accurately predicted or aligned. The low-pass frames will be very close to the original frames except some high frequency signals removed. After the lifting-based wavelet transform given in (2), some side information that represents the pixel mapping will also be produced besides the high-pass and low-pass frames. In the sense of compression, the accuracy of pixel mapping and the cost coding side information should be balanced. Therefore, the key problem of the generic lifting transform lies in how to get the pixel mapping $\boldsymbol{M}_{2k\to 2k+1}$ and $\boldsymbol{M}_{2k+1\to 2k}$ that usually describe the same motion but with reverse directions.

The generic lifting transform has been extensively studied and applied to the temporal decomposition of video sequence in the field of SVC [28]. Similar to the traditional hybrid video coding, the block-based translational model is normally adopted in the temporal decomposition, where each block has a rigid translational motion vector [25]. In this case, the pixel mapping can be taken as the block-based motion-compensated prediction (MCP). However, there are still some issues to be considered when the generic lifting transform is used in temporal decomposition. The first one arises from the sub-pixel MCP, which leads to the ambiguity in representing the pixel mapping and its inverse with only one set of rigid motion vectors. The second one lies in that, different from the traditional hybrid coding schemes, both the prediction and updating steps in lifting should be considered in order to achieve high coding efficiency. As claimed in [28], 3D wavelet coding schemes should take more advantages of true motion estimation than the hybrid ones. These open issues likewise exist and may be more difficult to be solved when the generic lifting transform is performed on the video sequence with special motion models, e.g., the multi-view video discussed in this paper.

B. Global Warping-based DCVF

Although the generic lifting transform with the block-based translational model can be directly used for the decomposition of the view video sequence, the performance cannot be guaranteed and may be even far from optimum without considering the inherent features of the multi-view video. The global disparity is one of the major properties of the multi-view video. The view disparity is similar to the temporal motion in the sense that they both represent the displacements between adjacent frames, whereas the properties and the inherent motion model are different. For example, it is well known that the disparity that represents the difference between two adjacent views is usually very compact. With this feature, it is well expectable to deduce a more effective lifting transform for view decomposition, which should tackle the problems similar to that also exist in the MCTF. As a matter of fact, the global disparity has been successfully used in the traditional hybrid MVC schemes. It is also desirable to practice it in the field of lifting-based view filtering.

Therefore, we propose a warping-based lifting transform as DCVF based on the assumption that the disparity between views can be perfectly represented by a global disparity model. Since this assumption is not always true for the natural video due to the distortion of cameras and the scene depth, the practical solution should also consider the local disparity model that will be discussed in the next subsection. Similar to the MCTF, the key problem of the DCVF also lies in the pixel mapping. In the proposed scheme, we use the six-parameter affine transformation to represent the global disparity. Although sometimes the global disparity parameters can be directly derived from the camera parameters, we still employ the global motion estimation (GME) technique proposed in [29] to calculate them for generality. To avoid the effects of distorted image areas from the wide-angle lens of the cameras, a 16-pixel wide boundary of each frame is excluded when calculating the model parameters. Complying with the generalized lifting transform, the lifting steps in the global warping-based DCVF consists of the prediction and updating steps.

In DCVF, $\boldsymbol{I}_{2k+1}$, $\boldsymbol{I}_{2k}$ and $\boldsymbol{I}_{2k+2}$ are frames from different views at the same time instant, where $\boldsymbol{I}_{2k+1}$ comes from an odd view, and $\boldsymbol{I}_{2k}$ and $\boldsymbol{I}_{2k+2}$ come from neighboring even views. In the prediction step, the key problem lies in the definition of $\boldsymbol{M}_{2k\to 2k+1}$ and $\boldsymbol{M}_{2k+2\to 2k+1}$. Here we take $\boldsymbol{M}_{2k\to 2k+1}$ as an example. Suppose $\boldsymbol{p} = (x, y)$ is the location of one pixel in $\boldsymbol{I}_{2k+1}$. The vector $\boldsymbol{X} = \{a_1, a_2, a_3, b_1, b_2, b_3\}^T$ consists of the affine transformation parameters from $\boldsymbol{I}_{2k+1}$ to $\boldsymbol{I}_{2k}$. $\boldsymbol{T}(\boldsymbol{p}, \boldsymbol{X})$ denotes the transformation of $\boldsymbol{p}$ with respect to $\boldsymbol{X}$, resulting in the location $\boldsymbol{p}'$ = (x', y') in $\boldsymbol{I}_{2k}$, i.e. $\boldsymbol{p}' = \boldsymbol{T}(\boldsymbol{p}, \boldsymbol{X})$. Particularly, we have:

$$\begin{pmatrix} x' \\ y' \end{pmatrix} = \begin{pmatrix} a_1 & a_2 \\ b_1 & b_2 \end{pmatrix}\begin{pmatrix} x \\ y \end{pmatrix} + \begin{pmatrix} a_3 \\ b_3 \end{pmatrix}. \tag{3}$$

Further, we can derive the predicting value for the pixel at $\boldsymbol{p}$ in $\boldsymbol{I}_{2k+1}$ from $\boldsymbol{I}_{2k}$ as:

$$\boldsymbol{M}_{2k\to 2k+1}(\boldsymbol{I}_{2k})[\boldsymbol{p}] = \boldsymbol{I}_{2k}[\boldsymbol{T}(\boldsymbol{p}, \boldsymbol{X})]\,. \tag{4}$$

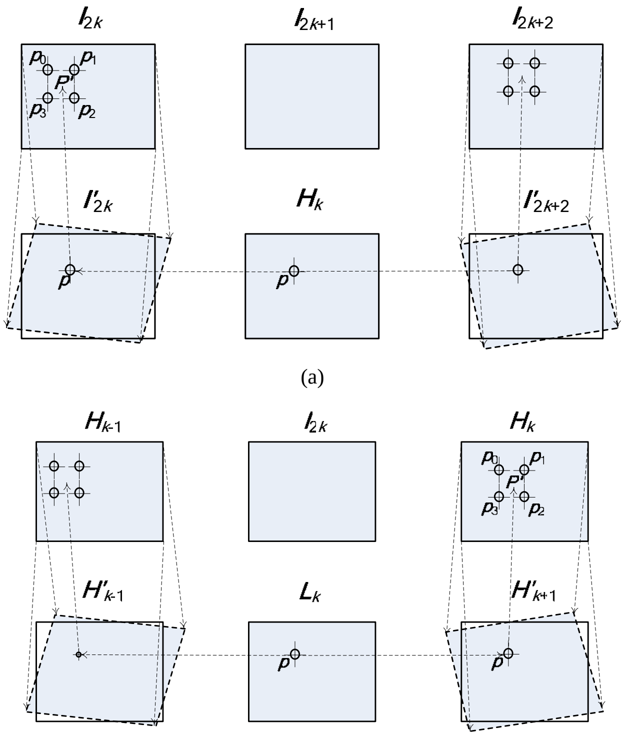
Since (x', y') usually corresponds to a floating coordinate in $\boldsymbol{I}_{2k}$, we calculate $\boldsymbol{I}_{2k}[\boldsymbol{p}']$ with bilinear interpolation. Finally the high-pass frame $\boldsymbol{H}_k$ is produced.

