

Fully Scalable Multiview Wavelet Video Coding

Yu Liu and King Ngi Ngan

Department of Electronic Engineering, The Chinese University of Hong Kong, Hong Kong SAR

Email: yliu@ieee.org, knngan@ee.cuhk.edu.hk

Abstract—This paper presents a 2-D pipeline-based locally adaptive inter-view-temporal lifting scheme for scalable multiview video coding. Instead of global temporal and inter-view correlation analysis based on the whole picture, the proposed method adopts locally adaptive inter-view-temporal correlation analysis based on the macroblocks. In addition, in order to reduce the memory requirement and computational complexity and remove both of temporal and view boundary effects, a new 2-D pipeline-based lifting scheme is proposed. Experimental results show that the proposed scheme always outperforms the other tested schemes in coding efficiency, while providing several advantages, such as view scalability, pipeline-based processing, and moderate complexity.

I. INTRODUCTION

Multiview video captured simultaneously by several cameras from different viewpoints, enables a wide variety of future multimedia applications such as free-viewpoint video, 3-D television, and photo-realistic rendering of 3-D scenes. However, it also results in huge amounts of data to be stored or transmitted to the user. Thus, efficient compression techniques are essential for realizing such applications. The straightforward solution for this would be to encode all the video signals independently using state-of-the-art video codecs such as H.264/AVC [1] or wavelet-based video codec VIDWAV [2]. However, multiview video contains a large amount of inter-view statistical dependencies, since all cameras capture the same scene from different viewpoints. These can be exploited by combined temporal/inter-view prediction, where images are not only predicted from temporally neighboring images but also from corresponding images in adjacent views, referred to as multiview video coding (MVC).

With the increasing diversity of network bandwidth, storage capability and display capability, one of important multiview video coding research topics is to achieve full scalability in all four dimensions: quality, spatial, temporal and view dimensions. Most scalable video codecs, such as wavelet-based SVC [2] or H.264/AVC-based JSVM [3], can achieve part of the requirement, but view scalability is missing. Although the upcoming H.264/AVC-based JMVM [4] offers some degree of temporal and view scalability, spatial or quality scalability is not supported at all. As an alternative to H.264/AVC-based JMVM, 4-D wavelet-based MVC promises to be an elegant way to provide such full scalability feature with non-redundant high-dimensional subband decomposition.

Schemes for wavelet-based MVC [5], [6] have been devised. They all have a common principle that the correlation of

multiview video is exploited by motion-compensated temporal filtering (MCTF) in the temporal domain, disparity-compensated view filtering (DCVF) in the view domain and 2-D discrete wavelet transform (DWT) in the spatial domain. The inherent scalability of such subband/wavelet decomposition is appealing. Although those MVC schemes are built upon the wavelet-based SVC [2], unfortunately no view scalability [5] or limited view scalability [6] is supported because of the fact that the global MCTF/DCVF selection is performed on the GoGOP (Group of GOP) basis. An attempt to provide full view scalability for wavelet-based MVC is made by Garbas et al. [7], but the coding efficiency is not satisfying because of rather simplistic experimental setup.

In this paper, we present a wavelet-based scalable MVC algorithm with full scalability in all four dimensions. The main contribution of this paper consists of two parts: (1) a locally adaptive prediction model based on macroblock is employed to exploit the inter-view-temporal correlation in the multiview video; and (2) a 2-D pipeline-based lifting scheme is proposed to reduce the memory requirement and computational complexity and remove both of temporal and view boundary effects.

The remainder of this paper is organized as follows. Section II first reviews the related work and then describes our proposed 2-D pipeline-based locally adaptive inter-view-temporal lifting scheme. The experimental results are presented in Section III, and Section IV concludes the paper.

II. PROPOSED METHOD

A. Related Work

Wavelet-based video coding provides an elegant and flexible way for the MVC framework by using high-dimensional wavelet transform. The most straightforward solution is the simulcast-based scalable wavelet video coding [2], which only exploits the temporal correlation and encodes all video views independently. Fig. 1(a) illustrates the simulcast-based wavelet decomposition structure along the temporal and view axes. As mentioned above, there exists a large amount of inter-view correlation in the multiview video. In order to achieve efficient coding performance, exploitation of these inter-view dependencies is indispensable.

In [5], Yang et al. first presented a regular decomposition structure and then proposed a more efficient decomposition structure derived from the global temporal and inter-view correlation analysis. The former is based on the assumption that the temporal correlation is always stronger than the inter-view one. Thus, in the regular decomposition structure, the DCVF

*This work was partially supported by a grant from the Chinese University of Hong Kong Direct Grant for Research Scheme (Project 2050383).

