Scalable distributed video coding based on block SW-SPIHT

Anhong Wang (王安红)[1,2], Yao Zhao (赵 耀)[1], Zhenfeng Zhu (朱振峰)[1], and Hao Wang (王 浩)[1]

[1] *Institute of Information Science, Beijing Jiaotong University, Beijing 100044*

[2] *Taiyuan University of Science and Technology, Taiyuan 030024*

Received September 5, 2006

Nowadays, distributed source coding (DSC) and distributed video coding (DVC) have been receiving more and more attention due to the distinct contributions to the easy encoding. At the same time, with more new requirements coming forth in the current network communication, the scalability of bit stream has been a new focus in the real applications. A scalable DVC scheme is presented without requiring layered coding in which the main attributions of DVC, namely the capabilities of easy encoding and robustness, are inherited remarkably and the property of scalability is also integrated simultaneously. Based on the block Slepian-Wolf set partitioning in hierarchical trees (SW-SPIHT), the Wyner-Ziv frames are encoded to get the scalable bit stream. In addition, the binary motion searching is explored at the decoder with the help of a rate-variable 'hash' from the encoder to improve the performance of the whole system. The final experimental results show that our system has higher peak signal-to-noise ratio (PSNR) than the pixel-domain DVC at the high bit rate. What is more, the scalability in signal-to-noise ratio (SNR) is also achieved satisfactorily.

OCIS codes: 100.2000, 040.7290, 330.7310, 100.7410.

Presently, the easy encoding is required by the friendly up-linking multimedia services. Conventional MPEG and H.26* cannot meet this need because of the complex motion estimation at the encoder. Based on the Slepian-Wolf[1] and Wyner-Ziv[2] theories, which have set solid foundation for easy encoding, distributed source coding (DSC) and distributed video coding (DVC) have shown great potential and achieved almost the same coding performance by exploiting dependences between sources at the decoder. Since then, lots of related works have been put forward. In Ref. [3], a syndrome-based PRISM scheme was proposed. The similar scheme taken by Anne and Girod can be referred to Ref. [4]. Based on the works in Refs. [3,4], some improvements have been exploited as shown in Refs. [5,6]. But unluckily, the aforementioned strategies only show DVC's efficiency in the view of easy encoding and robustness without considering the scalability of bit stream.

As a matter of fact, the scalability of bit stream has been considered as a crux in many real applications, for example, a set of heterogeneous mobile receivers may have various computational and display capabilities and/or channel capacities. However, only some tentative schemes have been proposed for scalable DVC, such as Refs. [7—9]. And these schemes are all built on a layered video framework, in which one standard video coding scheme is treated as the base layer. Particularly, the non-complete intra-frame encoding with motion estimation at the base layer is still adopted, which will bring some negative influences inevitably on the property of easy encoding at the encoder. In addition to this, the demerit of fragility to the lossy channel at the base layer is distinctly obvious because of the prediction shift in motion compensation.

In this paper, we will give more considerations to the scalable DVC and try to preserve the properties of easy encoding and robustness. A complete intra-frame encoding model based on the block Slepian-Wolf set partitioning in hierarchical trees (SW-SPIHT) is proposed for Wyner-Ziv frames. Similar to SPIHT, the block SW-SPIHT is provided with the embedded bit stream. And this embedded bit stream can possess more flexibly truncated rates than that in the layered coding. Enlightened by Ref. [10], which has applied SW-SPIHT to distributed hyperspectral imagery successfully and shown better performance than intra-frame SPIHT, we extend the idea of SW-SPIHT to wavelet block and develop a block SW-SPIHT technique. Additionally, a binary motion searching (BMS) at decoder with rate-adaptive 'hash' is proposed for block SW-SPIHT to improve the performance of the whole system. The rate-adaptive 'hash' in our case is based on some parity bits from a rate compatible channel coding, which is different from the fixed-rate 'hash' in Ref. [11]. Moreover, the complete intra-frame encoding takes on property of robustness. What we should note is that the 'hash' here refers to a kind of encoding-related information representation. Once sent to decoder, those assistant information contained in 'hash' can be expected reliably to be great helpful for motion searching at decoder.