In the updating step, we have to define $\boldsymbol{M}_{2k-1\to 2k}$ and $\boldsymbol{M}_{2k+1\to 2k}$. Here $\boldsymbol{M}_{2k+1\to 2k}$ is taken as an example. Suppose $\boldsymbol{p} = (x, y)$ is the location of one pixel in $\boldsymbol{H}_k$. The vector $\boldsymbol{Y}$ consists of the affine transformation parameters from $\boldsymbol{I}_{2k}$ to $\boldsymbol{H}_k$. Since $\boldsymbol{H}_k$ and $\boldsymbol{I}_{2k+1}$ have the same coordinate system, $\boldsymbol{Y}$ is exactly the inverse of $\boldsymbol{X}$ that has been used in the prediction step. Similarly, we can derive the updating value for the pixel at $\boldsymbol{p}$ in $\boldsymbol{I}_{2k}$ from $\boldsymbol{H}_k$ to:

$$\boldsymbol{M}_{2k+1\to 2k}(\boldsymbol{H}_k)[\boldsymbol{p}] = \boldsymbol{H}_k[\boldsymbol{T}(\boldsymbol{p}, \boldsymbol{Y})]\,. \tag{5}$$

Here we also calculate $\boldsymbol{H}_k[\boldsymbol{T}(\boldsymbol{p}, \boldsymbol{Y})]$ with bilinear interpolation. Finally, the low-pass frame $\boldsymbol{L}_k$ is produced.

The above prediction and updating steps compose the global warping-based DCVF. Fig. 1 intuitively illustrates how the two lifting steps are performed based on the global warping. For the



(a)

(b)

Fig. 1. Global warping-based DCVF, including (a) prediction, and (b) updating.

prediction step, $\boldsymbol{I}_{2k}$ is first warped towards $\boldsymbol{I}_{2k+1}$ with respect to the affine transformation parameters so as to generate the warped frame $\boldsymbol{I}'_{2k}$. Low-pass extrapolation (LPE) padding scheme is employed to fill the blank regions in $\boldsymbol{I}'_{2k}$ that do not correspond to any pixels in $\boldsymbol{I}_{2k}$. After the warping, $\boldsymbol{I}'_{2k}$ and $\boldsymbol{I}_{2k+1}$ are in the common coordinate system. Then, the prediction of the pixel in $\boldsymbol{I}_{2k+1}$ can be directly derived from the co-located pixel in $\boldsymbol{I}'_{2k}$. The updating step can also be performed based on the global warping. The proposed global warping-based DCVF can guarantee perfect reconstruction, namely, transformed disparity frames with quantization can perfectly reconstruct original frames by corresponding inverse transform.

C. Adaptive Global and Local Disparity Model in DCVF

As discussed above, the global warping-based DCVF owns many advantages in the decomposition of the view video sequence. These advantages base on the assumption that the disparity between two adjacent views can be perfectly represented by the global disparity model. However, any global disparity model cannot guarantee the perfect matching for all regions, even though in most cases they dominate the accuracy of pixel mapping between the two views. As verified in the traditional MVC schemes with disparity-compensated prediction, considering the local disparity information can further refine the matching accuracy between two adjacent views. It is also true in the lifting-based view filtering. Therefore, we employ the adaptive global and local disparity model in the proposed lifting-based DCVF.

Similar to the previous schemes, we also divide the frame into MBs. Thus, pixels of each MB can have its unique motion data, and hereby the pixel mapping of one MB may differ from another. Each MB corresponds to either local disparity model or global disparity model. If the local disparity mode is selected for an MB, the local disparity vectors equivalent to the local motion vectors of the block-based translational model are used to represent the pixel mapping, and the MB is referred to as the local disparity compensation (LDC) MB hereafter; otherwise, the affine transformation parameters with respect to the whole frame are used, and the MB is referred to as the global disparity compensation (GDC) MB hereafter. The MB type can be selected based on the tradeoff between the matching accuracy and the coding cost of side information. Actually, more MB modes should be defined by further classifying the LDC and GDC MBs for the purpose of improved coding efficiency, as will be discussed in Section IV.

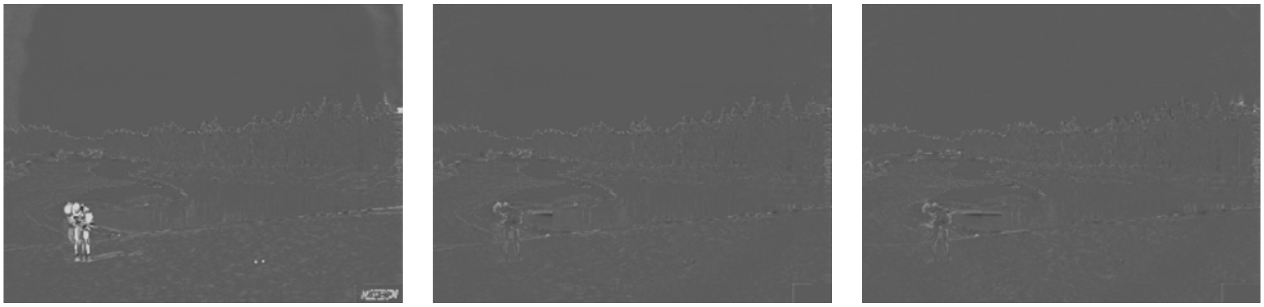
Now, the problem is how to implement the lifting steps on the LDC and GDC MBs within the same framework. The prediction step on the LDC MBs is the same as that in MCTF. In particular, we employ the barbell lifting transform [25] for the LDC MBs. The prediction step on the GDC MBs involves the affine disparity model and bilinear interpolation, which can be taken as one kind of the barbell lifting transforms. The warping of the entire reference frame is only helpful for the MB type selection but unnecessary in the actual prediction step. In this way, the prediction steps on the LDC and GDC MBs can be implemented within the same framework but with the different disparity parameters.

In the updating step, we employ the energy distributed update (EDU) technique developed in [30], in which every pixel in the high-pass frame $\boldsymbol{H}_k$ (corresponding to $\boldsymbol{I}_{2k+1}$) updates at most four pixels in $\boldsymbol{I}_{2k}$ (corresponding to the low-pass frame $\boldsymbol{L}_k$) under the energy distribution constraint. To implement the updating steps on the LDC and GDC MBs within the same framework, we derive the EDU-based updating step for the GDC MBs as follows. Suppose $\boldsymbol{p} = (x, y)$ denotes the location of one pixel in a GDC MB of the high-pass frame $\boldsymbol{H}_k$. Thus, $\boldsymbol{p}$ corresponds to a floating location $\boldsymbol{p}'$ in $\boldsymbol{I}_{2k}$ with respect to the affine transformation parameters $\boldsymbol{X}$ from $\boldsymbol{H}_k$ (i.e., $\boldsymbol{I}_{2k+1}$) to $\boldsymbol{I}_{2k}$. According to the routine of EDU-based updating, $\boldsymbol{H}_k[\boldsymbol{p}]$ should be scaled and added to each of the four pixels around $\boldsymbol{p}'$ in $\boldsymbol{I}_{2k}$, i.e., $\boldsymbol{p}_i'$, $i = 0, \ldots, 3$, where the scaling weight w_i depends on the distance between $\boldsymbol{p}_i$ and $\boldsymbol{p}$. We calculate it as:

$$w_i = \begin{cases} (1-|x-x_i|)\cdot(1-|y-y_i|) & \text{if } |x-x_i| \le 1 \ \& \ |y-y_i| \le 1 \\ 0 & \text{otherwise} \end{cases}, \quad (6)$$