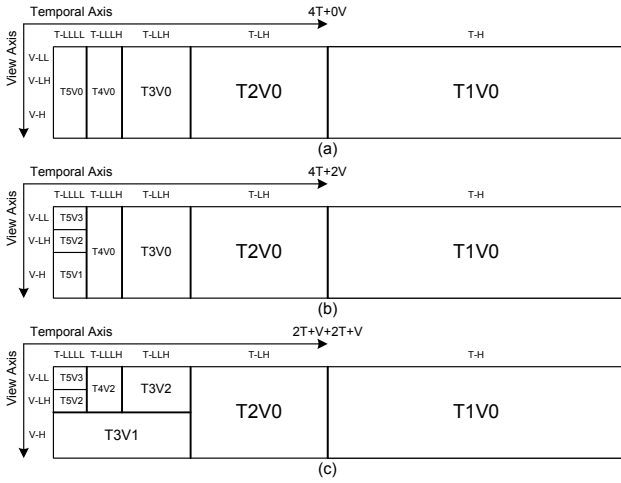


Fig. 1. Illustrations of (a) simulcast, (b) regular and (c) adaptive wavelet decomposition structures along the temporal and view axes

is performed after the multilevel MCTF has fully removed the temporal correlations. Fig. 1(b) shows an example of this structure, which consists of four levels of MCTF followed by two levels of DCVF. However, the above assumption 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 better coding efficiency is to derive the optimal decomposition structure based on the coding cost. Therefore, a more efficient adaptive decomposition structure is further proposed by Yang et al. [5] and Garbas et al. [6]. Instead of fixing the performing order of MCTF and DCVF, the MCTF and DCVF are interleaved based on the following coding cost,

$$H\text{Cost} = \text{SAD}(H) + \lambda R(MV) \quad (1)$$

where $\text{SAD}(H)$ is the sum of absolute difference (SAD) of samples in high-pass subbands H ; $R(MV)$ is the bit rate for coding motion or disparity information. Based on the analysis of the coding cost, the decomposition structure is adaptive to video content, so as to achieve high coding efficiency. Fig. 1(c) shows an example of this adaptive structure, where the MCTF and DCVF are interleaved together. Although the adaptive decomposition scheme achieves higher coding efficiency than the regular one, it still has some limitations because it adopts the global temporal and inter-view correlation analysis based on the whole picture. However, any global correlation analysis model cannot guarantee the perfect matching for all regions, even though in most cases, they dominate the correlation in the whole picture.

B. Locally Adaptive Inter-View-Temporal Structure

Fig. 2 analyzes the statistical properties of temporal and inter-view correlation on the macroblock level [8]. The purpose of the analysis is to determine by what percentage a rate-distortion optimized encoder would choose either one of these prediction modes, if all of temporal and inter-view neighbor pictures are available, as shown in Fig. 2(a). Fig. 2(b) gives the

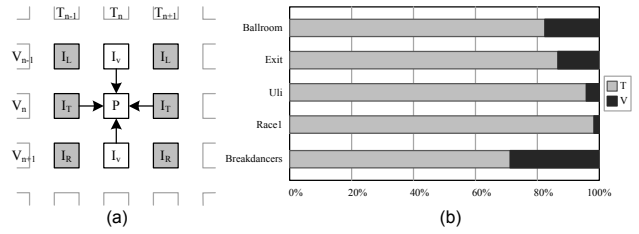


Fig. 2. Analysis of temporal and inter-view correlation. (a) Prediction modes with first order inter-view and temporal neighbor pictures, (b) Probabilities of prediction modes (T: temporal mode, V: inter-view mode) [8]

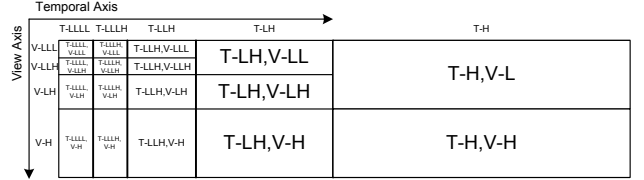


Fig. 3. Illustration of proposed locally adaptive inter-view-temporal wavelet decomposition structure along the temporal and view axes

results of this analysis over five multiview data sets, where the percentage values give the portion of blocks that are chosen for prediction with the reference picture of the corresponding mode. The results of this analysis over several multiview data sets show that sometimes inter-view prediction is more efficient than temporal prediction for a significant number of blocks, although for all sequences temporal prediction is the most often chosen mode. As verified by investigating the statistical dependencies between the temporal and inter-view pictures, considering a local correlation analysis model could further improve the coding efficiency.