The proposed scalable DVC for Wyner-Ziv frame is shown in Fig. 1, in which the even frame X_{2i} is the Wyner-Ziv frame and the odd frames X_{2i-1} and X_{2i+1} act as the key frames. To the key frames, the conventional SPIHT can be used, while to the Wyner-Ziv frame, the coding process is based on the following steps.

1) Intraframe encoding.

Step 1. Generating the wavelet blocks (WBs). The module of 'Generating WBs' refers to rearranging the wavelet coefficients to form cross-scale wavelet block (WB) as shown in Fig. 2. That is to say, a 3-scale (it can be extended to multi-scale easily) discrete wavelet

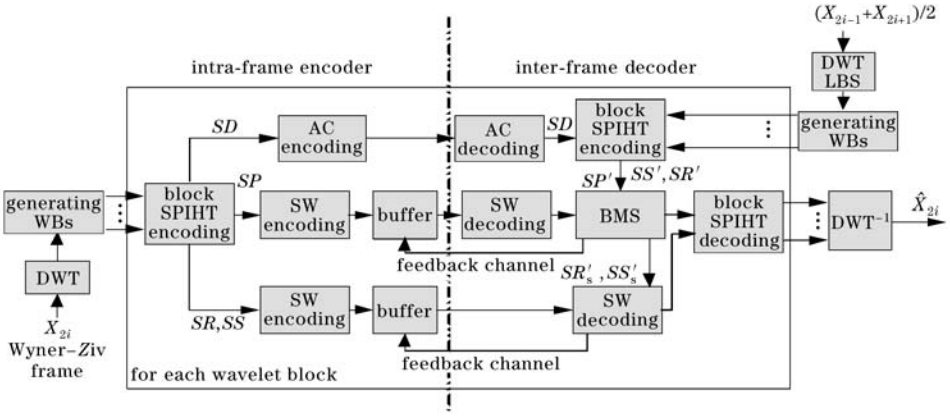

Fig. 1. Scalable DVC for Wyner-Ziv frames.

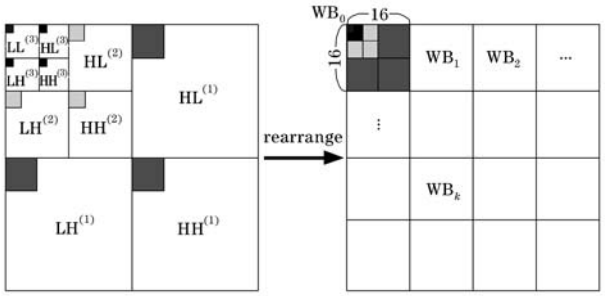

Fig. 2. Illustration of WB generation.

transformation (DWT) is implemented on Wyner-Ziv frames and the resulting DWT coefficients are subsequently partitioned into 16×16 cross-scale WBs. In each WB, there are four 2×2 coefficients vectors extracted from $\mathrm{LL}^{(3)}$, $\mathrm{HL}^{(3)}$, $\mathrm{LH}^{(3)}$, and $\mathrm{HH}^{(3)}$, three 4×4 coefficients vectors from $\mathrm{HL}^{(2)}$, $\mathrm{LH}^{(2)}$, and $\mathrm{HH}^{(2)}$, and three 8×8 coefficients vectors from $\mathrm{HL}^{(1)}$, $\mathrm{LH}^{(1)}$, and $\mathrm{HH}^{(1)}$. Obviously, since a WB is composed of several wavelet trees, it is convenient to apply the SPIHT to it.

Step 2. Block SPIHT encoding. 'Block SPIHT encoding' module denotes the SPIHT for WB, which is one contribution of this paper. 'Block SPIHT encoding' will output the tree distribution information SD, significance information SP, sign information SS, and the refinement information SR for each WB, which is similar to the conventional SPIHT encoding. But note that all WBs in the same wavelet image use the same original threshold, the maximum value of the whole wavelet image, for their block SPHIT. And this maximum value is sent to decoder. What mentioned above is named as block SPIHT, and some studies have shown that there is almost the same performance as the conventional SPIHT for the whole image, which sets solid foundation for combining block SPIHT with Slepian-Wolf theory.

Step 3. Arithmetic-encoding SD and sending it to decoder. In Fig. 1, the modules of 'AC encoding' and 'AC decoding' mean the arithmetic encoding and decoding respectively.