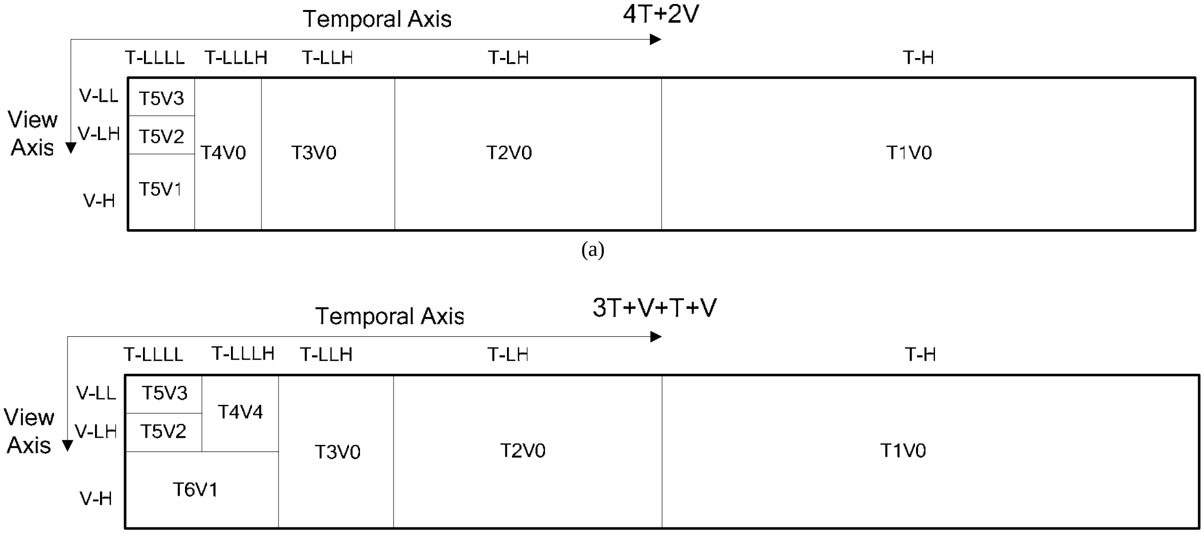
$w_i = 0$ indicates that $\boldsymbol{H}_k[\boldsymbol{p}]$ does not update the pixel at location $\boldsymbol{p}_i'$ in $\boldsymbol{I}_{2k}$.

Fig. 2 shows the high-pass frames produced from one-level decomposition of a view sequence, using the DCVF with GDC only, the DCVF with both GDC and LDC and the MCTF, respectively. As shown in Fig. 2(a), the energy at the foreground region is still relatively large because the affine transformation model cannot well compensate foreground



(a) (b) (c)

Fig. 2. High-pass frames from the one-level decomposition of the view sequence Golf2, using (a) DCVF with GDC only, (b) DCVF, and (c) MCTF.



(a)

(b)

Fig. 3. Illustration of decomposition structures along the temporal and view directions: (a) 4T+2V and (b) 3T+V+T+V.

regions due to the different scene depths. In this case, the LDC mode is required. As shown in Fig. 2(b), the whole frame including both the foreground and background regions presents very low energies, which is desirable in the sense of coding efficiency. Moreover, while the high-pass frame produced using the DCVF has the energies no more than that produced using the MCTF, as shown in Fig. 2(c), the DCVF also requires much fewer disparity parameters than the MCTF.

III. Adaptive Decomposition Structure

As discussed above, the proposed DCVF and the traditional MCTF can be adaptively utilized to decompose the multi-view video along the temporal and view directions, respectively. The problem raised here is the efficiency of the overall decomposition structure. In this section, we will first present a regular decomposition structure, and then propose a more efficient decomposition structure derived from the correlation analysis.

A. Regular 4D Wavelet Decomposition

The previous investigation on the 3D wavelet-based video coding has proven that it is only necessary to perform the wavelet decomposition on the original or the low-pass subband signals in the multi-level MCTF. Accordingly, it is also unnecessary to decompose the high-pass subband signals in the multi-level DCVF. Since the temporal correlation of the multi-view video is usually stronger than the view correlation, we have proposed a regular decomposition structure, in which the DCVF is done after the multi-level MCTF has fully extracted the temporal correlations [31]. Fig. 3(a) shows an example, which consists of four levels of MCTF followed by two levels of DCVF. For convenience, we refer to this decomposition structure as 4T+2V, where T and V represent the temporal decomposition and the view decomposition, respectively. Obviously, only the last subband, e.g., the subband at (T-LLLL, V-LL) in Fig. 3(a), is composed of the low-pass frames, and all other subbands of the high-pass frames.

However, the above decomposition structure cannot always achieve the optimized coding efficiency, because the assumption that the temporal correlation is always stronger than the view one is not always true. Moreover, it is also very hard to decide the actual number of levels of MCTF as well as that of DCVF prior to the coding. A reasonable way to achieve the best coding efficiency is to derive the optimum decomposition structure based on the coding cost. Suppose that

TABLE I
THE COST VALUES FOR THE ADAPTIVE DECOMPOSITION STRUCTURE

Sequence	Direct.	1st	2nd	3rd	4th	5th	6th	Structure
Flamenco1	Tempo.	305.2	416.6	564.9	735.9	745.8	926.6	3T+V+2T
	View	732.2	722.7	715.3	714.9	970.4	975.2	
Golf2	Tempo.	246.9	233.7	308.8	372.5	689.5	1119.0	5T+V
	View	902.9	900.1	900.8	898.0	903.9	914.7	
Race1	Tempo.	531.3	759.0	1089.2	1486.7	1508.6	1981.0	3T+V+T+V
	View	1201.5	1201.9	1201.9	1202.5	1595.5	1577.0	

the MCTF is performed at the beginning. The temporal distance between two low-pass frames increases by a factor of two after one-level MCTF, and the average coding cost in terms of the following MCTF increases as well. Since the MCTF and the DCVF are approximately independent, the average coding cost in terms of the DCVT almost remains the same. Therefore, it is possible that the DCVF will be selected after one or several levels of MCTF. Moreover, it is also possible that the MCTF will be selected once again after one or several levels of DCVF. In most cases, the MCTF and the DCVF should be interleaved based on coding cost.

B. *Adaptive Decomposition Structure*

As discussed above, it is reasonable to have the interleaved decomposition structure in the sense of coding efficiency. The temporal first (T-first) and view first (V-first) decomposition structures can be taken as the special cases of the interleaved decomposition structure. The remaining problem is how to define a criterion to decide the optimum decomposition structure. Similar to the motion compensation, it is desirable that the high-pass subband samples are close to zero in the temporal or view decomposition. Then, we define the cost function:

$$HCost = SAD(H) + \lambda R(MV) \ , \tag{7}$$

to evaluate the efficiency of de-correlation after one-level decomposition. $SAD(H)$ is the sum of absolute difference (SAD) of samples in the high-pass subband H obtained from the decomposition. $R(MV)$ is the bit rate for coding the motion or disparity information. λ is the Lagrangian parameter, which is the same as that used in rate-distortion optimized motion/disparity estimation.

Based on the analysis of the coding cost, the decomposition structure is adaptive to video content, so as to achieve high coding efficiency. We take the decomposition structure shown in Fig. 3(a) as an example. After the first three levels of MCTF, the multi-view low-pass subband subsequence is obtained. Then, the motion between the adjacent frames becomes very large, which makes the next MCTF inefficient. Perhaps, performing the DCVF is a better choice than another MCTF at this time. In other words, the optimum decomposition structure can be obtained through alternatively selecting the decomposition that can de-correlate the multi-view video to the maximum extend. Fig. 3(b) depicts an example of the proposed adaptive decomposition structure, where the MCTF and the DCVF are interleaved together.