Instead of the global temporal and inter-view correlation analysis based on the whole picture, our proposed wavelet decomposition structure adopts the locally adaptive inter-view-temporal correlation analysis based on the macroblocks. Each macroblock in the picture can be predicted by either temporal prediction mode or inter-view prediction mode according to the local macroblock-based cost function. If the temporal prediction mode is selected for a macroblock, the MCTF is performed on this macroblock. Otherwise, the DCVF is performed. Now, the problem is how to implement the lifting steps of the MCTF and DCVF on the macroblock level within the same framework. In order to implement it, the weighted lifting [9] is employed for the MCTF or DCVF. That is to say, in the update step, the high-pass coefficient will be updated exactly to the pixels they are predicted under the weight distribution constraint.

Fig. 3 shows an illustration of the proposed locally adaptive inter-view-temporal wavelet decomposition structure. The major difference between the proposed structure and conventional structures is that our inter-view transform is totally incorporated with temporal transform on the macroblock level within the same framework, leading to the proposed 2-D locally adaptive inter-view-temporal transform. The number of levels of inter-view transform at each temporal subband is related to that of temporal transform, but not in excess of the maximum predetermine number of levels of inter-view transform. In

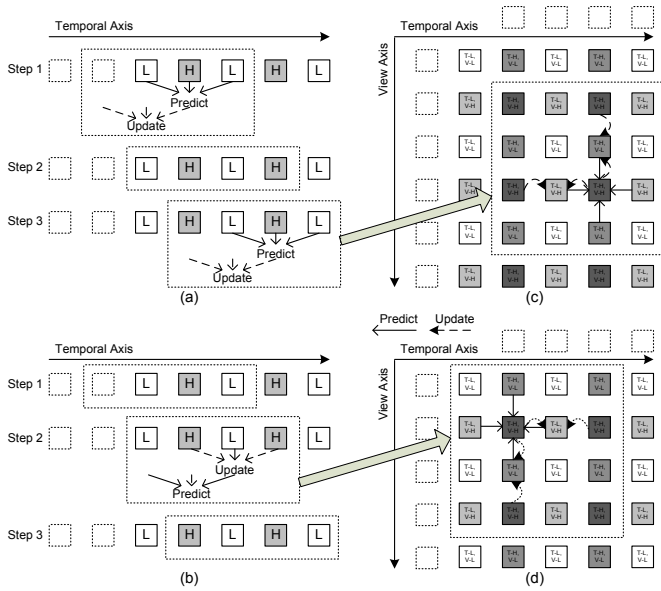


Fig. 4. Pipeline processing-based lifting schemes for 1-D case of (a) forward and (b) inverse temporal transform, and 2-D case of (c) forward and (d) inverse inter-view-temporal transform

Fig. 3, T- and V- denote temporal filtering and view filtering, respectively, while H and L denote high-pass and low-pass. The combination of the notations denotes the subband type. For example, (T-H,V-H) denotes that the subband is temporal high-pass and view high-pass filtered.

C. 2-D Pipeline-Based Lifting Scheme

The adaptive wavelet decomposition scheme [5], [6] in Fig. 1(c) is based on the hierarchical wavelet structure and the selection of optimal decomposition structure is performed on the basis of GoGOP, which contains the GOPs in all views at the same time interval. This significantly increases the memory requirement and computational complexity. In addition, it suffers from the problem of temporal boundary effects across GOP and thus its coding performance depends on the temporal GOP size.

In order to reduce the memory requirement and computational complexity and remove both of temporal and view boundary effects, a new 2-D pipeline-based lifting scheme is proposed. This 2-D lifting scheme is an extension of 1-D pipeline-based lifting scheme [10], which does not physically break the sequence into temporal GOPs but processes the sequence without intermission. As shown in Fig. 4(a), at the encoder side, the 1-D scheme uses exactly four frames for buffering, and then pushes video frames into the buffer one by one and outputs a wavelet transformed frame immediately whenever it is available. Fig. 4(b) shows the corresponding 1-D inverse lifting scheme at the decoder side.