Step 4. Slepian-Wolf encoding SP, SS, and SR, and storing all of the parity bits in buffer. 'SW encoding' module is the Slepian-Wolf encoding based on a rate compatible channel coding with feedback. All of the outputted parity bits of the channel coding are stored in the buffer and the encoder will send the parity bits in batches on the demands from feedback channel.

2) Inter-frame decoding.

Step 1. Interpolating and generating the side WBs. In this paper, the averagely interpolated frame $(x_{2i-1} + x_{2i+1})/2$ is used as the initial side information. In order to improve the performance of the whole system, the wavelet-domain motion estimation based on DWT with low-band-shift (LBS)[12] is implemented. The construction of referenced WBs for motion searching is based on the idea of the wavelet domain motion searching. The side WBs co-located at some frames without LBS are referred to as the side information with non motion searching, and the other referenced WBs are referred as the one with motion searching.

Step 2. Block SPIHT encoding. This process is similar to that at encoder. But, there are differences from that at encoder, i.e., during the 'block SPIHT encoding' process at decoder, the decompressed SD is used to develop SP', SS', and SR' of the side WB.

Step 3. Block SW-SPIHT. At decoder, the decompressed SD is used to conduct block SPIHT for the side WB and we can get information SP', SS', and SR' for the side WB; Then, encoder sends the parity bits in batches on the demands from feedback channel and the channel decoding is repeated using the received parity bits and the side information SP', SS', and SR'. After channel decoding succeeds, the main sequences SP, SS, and SR are recovered losslessly. We name the overall procedure mentioned above block SW-SPIHT. Here, the goal of compression can be achieved correspondingly since fewer parity bits than the original are sent to decoder. And a presumption is set that the information of SPIHT is dependent if the WBs, the wavelet tree, are similar to each other. It is the dependence between main and side WBs that makes the compression of the SW-SPIHT possible.

Step 4. BMS with rate-variable 'hash' at decoder. The aforementioned block SW-SPIHT compresses the SPIHT stream by using the dependence between the main and co-located side WBs. But the dependence assumption is not workable all the time generally for the reason of WB's motion. Hence, the BMS with rate-variable 'hash'

at decoder can be implemented to find the matched side WB among the referenced WBs and to compress SP, SS, and SR further.

The variable 'hash' is denoted as the compressed SD and the partial parity bits of SP. What should be emphasized is that the implementation of the motion searching in DVC is moved to the decoder. But at the decoder, there is no any information available on current frame being encoded. Thus, it is necessary to send some more efficient information representing the current frame to the decoder in order to assist the motion estimation. This representation is just equivalent to the function of 'hash'. In Ref. [4], the 'hash' bits consist of a very coarsely sub-sampled and quantized version of the pixel block. And in Ref. [11], the quantized higher frequency coefficients of DCT act as 'hash'. But these 'hash's are fixed-rate. In our work, we try to take SD and SP as the 'hash' for assistant to doing motion searching. Since SD and SP are the bit-plane information of wavelet coefficients, accurate motion estimation can be obtained via these bit plane information. Besides, the arithmetic codec in our scheme is adopted for SD compression and a SW codec with feedback for SP. With the feedback imported, SP can be compressed by the least number of parity bits with searching over all the referenced side WBs, in which the 'hash' bits vary with the dependence of WBs.

Specially, giving a main WB with information SD, SP, SS, SR, let B_i be the i-th referenced side WB corresponding to SD, SP'_i, SS'_i, SR'_i, $i \in (1, \cdots, N)$ and N be the number of referenced WBs. In our case, SP'_i, SS'_i, and SR'_i can be obtained expediently by block SPIHT encoding mentioned above with the decompressed SD as the distributed information. The detailed searching procedure is as follows.

1) SW-encoding SP and storing all of its parity bits in buffer.

2) Sending the first part of parity bits of SP to decoder and judging if SP can be decoded correctly from SP'_1. If decoding succeeds, go to the step 5); otherwise, $i = i + 1$ and try the next referenced WB.

3) Sending more partial parity bits if SP cannot be decoded correctly by all referenced WBs in step 2).

4) Repeating steps 2) and 3) until SP is recovered with the least parity bits.

5) Finding a side B_s whose SP'_s is the nearest to the recovered SP in the Hamming distance space and getting the corresponding SS'_s and SR'_s.