Table I shows some statistics from the MPEG testing multi-view video sequences. The cost values of the temporal and view decompositions at each level are respectively calculated using (7). These results are obtained on a group of GOPs (GoGOP), which contains the GOPs in all views at the same time interval. The view or temporal decomposition with the smallest cost value is selected for the current level of decomposition. Then, the next level of decomposition can be performed on the current low-pass subband frames. Obviously, the statistics in Table 1 show that the cost value of the view direction remains almost unchanged if temporal decomposition is selected, and the cost value of the temporal direction also remains unchanged if view decomposition is selected. Conclusively, the statistics comply with the above analysis on the proposed adaptive decomposition structure in terms of coding efficiency.

IV. SUBBAND MULTI-VIEW VIDEO CODING

Based on the adaptive decomposition structure, we propose an efficient MVC scheme to enhance the functionalities of flexibility, extensibility and scalability. In this section, we will first give an overview of the proposed scheme, and then discuss the MB mode design and the subband coding in detail.

A. *The Proposed MVC Scheme*

The proposed MVC scheme is built up upon the wavelet-based scalable video coding in [25], which has been adopted as the common reference software for the MPEG exploration activity on inter-frame wavelet coding. The overall structural diagram of the proposed scheme is shown in Fig. 4. Particularly, the adaptive lifting-based wavelet decomposition is performed on the GoGOP basis, which contains the GOPs in all views at the same time interval. After the coding cost is calculated at each level of decomposition, either DCVF or MCTF is performed, resulting in the low-pass and high-pass subbands. The low-pass subband can be further decomposed until the last level of decomposition is reached. The high-pass subbands and the last low-pass subband are sent to the buffer for the spatial wavelet decomposition, followed by the entropy coding. Note that in the proposed DCVF, more MB modes are designed in terms of the LDC and GDC MBs. The mode and motion/disparity information associated with each MB are also encoded. 3D-ESCOT in [24] is extended to code the 4D wavelet coefficients. Moreover, a rate-distortion weight is assigned to each temporal-view-spatial subband for overall bit allocation.

B. *MB Modes*

It is a common way to define a number of MB modes in terms of the motion-compensated prediction types in most video coding schemes. The MB modes in the conventional MCTF include *FwD*, *BwD*, *DirecT*, *Skip* and *BiD* [25]. Due to the existence of the GDC MBs in the proposed DCVF, it is inefficient to directly employ the MB modes designed for the MCTF. Instead, we introduce some new MB modes for the GDC MBs and exclude some others for the LDC MBs.

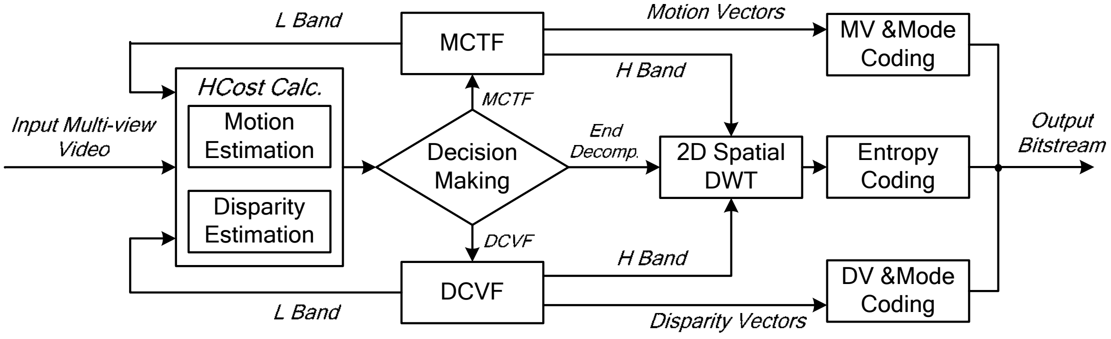


Fig. 4. Block diagram of the proposed MVC coding scheme.

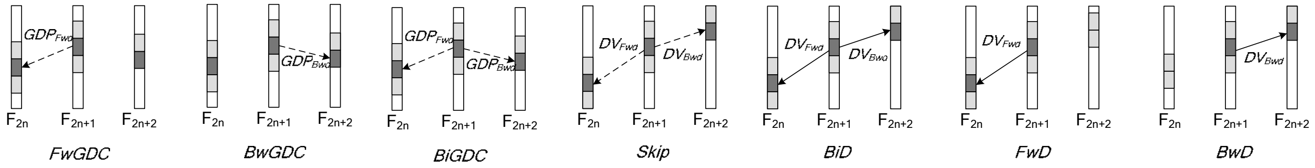
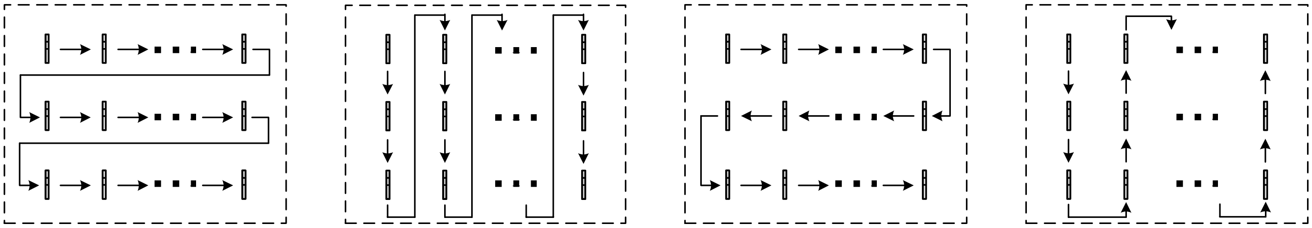


Fig. 5. Illustration of seven MB modes in DCVF.



(a) (b) (c) (d)

Fig. 6. Scan orders of a temporal-view subband. (a) T-scan; (b) V-scan; (c) head-to-tail scan for MCTF-decomposed subband; and (d) head-to-tail scan for DCVF-decomposed subband.

TABLE II
WEIGHT MATRIX FOR GLOBAL RATE-DISTORTION OPTIMIZATION TRUNCATION

	V (w/o filt.)	V-H	V-LH	V-LL
T-H	0.7188	---	---	---
T-LH	0.9219	---	---	---
T-LLH	1.5859	---	---	---
T-LLLH	3.0430	---	---	---
T-LLLL	---	7.6822	9.8525	29.3906

Particularly, we introduce three new MBs modes for the GDC MBs, i.e., *FwGDC*, *BwGDC* and *BiGDC*, where the forward, backward and bidirectional global disparity-compensated predictions are used, respectively. Since the disparity information can be derived from the global disparity parameters coded at the frame level, no local disparity vectors are encoded for any of the three modes. Moreover, we exclude the *DirecT* mode while preserving the other four modes for the LDC MBs because the disparity in the consecutive view frames is not as continuous as the motion of the temporal frames. As a summary, there are totally seven MB modes designed for the GDC and LDC MBs in the proposed DCVF, as shown in Fig. 5.

The rate-distortion optimization (RDO) technique in [25] is employed to select the best MB mode. To follow the syntax in the original codec, we also employ the Golomb-Rice codes for the MB mode coding. Towards this goal, the seven modes are ordered based on the probabilities in statistics. Since *FwGDC* and *BwGDC* have very similar probabilities in statistics, it does not make sense to assign different code lengths to them. Accordingly, we merge the two modes into one mode *SiGDC*. If *SiGDC* is coded, an additional flag has to be signaled to indicate the alternative of the two modes. Moreover, when the right reference is unavailable, only *FwD* and *FwGDC* modes are used, and one bit is enough to signal the MB mode. In the temporal decomposition, only few frames do not have the right references. However, in the view decomposition, the number of frames not having the right references may occupy a large portion. For example, for the two-level DCVF of an eight-frame view sequence, one third of the prediction steps may not have the right references.