Here, we extend the pipeline-based lifting scheme from 1-D temporal transform to 2-D inter-view-temporal transform. The proposed 2-D pipeline-based lifting scheme uses at most 4×4 frames for buffering. The video frames at the temporal axis are first pushed into the 4×4 buffering window sequentially like 1-D case, followed by those at the view axis. In other

words, the 2-D pipeline-based lifting scheme consists of two separable 1-D pipeline processes along the temporal or view axis. The temporal-axis pipeline process and the view-axis pipeline process are interleaved together. Therefore, the multiview sequence can be processed continuously without being broken into the GoGOP. Fig. 4 (c) and (d) show the corresponding forward and inverse 2-D pipeline-based lifting schemes. At the encoder side, the 2-D forward lifting scheme first predicts the (T-H,V-H) picture at the right-bottom side of the 4×4 buffering window from its four neighboring inter-view and temporal pictures, and then updates its preceding inter-view picture (T-L,V-H) and temporal picture (T-H,V-L). On the other hand, at the decoder side, the update step of the inter-view and temporal pictures are first performed, followed by the predict step of the (T-H,V-H) picture at the left-top side.

For multilevel wavelet transform, the *push* and *pull* models are adopted at the encoder and decoder sides, respectively. At the encoder side, the input frames in one level are pushed into the buffer of that level and once the outputs are ready, they are pushed into the next-level buffer or the final output buffer. At the decoder side, since the output frames should be decoded in natural order but the inputs are not, the decoder algorithm decides which frames should be pulled into the buffer of one level from the next-level buffer and the request is sent whenever the wavelet frames are needed.

III. EXPERIMENTAL RESULTS

In the experiments, two multiview QVGA (320×240) video sequences are used: *Ballroom* and *Race1*. Each multiview video includes eight views with frame rate of 25 Hz or 30 Hz, respectively, captured by a parallel camera array with 20 cm spacing. Different temporal/inter-view wavelet decomposition schemes are compared and are all built upon the wavelet-based SVC [2]. The coding results are obtained by encoding the first 128 frames of each sequence once to generate only one bitstream for one multiview video (including eight views), and then decoding is done by truncation of the embedded bitstream at the respective bit rate, spatial resolution, frame rate and even view rate (if any). The PSNR is averaged over all views and bit rate is given per view.

Fig. 5 shows the rate distortion (R-D) curves with respect to the coding results on the two multiview videos with eight views. We can observe that the proposed scheme always achieves the best coding performance amongst the tested schemes. The coding gain can be up to 1.18 dB and 0.88 dB at low bit rates over the regular and adaptive schemes. Compared with the simulcast scheme, significant improvement in coding performance is achieved and up to 2.49 dB gain is observed at low bit rate. Fig. 6 presents the performance comparison between the proposed scheme and the simulcast scheme for the two multiview videos at different bit rates, frame rates and view rates. As seen from the R-D curves at different test points, the proposed scheme outperforms consistently the simulcast scheme. Even for the two-view case, the proposed scheme still shows advantages over the simulcast scheme. Experimental results show that our proposed locally adaptive lifting scheme

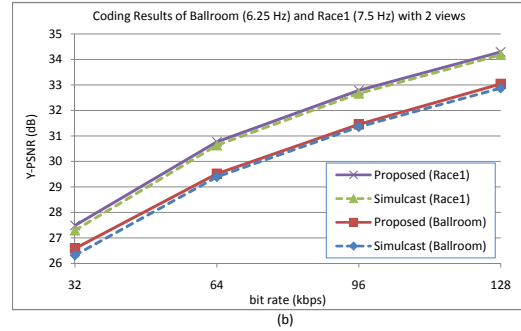
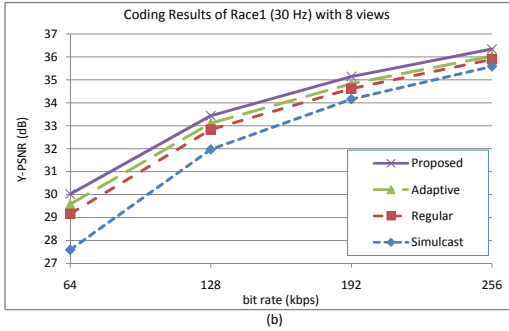
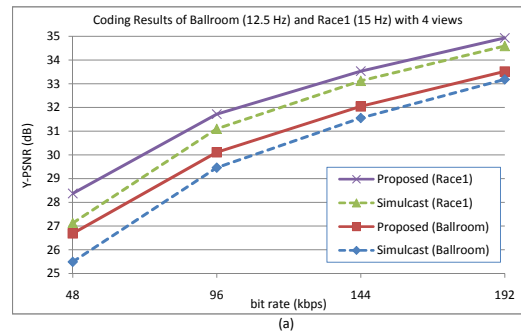
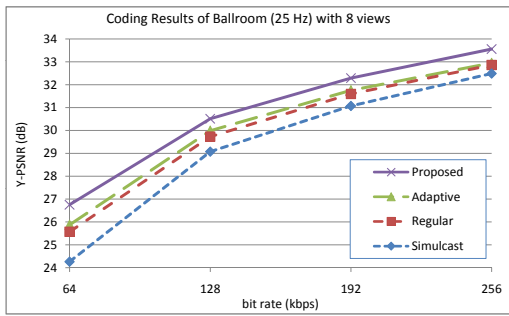


Fig. 5. Performance comparison of different wavelet decomposition schemes for tested multiview sequences (a) *Ballroom* and (b) *Race1*.