Summarily, as mentioned above, the proposed scalable DVC shares the following three properties and advantages:

1) Easy encoding. The Wyner-Ziv encoding scheme consists of a DWT, a SPIHT, and two channel codings, which makes the complexity of the encoding similar to the conventional intra-frame encoding. However, although with lower computational complexity of encoding than the conventional inter-frame one, more computational expenses are involved in the procedure of decoding than conventional ones because of the channel decoding and the BMS at decoder.

2) More general robustness. Since only BMS is taken at decoder without implementing any motion searching at encoder, the prediction shift in conventional inter-frame encoding can be avoided evidently, which leads to the property of more general robustness.

3) Scalability. With the embedded bit stream like conventional SPIHT, the bit stream can be truncated flexibly. While it shows scalable property related to signal-to-noise ratio (SNR), it can be obviously extended to temporal and frequency domain as well.

Now, we do comparison experiments. Similar to Ref. [13], only the luminance is tested for the first 101 frames of QCIF Carphone and 361 frames of Foreman with 30-Hz frame rate. The even frames are encoded by Wyner-Ziv model, in which the SW codec is based on an accumulative low density parity code (LDPC)[14]. The easy average-interpolation of adjacent frames is taken as the side information with assumption of odd frames decoded completely. During the process of BMS for SW-SPIHT, the block searching region is set to be $dx = \pm 16$, $dy = \pm 16$. Out of the consideration of comparison, only the peak signal-to-noise ratios (PSNRs) of the even frames are averaged and the frame rate is 15 Hz. To make overall evaluation for our proposed scalable DVC framework, four aspects of experiments include:

1) Inter-frame coding performance of H.263+ (I-B-I-B); 2) Performance comparison of conventional SPIHT and the proposed block-SPIHT; 3) Proposed block SW-SPIHT without BMS testing the capability of block SW-SPIHT; 4) Proposed block SW-SPIHT with BMS showing the efficiency of the BMS with variable 'hash'.

The evaluation on the quality of the recovered image is shown in Fig. 3, from which we can see that the block SPIHT keeps almost the same performance as SPIHT, and the SW-SPIHT can bring more than 2 dB improvement than SPIHT because of utilizing of dependence at

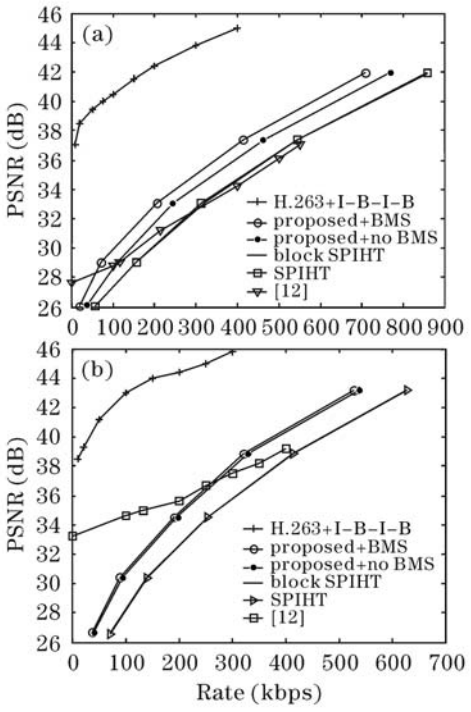

Fig. 3. Rate-distortion curves of the proposed scalable DVC framework. (a) Foreman sequence (361 frames); (b) Carphone sequence (101 frames).

Fig. 4. Original 2nd frame of Foreman sequence (left) and the reconstructed one based on our scheme with 100% data (right).

the decoder. Additionally, the BMS at decoder gives more than 1 dB improvement for block SW-SPIHT. Nevertheless, for Carphone sequence, the BMS does not work well as expected due to its less motion.

Moreover, our system has more than 2 dB improvement than Anne's pixel-domain DVC for Foreman sequence when only the easy average interpolation is used at decoder. This is because that the DWT and SPIHT can exploit the dependence better at the encoder and the BMS is implemented at the decoder. What is more, our scheme is scalable and the truncated rate points are flexible similar to SPIHT for intra-frame coding.