C. Subband Coding

Based on the proposed MVC framework, the multi-view video can be decomposed into 4D wavelet coefficients. For entropy coding, one of the key problems is how to efficiently organize these coefficients. As shown in Fig. 3, every

(a) (b) (c)

Fig. 7. The 1st frame of the 3rd view of (a) Flamenco1, (b) Golf2, and (c) Race1.

TABLE III
TEST SEQUENCES

Sequence	Class	Image Property	Camera Parameter	Bit rate (kbps)
Flamenco1	A (easy)	S: 320x240 F: 30fps	N: 8, D: ~20cm, A: 1D parallel	64, 128, 192, 256
Golf2	A (easy)	S: 320x240 F: 30fps	N: 8, D: ~20cm, A: 1D parallel	64, 128, 192, 256
Race1	B (difficult)	S: 320x240 F: 30fps	N: 8, D: ~20cm, A: 2D parallel	128, 256, 384, 512

TABLE IV
MPEG-4 AVC/H.264 CODING PARAMETERS

Feature / Tool / Setting	Coding Parameters
Rate control	Yes
R-D optimization	Yes
Specific settings	Loop filter, CABAC
Search range	±32 for CIF/VGA
# Reference picture	5
I-frame period	128 frames
GOP Structure	IBBP…

temporal-view subband from the adaptive DCVF and MCTF is composed of two dimensional coefficient frames, i.e., one temporal dimension and one view dimension. For simplicity, we scan these coefficient frames into one sequence. In other words, we re-organize the 4D wavelet coefficients into 3D representations, so that the entropy coding in 3D-ESCOT can be directly utilized. Since the 3D-ESCOT encodes one coefficient frame based on the context from its adjacent frames, the scan order may have a close impact on the context model as well as the coding efficiency. After carefully studying the various scan orders, we propose an efficient scan scheme in terms of the actual decomposition structure.

Particularly, we consider three scan orders, i.e. the T-scan, the V-scan and the head-to-tail scan, as shown in Fig. 6. For the subband generated from the MCTF, the temporal correlation is stronger than the view correlation. For the subband generated from the DCVF, the view correlation is stronger than the temporal correlation. Therefore, the efficiency of T-scan or V-scan depends on the actual decomposition structure. Moreover, both T-scan and V-scan contain the transition points with very large skips. Obviously, the entropy coding is inefficient if the context crosses these transition points. Comparing with the T-scan or V-scan, we observe a performance gain of the head-to-tail scan, in which the subband generated from MCTF or DCVF is scanned in temporal or view order for the even rows or columns and in inverse temporal or view order for the odd rows and columns.

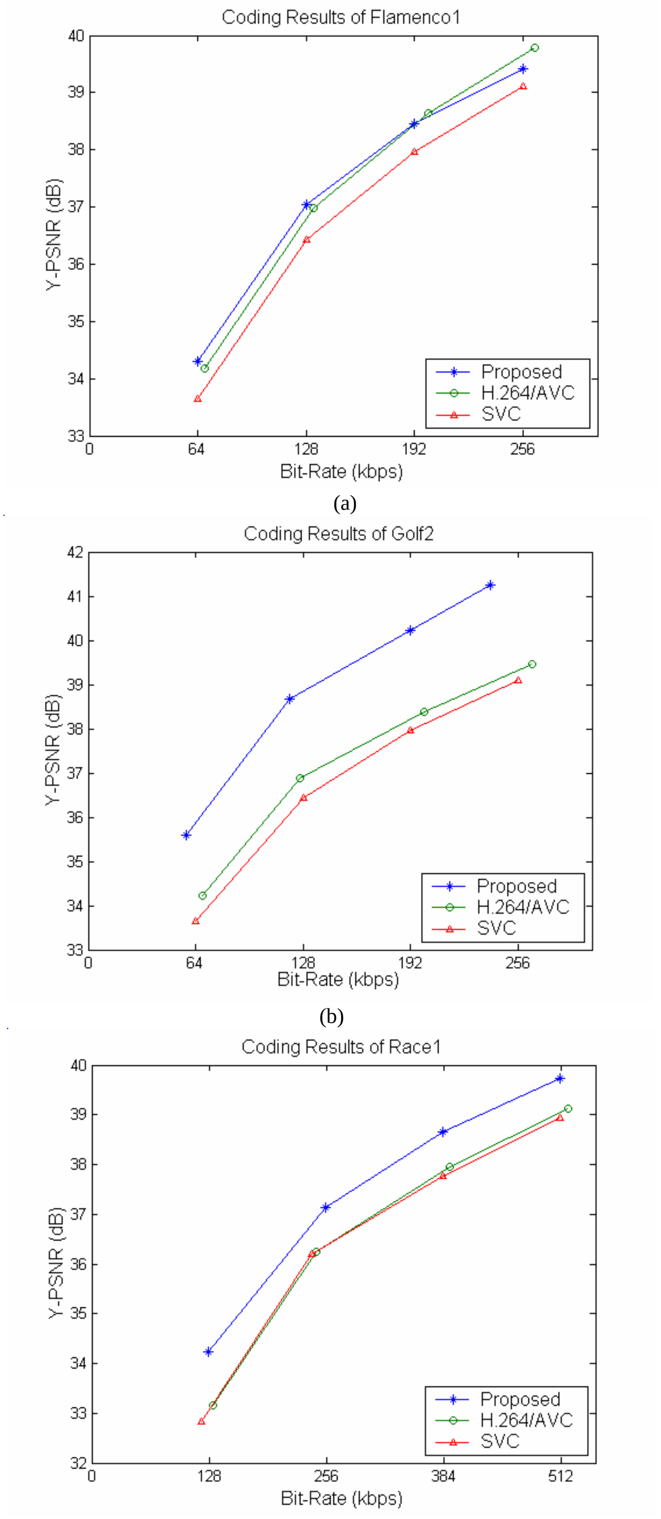
A rate-distortion (R-D) optimized truncation technique is used to assemble the compressed bits into multiple layers so as to improve the overall coding performance. The basic idea is to truncate the compressed bits of the different subbands at the same slope of the R-D curves by a weight matrix. The definition of the weight matrix depends on the actual decomposition structure. We take the 4T+2V structure as an example. Suppose the 5/3 filter, i.e., (1/2, 1, 1/2) for the low-pass synthesis and (-1/8, -1/4, 3/4, -1/4, -1/8) for the high-pass synthesis, is used for both MCTF and DCVF. Based on these filter coefficients, the transform gain of every subband can be calculated in terms of the one-level or multi-level DWT. For example, W_L = 1.5 and W_H = 0.7188 for the low-pass and high-pass subbands in the one-level DWT; and W_{LLLL} = 10.6875 and W_{LLLH} = 3.0430 for the low-pass and high-pass subbands in the four-level DWT. To extend it to the 4D wavelet decomposition structure, the weight of the subband from both temporal and view decompositions is defined as the multiplication of the two weights from the temporal and view decompositions. For example, the weight of subband (T-LLLH, V-H) is defined as $W_{LLLH} \times W_H = 7.6822$. Table II gives the weight for every subband from the 4T+2V decomposition.