Fig. 6. Performance comparison between the proposed scheme and the simulcast scheme with different quality, temporal and view scalability.

can exploit efficiently local inter-view-temporal correlations on the macroblock level. It is worth pointing out that, in the experiments, we do not demonstrate the coding results for spatial scalability. The reason is that all of these schemes naturally inherit this scalability from wavelet-based SVC.

The analysis of memory and complexity is discussed as follows. It is certain that the simulcast scheme has always the least requirement of memory and computation among all the tested schemes, since it can exploit only the temporal correlation. The regular scheme can be implemented in the similar way as simulcast scheme, but only exploits the inter-view correlation at the lowest-pass temporal subbands. Therefore, it only slightly increases the requirement of memory and computation than the simulcast scheme. As discussed above, the adaptive scheme is performed on the GoGOP basis, and thus it significantly increases the memory and computational complexity. For example, when the temporal GOP size is set to 128 and the view GOP size to 8, the frame buffer size is set to at least 128×8 frames. The extra cost of the adaptive scheme is so huge that most of applications cannot afford the potentially huge memory and computational cost. However, this situation does not happen to our proposed scheme, since the 2-D pipeline-based inter-view-temporal lifting is performed on the macroblock basis and at most 4×4 frames are involved for one-level transform.

IV. CONCLUSION

In this paper, we present a new inter-view-temporal lifting-based wavelet coding technique for fully scalable multiview wavelet video coding. Compared with conventional wavelet-based MVC schemes, the proposed scheme can provide the following advantages: (1) it provides an important scalability

feature, i.e., view scalability, where conventional schemes do not support well; (2) it implements the 2-D pipeline-based lifting and thus removes both of temporal and view boundary effects; and (3) it improves further the coding efficiency by exploiting local inter-view-temporal correlation. Experimental results show that the proposed scheme consistently outperforms other wavelet-based MVC schemes in coding efficiency, while providing several advantages, such as view scalability, pipeline processing feature, and moderate complexity.

REFERENCES

- [1] ITU-T and ISO/IEC, "Advanced video coding for generic audiovisual services, ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC," 2005
- [2] R. Xiong, X. Ji, D. Zhang, J. Xu, G. Pau, M. Trocan and V. Botreau, "VIDWAV Wavelet Video Coding Specifications 2.0," *ISO/IEC JTC1/SC29/WG11/M12339*, 2005
- [3] J. Reichel, H. Schwarz, and M. Wien, "Joint Scalable Video Model (JSVM) 8.0," *ISO/IEC JTC1/SC29/WG11/8456*, Hangzhou, China, 2006
- [4] A. Vetro, P. Pandit, H. Kimata, and A. Smolic, "Joint Multiview Video Model (JMVM) 7.0," *ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6, JVT-Z207*, Antalya, TR, 20008
- [5] W. Yang, Y. Lu, F. Wu, J. Cai, K.N. Ngan, and S. Li, "4-D wavelet-based multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, pp.1385-1396, Nov. 2006.
- [6] J. Garbas and A. Kaup, "Wavelet-based multi-view video coding with spatial scalability," International Workshop on Multimedia Signal Processing (MMSP), Crete, Greece, Oct. 2007
- [7] J. Garbas, U. Fecker, T. Troger and A. Kaup, "4D scalable multi-view video coding using disparity compensated view filtering and motion compensated temporal filtering," International Workshop on Multimedia Signal Processing (MMSP), Victoria, Canada, Oct. 2006
- [8] P. Merkle, A. Smolic, K. Muller, and T. Wiegand, "Efficient prediction structures for multiview video coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol.17, no.11, pp.1461-1473, Nov. 2007
- [9] Y. Liu and K. N. Ngan, "Weighted adaptive lifting-based wavelet transform," *IEEE Int. Conf. Image Process.*, San Antonio, USA, 2007
- [10] J. Xu, Z. Xiong, S. Li, and Y.-Q. Zhang, "Memory-Constrained 3-D Wavelet Transform for Video Coding Without Boundary Effects," *IEEE Trans. Circuits Syst. Video Technol.*, vol.12, no.9, pp.812-818, Sep. 2002