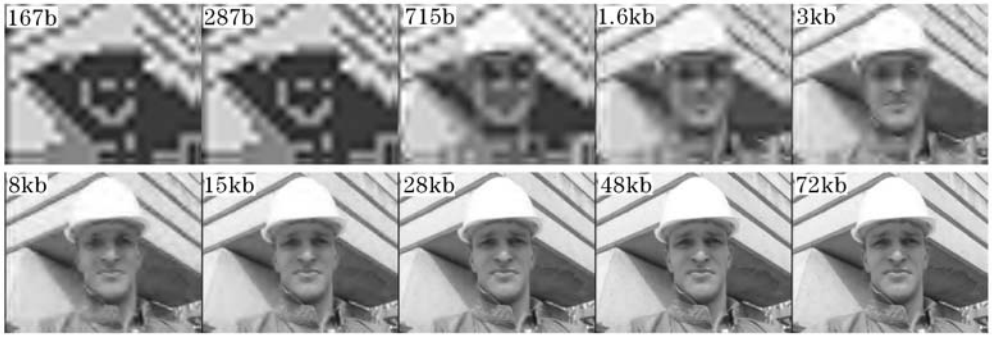

Fig. 5. Scalable decoded Wyner-Ziv frame (2nd frame) of Foreman sequence with different bits.

One point should be noted is that the proposed scalable DVC has lower PSNR at low bit rate. The reason is that only the intra-frame decoding is taken during reconstruction stage without side information applied. In addition, as far as the computational expense at decoder is concerned, the efficiency embodied in our system is much higher with the consideration of motion searching, which yet makes no influence on the capability of easy encoding. We give the original 2nd frame of Foreman sequence and the reconstructed one based on our scheme with 100% data in Fig. 4, which verifies the higher efficiency of the proposed scheme. Finally, the scalable decoded frames at different bits are shown in Fig. 5.

In conclusion, we have introduced a scalable DVC based on the embedded bit stream of block SPIHT in this paper. Aside from its easy implementation, the main property of DVC can be preserved at one time. Although the current experimental results are encouraging, much endeavor remains to be done. For example, the feedback in Slepian-Wolf coding will make the system invalid when there is no feedback channel provided. For the aspect of robustness, the SPIHT is yet sensitive to channel error, and the robustness won't be guaranteed even a bit errors.

This work was supported in part by the National Natural Science Foundation of China (No. 90604032, 60373028), the Specialized Research Fund for the Doctoral Program of Higher Education, and the Program for New Century Excellent Talents in University, the Specialized Research Foundation of BJTU, and the "973" Program of China (No. 2006CB303104). A. Wang's e-mail address is wah_ty@yahoo.com.cn, and Y. Zhao's e-mail address is yzhao@center.njtu.edu.cn.

References

1. D. Slepian and J. K. Wolf, IEEE Trans. Information Theory **19,** 471 (1973).
2. A. D. Wyner and J. Ziv, IEEE Trans. Information Theory **22,** 1 (1976).
3. R. Puri and K. Ramchandran, in *Proceedings of ICASSP'2003* 856 (2003).
4. A. Aaron, D. Varodayan, and B. Girod, in *Proceedings of PCS'06* 235 (2006).
5. J. E. Fowler, M. Tagliasacchi, and B. Pesquel-Popescu, Wavelet-based distributed source coding of video, http://www.ee.bilkent.edu.tr/∼signal/defevent/papers/cr1535.pdf.
6. A. Wang, Y. Zhao, and L. Wei, in *Proceedings of ICIP* 241 (2006).
7. Q. Xu and Z. Xiong, IEEE Trans. Image Processing **15,** 3791 (2006).
8. A. Sehgal, A. Jagmohan, and N. Ahuja, in *Proceedings of PCS'04* (2004).
9. M. Tagliasacchi, A. Majumdar, and K. Ramchandran, in *Proceedings of PCS'04* (2004).
10. C. Tang, N. Cheung, A. Ortega and C. S. Raghavendra, in *Proceedings of DCC'05* 437 (2005).
11. A. Aaron, S. Rane, and B. Griod, in *Proceedings of ICIP* 3097 (2004).
12. H.-W. Park and H.-S. Kim, IEEE Trans. Image Processing **9,** 577 (2000).
13. A. Aaron, S. Rane, E. Setton, and B. Griod, in *Proceedings of Asilomar Conference on Signals and Systems* 3 (2002).
14. D. Varodayan, A. Aaron, and B. Girod, in *Proceedings of Asilomar Conference on Signals, Systems and Computers* 1203 (2005).