V. EXPERIMENTAL RESULTS

A. *Evaluation of Overall Coding Performance*

To verify the performance of the proposed MVC scheme, we present the experimental results on three MPEG 3DAV test multi-view videos [32], as described in Table III. Each multi-view video includes 8 views with 256 frames per view. Some sample frames from the three multi-view videos are shown in Fig. 7. In the testing of the proposed MVC scheme, the adaptive decomposition structure is used. The temporal GOP size is set to be 128, and the maximum number of levels for the temporal decomposition can be up to 7. The view GOP size is set to be 8, and the maximum number of levels for the view decomposition can be up to 3. In particular, we set the total number of levels for the temporal and view decompositions to be 6.

Above all, the proposed MVC scheme is compared with the



(a)

(b)

(c)

Fig. 8. Comparison of coding performance with simulcast coding schemes for (a) Flamenco1, (b) Golf2, and (c) Race1.

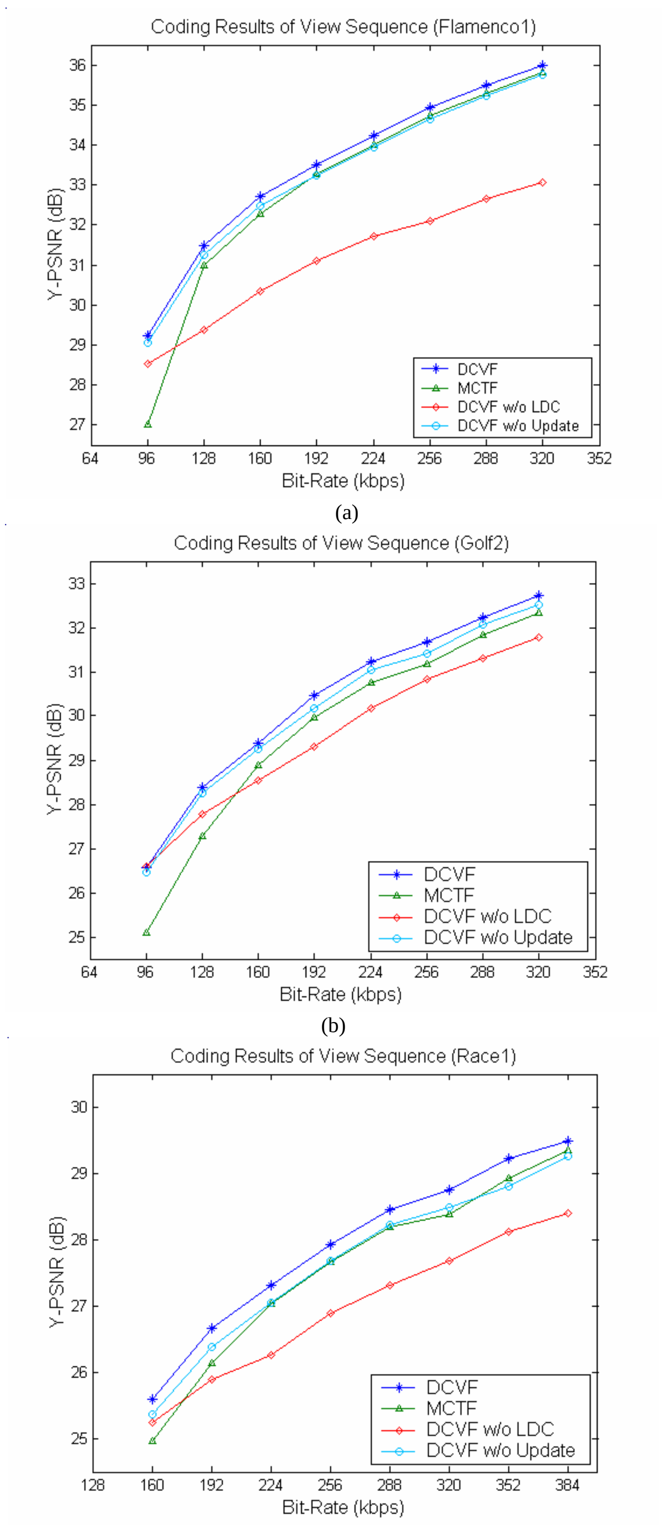
TABLE V
CODING PERFORMANCE OF THE THREE DIFFERENT DECOMPOSITION STRUCTURES

Sequence	Decomposition Structure	Bit-rate	PSNR	*HCost*
Flamenco1 128kbps	T-first (4T+2V)	127.92	38.03	483.90
	V-first (2V+4T)	128.00	34.89	965.39
	T-only (6T)	127.98	37.60	669.21
	Interleaved (3T+V+2T)	127.96	38.14	465.62
Golf2 128kbps	T-first (4T+2V)	127.94	39.56	332.79
	V-first (2V+4T)	127.89	35.20	806.62
	T-only (6T)	127.96	39.54	493.99
	Interleaved (5T+V)	127.99	39.76	324.30
Race1 256kbps	T-first (4T+2V)	255.85	34.17	572.21
	V-first (2V+4T)	255.98	32.46	1165.93
	T-only (6T)	255.83	33.41	1117.21
	Interleaved (3T+V+T+V)	255.96	34.28	562.98

SVC-based simulcast video coding, in which every view is independently coded using the SVC codec in [25]. We set the same test conditions for the proposed MVC scheme and the original SVC codec. Either scheme generates only one bit-stream for one multi-view video, and achieve the coding results at all testing rates with the bit-stream truncation. Fig. 8 shows the rate-distortion curves with respect to the results on the three multi-view videos. We can observe that the proposed algorithm always outperforms the SVC simulcast video coding. The coding gain is around 1 dB in average and can be up to 2 dB at low rates. Actually, the coding gain of the proposed algorithm mainly comes from the adaptive decomposition structure with the proposed DCVF. Since the global disparity model in the proposed DCVF can achieve more coding gains at lower rates, sometimes the overall coding gain of the proposed MVC scheme slightly decreases at high rates.

Moreover, the proposed MVC scheme is also compared with the MPEG-4 AVC/H.264 based simulcast video coding, in which the test model JM7.6 is selected as the reference software. Table IV shows the coding parameters used to generate the H.264 anchors [32]. For the fair comparison, the H.264 anchors are also encoded with an I-frame period of 128 frames. The rate-distortion curves in terms of these H.264 anchors are also shown in Fig. 8. The coding gain of the proposed MVC scheme over the H.264 simulcast coding is 1.3 dB in average and can be up to 2 dB. However, for Flamenco1, the H.264 simulcast coding slightly outperforms the proposed scheme at the highest rate. Actually, at this testing rate, the H.264 simulcast coding also outperforms the SVC simulcast coding about 1 dB. We observe that there exists some illumination changing across the temporal frames in Flamenco1. In this case, the MCTF in SVC is less efficient than the directional spatial prediction in H.264.

The memory cost and computing complexity are discussed as follows. Since the simulcast coding scheme can off-line process the multi-view video view by view, its memory cost is always less than the proposed MVC scheme. Here we only discuss the case of processing the multiple views simultaneously. In this case, if the SVC simulcast coding uses hierarchal prediction structure in the view and temporal directions, the memory costs of the SVC simulcast coding and the proposed MVC scheme are very similar because both of



(a)

(b)

(c)

Fig. 9. Evaluation of the proposed DCVF. (a) Flamenco1, (b) Golf2, and (c) Race1.

schemes have the same decomposition structure. If the SVC simulcast uses P-prediction structure on some view and temporal frames, the memory costs are fewer than that of the proposed scheme. Compared with the H.264 simulcast coding, the proposed scheme needs more memory due to the hierarchical wavelet decomposition, which also is one major problem in the existing wavelet-based SVC schemes. As for the complexity issue, the extra cost of the proposed MVC scheme over the SVC simulcast coding mainly lies in the encoder part, i.e., selecting the optimum decomposition structure, which can be further simplified in future works. Moreover, compared with the H.264 coding with 5 reference frames, our testing has shown that the two schemes have the similar running time.

B. Evaluation of Decomposition Structure

Since the proposed MVC scheme supports the adaptive decomposition structure, the coding performances with respect to the different decomposition structures are also evaluated, including the T-first, the V-first and the interleaved structures. The T-first or V-first structure can be taken as the special case of the interleaved structure. However, prior to the coding, we do not know how many levels of temporal and view decompositions are good enough in terms of the T-first and V-first structures. Particularly, we employ the 4T+2V and T-only (i.e. 6T) structures to evaluate the T-first case, and the 2V+4T structure to evaluate the V-first case. In the evaluation, the interleaved decomposition structure is obtained based on the cost function defined in Section III. Table 1 in Section III has shown the cost value at each level of decomposition in terms of the interleaved structure, which is calculated based on one GoGOP of every multi-view video. The experiments in this subsection are also performed on the same GoGOP.

Table V gives the distortion and the final cost value with respect to every decomposition structure at the same bit rate, which indicates that the cost value approximately represents the de-correlation efficiency as well as the final rate-distortion performance. As shown in Table 5, the interleaved decomposition structure has the best rate-distortion performance. Actually, supposing the decomposition structures are very close, the corresponding cost values as well as the overall rate-distortion performances are also very close, e.g. the 4T+2V structure and the 3T+V+T+V structure for Race1. In our experiments, the interleaved structure is decided by the cost function with respect to the Lagrange multiplier. The optimum decomposition structures derived from the cost function targeting at the different bit rates may be a little bit different. However, the effects on the overall coding efficiency are very limited. Therefore, we propose to the use the same decomposition structure that is selected targeting at a preferred bit rate for the coding of the multi-view video at any other bit rates.

C. Evaluation of DCVF

The efficiency of the proposed DCVF for the view decomposition is also evaluated, in which the MCTF is taken as a reference. The experiments are performed on the view sequences extracted from the above three multi-view videos. Fig. 9 shows the testing results. Obviously, the proposed DCVF outperforms the MCTF in all experiments. The coding gain can be up to 2 dB at low bit rates. It is reasonable because the global disparity model in the DCVF can save more bits of coding the disparity information. The lifting-based DCVF without updating step is also tested, whose performance is always worse than the DCVF with updating step, as shown in Fig. 9. Moreover, the DCVF without LDC modes is tested as well. As shown in Fig. 9, the gap between the DCVF with and without LDC modes increases while the bit rate increases, which indicates that the LDC in DCVF works well at the high bit rates and the GDC works well at the low bit rates.

VI. Conclusions

In this paper, we have proposed a flexible MVC framework using high dimensional wavelet, which rightly matches the inherent high dimension property of multi-view video. In particular, we have investigated the MVC coding from the following aspects. Firstly, a disparity compensated view filter with pixel alignment is proposed here, which can accommodate both global and local view disparities among view frames. The proposed DCVF and the existing MCTF unifies the view and temporal decompositions as a generic lifting transform. Secondly, a flexible decomposition structure based on the analysis of the temporal and the view correlations is proposed. The Lagrangian cost function is derived to determine the optimum decomposition structure. Thirdly, the major components of the proposed MVC coding are developed, including macroblock type design, subband coefficient coding, rate allocation, etc. Extensive experiments have been carried out on MPEG 3DAV test sequences and the superior performance of the proposed MVC framework has been demonstrated.

There also remains some future work. The proposed MVC scheme inherently supports the SNR scalability. One of its extensions is to support other scalabilities such as view scalability. Other future research directions include the further improvement of coding performance. Recently, an H.264-based MVC scheme has been proposed to MPEG, in which the inter-view-temporal prediction structure with hierarchical B pictures is utilized [33]. Actually, the hieratical B prediction structure across views is similar to the structure of view decomposition. Therefore, the proposed algorithms such as the adaptive decomposition structure may be suitable to be applied in the H.264-based MVC scheme; and meanwhile, some efficient modules in the H.264-based MVC scheme may also be suitable in the proposed 4D wavelet-based MVC scheme. The combination of the two MVC schemes is worthy of further study.

References

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

[2] ISO/IEC JTC1/SC29/WG11 N6501, “Requirements on multi-view video coding,” Redmond, USA, July 2004.

[3] S. Malassiotis and M. G. Strintzis, “Object-based coding of stereo image sequences using three-dimensional models,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 7, no. 6, pp. 892-905, Dec. 1997.

[4] E. Izquierdo and J.-R. Ohm, “Image-based rendering and 3D modeling: a complete framework,” *Signal Processing: Image Communication*, vol. 15, pp. 817–858, 2000.

[5] R. S. Wang, Y. Wang, “Multi-view video sequence analysis, compression, and virtual viewpoint synthesis,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp.397–410, April 2000.

[6] J.-R. Ohm, “Stereo/multiview encoding using the MPEG family of standards,” in *Proc. SPIE, Stereoscopic Displays and Virtual Reality Systems VI*, vol. 3639, pp. 242–253, Jan. 1999.

[7] MPEG-2 Multiview Profile, ISO/IEC 13818-2: AMD 3, 1996.

[8] Generic coding of audio-visual objects – Part 2: visual, ISO/IEC 14496-2: 2001, 2001.

[9] A. Puri, R. V. Kollarits and B. G. Haskell, “Basics of stereoscopic video, new compression results with MPEG-2 and a proposal for MPEG-4,” *Signal Processing: Image Communication*, vol. 10, pp.201-234, 1997.

[10] J. Lim, K. Ngan, W. Yang and K. Sohn, “Multiview sequence CODEC with view scalability,” *Signal Processing: Image Communication*, vol. 19, no. 3, pp. 239–256, Mar. 2004.

[11] W. Yang and K. Ngan, “MPEG-4 based stereoscopic video sequences encoder”, in *Proc. ICASSP 2004*, vol. 3, pp. 741–744, May 2004.

[12] Y. Choi, S. Cho, J. Lee, C. Ahn, “Field-based stereoscopic video codec for multiple display methods,” in *Proc. ICIP 2002*, vol. 2, pp. 253–256, NY, USA, Sept. 2002.

[13] X. Guo, Q. Huang, “Multiview video coding based on global motion model,” *Lecture Notes in Computer Sciences*, vol. 3333, pp. 665–672, Dec. 2004.

[14] G. Li, Y. He, “A novel multiview video coding scheme based on H.264,” in *Proc. ICICS-PCM 2003*, pp. 493–497, Singapore, Dec. 2003.

[15] ISO/IEC JTC1/SC29/WG11 M11700, “Responses received to CfE on multi-view video coding,” Hong Kong, China, Jan. 2005.

[16] N. Grammalidis, D. Beletsiotis, and M. G. Strintzis, “Sprite generation and coding in multiview image sequences,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 10, pp. 302–311, Mar. 2000.

[17] A. Secker and D. Taubman, “Motion-compensated highly scalable video compression using an adaptive 3D wavelet transform based on lifting,” in *Proc. ICIP 2001*, vol. 2, pp 1029-1032, Greece, Oct. 2001.

[18] L. Luo, J. Li, S. Li, Z. Zhuang, and Y.-Q. Zhang, “Motion compensated lifting wavelet and its application in video coding,” in *Proc. ICME 2001*, pp 365-368, Tokyo, 2001.

[19] L. Luo, F. Wu, S. Li, Z. Zhuang and Y.-Q Zhang, “A two-pass optimal motion threading technique for 3D wavelet video coding,” in *Proc. ISCAS 2002*, vol. 4, pp. 819–822, May 2002.

[20] R. Xiong, F. Wu, et al, “Exploiting temporal correlation with block-size adaptive motion alignment for 3D wavelet coding,” in *Proc. of SPIE VCIP 2004*, San Jose, CA, Jan. 2004.

[21] B. Girod, C.-L. Chang, P. Ramanathan, and X. Zhu, “Light field compression using disparity-compensated lifting,” in *Proc. IEEE ICASSP 2003*, vol. 4, pp. 760-763, Hong Kong, China, April 2003.

[22] C.-L. Chang, X. Zhu, P. Ramanathan, and B. Girod, “Inter-view wavelet compression of light fields with disparity-compensated lifting,” *in Proc. SPIE VCIP 2003*, pp.14–22, Lugano, Switzerland, July 2003.

[23] X. Tong, R. M. Gray, “Interactive rendering from compressed light fields,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 13, pp. 1080–1091, Nov. 2003.

[24] J. Xu, Z. Xiong, S. Li and Y.-Q. Zhang, “Three-dimensional embedded subband coding with optimized truncation (3D ESCOT),” *J. Applied and Computational Harmonic Analysis*, vol. 10, pp. 290–315, May 2001.

[25] R. Xiong, F. Wu, J. Xu, S. Li and Y.-Q. Zhang, “Barbell lifting wavelet transform for highly scalable video coding”, in *Proc. PCS 2004*, Dec. 2004.

[26] B. Pesquet-Popescu and V. Bottreau, “Three-dimensional lifting schemes for motion compensated video compression,” in *Proc. ICASSP 2001*, vol. 3, pp 1793–1796, Salt Lake City, USA, 2001.

[27] A. Secker and D. Taubman, “Lifting-based invertible motion adaptive transform (LIMAT) framework for highly scalable video compression,” *IEEE Trans. On Image Processing*, vol. 12, no. 12, pp.1530–1542, Dec. 2003.

[28] J.-R. Ohm, “Advances in scalable video coding,” *Proc. IEEE*, vol. 93, no. 1, pp. 42–56, Jan. 2005.

[29] F. Dufaux and J. Konrad, “Efficient, robust, and fast global motion estimation for video coding,” *IEEE Trans. Image Processing*, vol. 9, pp. 497–501, Mar. 2000.

[30] B. Feng, J. Xu, F. Wu, S. Yang, “Energy distributed update steps (EDU) in lifting based motion compensated video coding,” in *Proc. ICIP 2004*, pp. 2267–2270, Singapore, Oct. 2004.

[31] W. Yang, F. Wu, Y. Lu, J. Cai, K. Ngan, S. Li, “Scalable multi-view video coding using wavelet,” in *Proc. ISCAS 2005*, Kobe, Japan, May 2005.

[32] ISO/IEC JTC1/SC29/WG11 N6494, “Preliminary call for evidence on multiview video coding,” Redmond, USA, July 2004.

[33] K. Mueller, P. Merkle, A. Smolic and T. Wiegand, “Multiview coding using AVC,” MPEG2006/m12945, 75th MPEG meeting, Bangkok, Thailand, Jan. 2006.

Wenxian Yang received the B.Eng. degree from Zhejiang University, Hangzhou, China, in 2001 and the Ph.D. degree in Computer Engineering from Nanyang Technological University, Singapore, in 2006.

In 2004, Dr. Yang was with Microsoft Research Asia, Beijing, China for internship. From 2005 to 2006, she was a postdoctoral researcher in the French National Institute for Research in Computer Science and Control (INRIA-IRISA), France. She is now a research scholar in The Chinese University of Hong Kong. Her research interests include video compression, 3-D video compression and processing.

Yan Lu received his B.S., M.S. and Ph. D degrees in computer science from Harbin Institute of Technology, China, in 1997, in 1999 and in 2003, respectively.

From 1999 to 2000, Dr. Lu was a research assistant at the computer science department of City University of Hong Kong, Hong Kong SAR. From 2001 to 2004, he was with the Joint R&D Lab (JDL) for advanced computing and communication, Chinese Academy of Sciences, Beijing, China. Since April 2004, he has been with Microsoft Research Asia. His research interests include image and video coding, multimedia streaming, and texture compression.

Feng Wu (M’99-SM’06) received the B.S. degree in Electrical Engineering from XIDIAN University in 1992. He received the M.S. and Ph.D. degrees in Computer Science from Harbin Institute of Technology in 1996 and 1999, respectively.

Dr. Wu joined in Microsoft Research China as an associated researcher in 1999. He has been a researcher with Microsoft Research Asia since 2001. His research interests include image and video representation, media compression and communication, computer vision and graphics. He has been an active contributor to ISO/MPEG and ITU-T standards. Some techniques have been adopted by MPEG-4 FGS, H.264/MPEG-4 AVC and the coming H.264 SVC standard. He served as the chairman of China AVS video group in 2002~2004 and led the efforts on developing China AVS video standard 1.0. He has authored or co-authored over 100 conference and journal papers. He has about 30 U.S. patents granted or pending in video and image coding.

Jianfei Cai received his Ph.D degree from University of Missouri-Columbia in 2002.

Dr. Cai is an assistant professor with Nanyang Technological University, Singapore. His major research interests include digital media processing, multimedia compression, communications and networking technologies. He has published more than 50 technical papers in international conferences and journals. He has been actively participated in program committees of various conferences, and he is the mobile multimedia track co-chair for ICME 2006, the technical program co-chair for Multimedia Modeling (MMM) 2007 and the conference co-chair for Multimedia on Mobile Devices 2007. He is also an Associate Editor for IEEE Transactions on Circuits and Systems for Video Technology (T-CSVT).

King N. Ngan (M’79–SM’91–F’00) and holds a Ph.D. degree in Electrical Engineering from Loughborough University of Technology, U.K. He is a chair professor in the Department of Electronic Engineering, the Chinese University of Hong Kong, Hong Kong and was previously a full professor in the Nanyang Technological University, Singapore, and the University of Western Australia, Australia.

Professor Ngan is an associate editor of the Journal on Visual Communications and Image Representation, U.S.A., as well as an area editor of EURASIP Journal of Signal Processing: Image Communication, and served as an associate editor of IEEE Transactions on Circuits and Systems for Video Technology and Journal of Applied Signal Processing. He chaired a number of prestigious international conferences on video signal processing and communications and served on the advisory and technical committees of numerous professional organizations. He has published extensively including 3 authored books, 5 edited volumes and over 200 refereed technical papers in the areas of image/video coding and communications.

Professor Ngan is a Fellow of IEEE (U.S.A.), a Fellow of IEE (U.K.), and a Fellow of IEAust (Australia).

Shipeng Li received the B.S. and M.S. degrees from the University of Science and Technology of China (USTC), Hefei, China, in 1988 and 1991, respectively, and the Ph.D. degree from Lehigh University, Bethlehem, PA, in 1996, all in electrical engineering.

Dr. Li was with the Electrical Engineering Department, USTC, during 1991–1992. He was a Member of Technical Staff with Sarnoff Corporation, Princeton, NJ, during 1996–1999. He has been a Researcher with Microsoft Research Asia, Beijing, China, since May 1999 and has contributed some technologies in MPEG-4 and H.264. His research interests include image/video compression and communications, digital television, multimedia, and wireless communication.